A Generalized Markov-Chain Modelling Approach to $(1, \lambda)$-ES Linear Optimization: Technical Report

Alexandre Chotard, Martin Holena

A Generalized Markov-Chain Modelling Approach to $(1, \lambda)$-ES Linear Optimization: Technical Report

Alexandre Chotard [1] and Martin Holeňa [2]

[1] INRIA Saclay-Ile-de-France, LRI, alexandre.chotard@lri.fr University Paris-Sud, France
[2] Institute of Computer Science, Academy of Sciences, Pod vodárenskou věží 2, Prague, Czech Republic, martin@cs.cas.cz

Abstract. Several recent publications investigated Markov-chain modelling of linear optimization by a $(1, \lambda)$-ES, considering both unconstrained and linearly constrained optimization, and both constant and varying step size. All of them assume normality of the involved random steps, and while this is consistent with a black-box scenario, information on the function to be optimized (e.g. separability) may be exploited by the use of another distribution. The objective of our contribution is to complement previous studies realized with normal steps, and to give sufficient conditions on the distribution of the random steps for the success of a constant step-size $(1, \lambda)$-ES on the simple problem of a linear function with a linear constraint. The decomposition of a multidimensional distribution into its marginals and the copula combining them is applied to the new distributional assumptions, particular attention being paid to distributions with Archimedean copulas.

Keywords: evolution strategies, continuous optimization, linear optimization, linear constraint, linear function, Markov chain models, Archimedean copulas

1 Introduction

Evolution Strategies (ES) are Derivative Free Optimization (DFO) methods, and as such are suited for the optimization of numerical problems in a black-box context, where the algorithm has no information on the function f it optimizes (e.g. existence of gradient) and can only query the function's values. In such a context, it is natural to assume normality of the random steps, as the normal distribution has maximum entropy for given mean and variance, meaning that it is the most general assumption one can make without the use of additional information on f. However such additional information may be available, and then using normal steps may not be optimal. Cases where different distributions have been studied include so-called Fast Evolution Strategies [1] or SNES [2, 3] which exploits the separability of f, or heavy-tail distributions on multimodal problems [4, 3].

In several recent publications [5–8], attention has been paid to Markov-chain modelling of linear optimization by a $(1, \lambda)$-ES, i.e. by an evolution strategy in which λ children are generated from a single parent $\boldsymbol{X} \in \mathbb{R}^n$ by adding normally distributed n-dimensional random steps $\boldsymbol{M}$,

$$\boldsymbol{X} \leftarrow \boldsymbol{X} + \sigma \mathbf{C}^{\frac{1}{2}} \boldsymbol{M}, \text{ where } \boldsymbol{M} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n). \tag{1}$$

Here, σ is called step size, $\mathbf{C}$ is a covariance matrix, and $\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ denotes the n-dimensional standard normal distribution with zero mean and covariance matrix identity. The best among the λ children, i.e. the one with the highest fitness, becomes the parent of the next generation, and the step-size σ and the covariance matrix $\mathbf{C}$ may then be adapted to increase the probability of sampling better children. In this paper we relax the normality assumption of the movement $\boldsymbol{M}$ to a more general distribution H.

The linear function models a situation where the step-size is relatively small compared to the distance towards a local optimum. This is a simple problem that must be solved by any effective evolution strategy by diverging with positive increments of $\nabla f.\boldsymbol{M}$. This unconstrained case was studied in [7] for normal steps with cumulative step-size adaptation (the step-size adaptation mechanism in CMA-ES [9]).

Linear constraints naturally arise in real-world problems (e.g. need for positive values, box constraints) and also model a step-size relatively small compared to the curvature of the constraint. Many techniques to handle constraints in randomised algorithms have been proposed (see [10]). In this paper we focus on the resampling method, which consists in resampling any unfeasible candidate until a feasible one is sampled. We chose this method as it makes the algorithm easier to study, and is consistent with the previous studies assuming normal steps [11, 5, 6, 8], studying constant step-size, self adaptation and cumulative step-size adaptation mechanisms (with fixed covariance matrix).

Our aim is to study the $(1, \lambda)$-ES with constant step-size, constant covariance matrix and random steps with a general absolutely continuous distribution H optimizing a linear function under a linear constraint handled through resampling. We want to extend the results obtained in [5, 8] using the theory of Markov chains. It is our hope that such results will help in designing new algorithms using information on the objective function to make non-normal steps. We pay a special attention to distributions with Archimedean copulas, which are a particularly well transparent alternative to the normal distribution. Such distributions have been recently considered in the Estimation of Distribution Algorithms [12, 13], continuing the trend of using copulas in that kind of evolutionary optimization algorithms [14].

In the next section, the basic setting for modelling the considered evolutionary optimization task is formally defined. In Section 3, the distributions of the feasible steps and of the selected steps are linked to the distribution of the random steps, and another way to sample them is provided. In Section 4, it is shown that, under some conditions on the distribution of the random steps, the normalized distance to the constraint is a ergodic Markov chain, and a law of large numbers for Markov chains is applied. Finally, Section 5 gives properties

on the distribution of the random steps under which some of the aforementioned conditions are verified.

Notations

For $(a, b) \in \mathbb{N}^2$ with $a < b$, $[a..b]$ denotes the set of integers i such that $a \leq i \leq b$. For X and Y two random vectors, $X \overset{(d)}{=} Y$ denotes that these variables are equal in distribution, $X \overset{a.s.}{\to} Y$ and $X \overset{\mathcal{P}}{\to} Y$ denote, respectively, almost sure convergence and convergence in probability. For $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^n$, $\boldsymbol{x}.\boldsymbol{y}$ denotes the scalar product between the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, and for $i \in [1..n]$, $[\boldsymbol{x}]_i$ denotes the i^{th} coordinate of $\boldsymbol{x}$. For A a subset of $\mathbb{R}^n$, 1_A denotes the indicator function of A. For $\mathcal{X}$ a topological set, $\mathcal{B}(\mathcal{X})$ denotes the Borel algebra on $\mathcal{X}$.

2 Problem setting and algorithm definition

Throughout this paper, we study a $(1, \lambda)$-ES optimizing a linear function $f : \mathbb{R}^n \to \mathbb{R}$ where $\lambda \geq 2$ and $n \geq 2$, with a linear constraint $g : \mathbb{R}^n \to \mathbb{R}$, handling the constraint by resampling unfeasible solutions until a feasible solution is sampled.

Take $(\boldsymbol{e}_k)_{k\in[1..n]}$ a orthonormal basis of $\mathbb{R}^n$. We may assume ∇f to be normalized as the behaviour of an ES is invariant to the composition of the objective function by a strictly increasing function (e.g. $h : x \mapsto x/\|\nabla f\|$), and the same holds for ∇g since our constraint handling method depends only on the inequality $g(\boldsymbol{x}) \leq 0$ which is invariant to the composition of g by a homothetic transformation. Hence w.l.o.g. we assume that $\nabla f = \boldsymbol{e}_1$ and $\nabla g = \cos\theta \boldsymbol{e}_1 + \sin\theta \boldsymbol{e}_2$ with the set of feasible solutions $\mathcal{X}_{\text{feasible}} := \{\boldsymbol{x} \in \mathbb{R}^n | g(\boldsymbol{x}) \leq 0\}$. We restrict our study to $\theta \in (0, \pi/2)$. Overall the problem reads

$$\begin{array}{c} \text{maximize} \quad f(\boldsymbol{x}) = [\boldsymbol{x}]_1 \quad \text{subject to} \\ g(\boldsymbol{x}) = [\boldsymbol{x}]_1 \cos\theta + [\boldsymbol{x}]_2 \sin\theta \leq 0 \ . \end{array} \tag{2}$$

At iteration $t \in \mathbb{N}$, from a so-called parent point $\boldsymbol{X}_t \in \mathcal{X}_{\text{feasible}}$ and with step-size $\sigma_t \in \mathbb{R}_+^*$ we sample new candidate solutions by adding to $\boldsymbol{X}_t$ a random vector $\sigma_t \boldsymbol{M}_t^{i,j}$ where $\boldsymbol{M}_t^{i,j}$ is called a random step and $(\boldsymbol{M}_t^{i,j})_{i\in[1..\lambda],j\in\mathbb{N},t\in\mathbb{N}}$ is a i.i.d. sequence of random vectors with distribution H. The i index stands for the λ new samples to be generated, and the j index stands for the unbounded number of samples used by the resampling. We denote $\boldsymbol{M}_t^i$ a feasible step, that is the first element of $(\boldsymbol{M}_t^{i,j})_{j\in\mathbb{N}}$ such that $\boldsymbol{X}_t + \sigma_t \boldsymbol{M}_t^i \in \mathcal{X}_{\text{feasible}}$ (random steps are sampled until a suitable candidate is found). The i^{th} feasible solution $\boldsymbol{Y}_t^i$ is then

$$\boldsymbol{Y}_t^i := \boldsymbol{X}_t + \sigma_t \boldsymbol{M}_t^i \ . \tag{3}$$

Then we denote $\star := \text{argmax}_{i\in[1..\lambda]} f(\boldsymbol{Y}_t^i)$ the index of the feasible solution maximizing the function f, and update the parent point

$$\boldsymbol{X}_{t+1} := \boldsymbol{Y}_t^\star = \boldsymbol{X}_t + \sigma_t \boldsymbol{M}_t^\star \ , \tag{4}$$

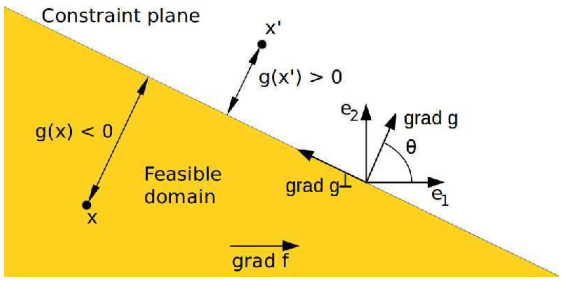

Fig. 1. Linear function with a linear constraint, in the plane spanned by ∇f and ∇g, with the angle from ∇f to ∇g equal to $\theta \in (0, \pi/2)$. The point $\boldsymbol{x}$ is at distance $g(\boldsymbol{x})$ from the constraint hyperplan $g(\boldsymbol{x}) = 0$.

where $\boldsymbol{M}_t^\star$ is called the selected step. Then the step-size σ_t, the distribution of the random steps H or other internal parameters may be adapted.

Following [5, 6, 11, 8] we define δ_t as

$$\delta_t := -\frac{g(\boldsymbol{X}_t)}{\sigma_t} \quad . \tag{5}$$

3 Distribution of the feasible and selected steps

In this section we link the distributions of the random vectors $\boldsymbol{M}_t^i$ and $\boldsymbol{M}_t^\star$ to the distribution of the random steps $\boldsymbol{M}_t^{i,j}$, and give another way to sample $\boldsymbol{M}_t^i$ and $\boldsymbol{M}_t^\star$ not requiring an unbounded number of samples.

Lemma 1. *Let a $(1, \lambda)$-ES optimize the problem defined in (2) handling constraint through resampling. Take H the distribution of the random step $\boldsymbol{M}_t^{i,j}$, and for $\delta \in \mathbb{R}_+^*$ denote $L_\delta := \{\boldsymbol{x} \in \mathbb{R}^n | g(\boldsymbol{x}) \leq \delta\}$. Providing that H is absolutely continuous and that $H(L_\delta) > 0$ for all $\delta \in \mathbb{R}_+$, the distribution $\tilde{H}_\delta$ of the feasible step and $\tilde{H}_\delta^\star$ the distribution of the selected step when $\delta_t = \delta$ are absolutely continuous, and denoting h, $\tilde{h}_\delta$ and $\tilde{h}_\delta^\star$ the probability density functions of, respectively, the random step, the feasible step $\boldsymbol{M}_t^i$ and the selected step $\boldsymbol{M}_t^\star$ when $\delta_t = \delta$*

$$\tilde{h}_\delta(\boldsymbol{x}) = \frac{h(\boldsymbol{x}) 1_{L_\delta}(\boldsymbol{x})}{H(L_\delta)} \quad , \tag{6}$$

and

$$\begin{aligned} \tilde{h}_\delta^\star(\boldsymbol{x}) &= \lambda \tilde{h}_\delta(\boldsymbol{x}) \tilde{H}_\delta((-\infty, [\boldsymbol{x}]_1) \times \mathbb{R}^{n-1})^{\lambda-1} \\ &= \lambda \frac{h(\boldsymbol{x}) 1_{L_\delta}(\boldsymbol{x}) H((-\infty, [\boldsymbol{x}]_1) \times \mathbb{R}^{n-1} \cap L_\delta)^{\lambda-1}}{H(L_\delta)^\lambda} \quad . \end{aligned} \tag{7}$$

Proof. Let $\delta > 0, A \in \mathcal{B}(\mathbb{R}^n)$. Then for $t \in \mathbb{N}, i = 1 \dots \lambda$, using the the fact that $(\boldsymbol{M}_t^{i,j})_{j\in\mathbb{N}}$ is a i.i.d. sequence

$$\begin{aligned}
\tilde{H}_\delta(A) &= \Pr(\boldsymbol{M}_t^i \in A | \delta_t = \delta) \\
&= \sum_{j\in\mathbb{N}} \Pr(\boldsymbol{M}_t^{i,j} \in A \cap L_\delta \text{ and } \forall k < j, \boldsymbol{M}_t^{i,k} \in L_\delta^c | \delta_t = \delta) \\
&= \sum_{j\in\mathbb{N}} \Pr(\boldsymbol{M}_t^{i,j} \in A \cap L_\delta | \delta_t = \delta) \Pr(\forall k < j, \boldsymbol{M}_t^{i,k} \in L_\delta^c | \delta_t = \delta) \\
&= \sum_{j\in\mathbb{N}} H(A \cap L_\delta)(1 - H(L_\delta))^j \\
&= \frac{H(A \cap L_\delta)}{H(L_\delta)} = \int_A \frac{h(\boldsymbol{x}) 1_{L_\delta}(\boldsymbol{x}) d\boldsymbol{x}}{H(L_\delta)} \ ,
\end{aligned}$$

which yield Eq. (6) and that $\tilde{H}_\delta$ admits a density $\tilde{h}_\delta$ and is therefore absolutely continuous.

Since $((\boldsymbol{M}_t^{i,j})_{j\in\mathbb{N}})_{i\in[1..\lambda]}$ is i.i.d., $(\boldsymbol{M}_t^i)_{i\in[1..\lambda]}$ is i.i.d. and

$$\begin{aligned}
\tilde{H}_\delta^\star(A) &= \Pr(\boldsymbol{M}_t^\star \in A | \delta_t = \delta) \\
&= \sum_{i=1}^{\lambda} \Pr(\boldsymbol{M}_t^i \in A \text{ and } \forall j \in [1..\lambda] \backslash \{i\}, [\boldsymbol{M}_t^i]_1 > [\boldsymbol{M}_t^j]_1 | \delta_t = \delta) \\
&= \lambda \Pr(\boldsymbol{M}_t^1 \in A \text{ and } \forall j \in [2..\lambda], [\boldsymbol{M}_t^1]_1 > [\boldsymbol{M}_t^j]_1 | \delta_t = \delta) \\
&= \lambda \int_A \tilde{h}_\delta(\boldsymbol{x}) \Pr(\forall j \in [2..\lambda], [\boldsymbol{M}_t^j]_1 < [\boldsymbol{x}]_1 | \delta_t = \delta) d\boldsymbol{x} \\
&= \int_A \lambda \tilde{h}_\delta(\boldsymbol{x}) \tilde{H}_\delta((-\infty, [\boldsymbol{x}]_1) \times \mathbb{R}^{n-1})^{\lambda-1} d\boldsymbol{x} \ ,
\end{aligned}$$

which shows that $\tilde{H}_\delta^\star$ possess a density, and with (6) yield Eq. (7). □

The vectors $(\boldsymbol{M}_t^i)_{i\in[1..\lambda]}$ and $\boldsymbol{M}_t^\star$ are functions of the vectors $(\boldsymbol{M}_t^{i,j})_{i\in[1..\lambda],j\in\mathbb{N}}$ and of δ_t. In the following Lemma an equivalent way to sample $\boldsymbol{M}_t^i$ and $\boldsymbol{M}_t^\star$ is given which uses a finite number of samples. This method is useful if one wants to avoid dealing with the infinite dimension space implied by the sequence $(\boldsymbol{M}_t^{i,j})_{i\in[1..\lambda,j\in\mathbb{N}}$.

Lemma 2. *Let a $(1, \lambda)$-ES optimize problem (2), handling the constraint through resampling, and take δ_t as defined in (5). Let H denote the distribution of $\boldsymbol{M}_t^{i,j}$ that we assume absolutely continuous, $\nabla g^\perp := -\sin\theta \boldsymbol{e}_1 + \cos\theta \boldsymbol{e}_2$, $\mathbf{Q}$ the rotation matrix of angle θ changing $(\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_n)$ into $(\nabla g, \nabla g^\perp, \dots, \boldsymbol{e}_n)$. Take $F_{1,\delta}(x) := \Pr(\boldsymbol{M}_t^i.\nabla g \le x | \delta_t = \delta)$, $F_{2,\delta}(x) := \Pr(\boldsymbol{M}_t^i.\nabla g^\perp \le x | \delta_t = \delta)$ and $F_{k,\delta}(x) := \Pr([\boldsymbol{M}_t^i]_k \le x | \delta_t = \delta)$ for $k \in [3..n]$, the marginal cumulative distribution functions when $\delta_t = \delta$, and C_δ the copula of $(\boldsymbol{M}_t^i.\nabla g, \boldsymbol{M}_t^i.\nabla g^\perp, \dots, \boldsymbol{M}_t^i.\boldsymbol{e}_n)$.*

We define

$$\mathcal{G} : (\delta, (u_i)_{i\in[1..n]}) \in \mathbb{R}_+ \times [0,1]^n \mapsto \mathbf{Q} \begin{pmatrix} F_{1,\delta}^{-1}(u_1) \\ \vdots \\ F_{n,\delta}^{-1}(u_n) \end{pmatrix} \ , \tag{8}$$

$$\mathcal{G}^\star : (\delta, (\boldsymbol{v}_i)_{i\in[1..\lambda]}) \in \mathbb{R}_+ \times [0,1]^{n\lambda} \mapsto \underset{\boldsymbol{G}\in\{\mathcal{G}(\delta,\boldsymbol{v}_i)|i\in[1..\lambda]\}}{\operatorname{argmax}} f(\boldsymbol{G}) \ . \tag{9}$$

Then, if the copula C_δ is constant in regard to δ, for $\mathbf{W}_t = (\boldsymbol{V}_{i,t})_{i\in[1..\lambda]}$ a i.i.d. sequence with $\boldsymbol{V}_{i,t} \sim C_\delta$

$$\mathcal{G}(\delta_t, \boldsymbol{V}_{i,t}) \overset{(d)}{=} \boldsymbol{M}_t^i \ , \tag{10}$$

$$\mathcal{G}^\star(\delta_t, \mathbf{W}_t) \overset{(d)}{=} \boldsymbol{M}_t^\star \ . \tag{11}$$

Proof. Since $\boldsymbol{V}_{i,t} \sim \mathbf{C}_\delta$

$$(\boldsymbol{M}_t^i.\nabla g, \boldsymbol{M}_t^i.\nabla g^\perp, \ldots, \boldsymbol{M}_t^i.\boldsymbol{e}_n) \overset{(d)}{=} (F_{1,\delta}^{-1}(\boldsymbol{V}_{1,t}), F_{2,\delta}^{-1}(\boldsymbol{V}_{2,t}), \ldots, F_{n,\delta}^{-1}(\boldsymbol{V}_{n,t})) \ ,$$

and if the function $\delta \in \mathbb{R}_+ \mapsto C_\delta$ is constant, then the sequence of random vectors $(\boldsymbol{V}_{i,t})_{i\in[1..\lambda],t\in\mathbb{N}}$ is i.i.d.. Finally by definition $\mathbf{Q}^{-1}\boldsymbol{M}_t^i = (\boldsymbol{M}_t^i.\nabla g, \boldsymbol{M}_t^i.\nabla g^\perp, \ldots, \boldsymbol{M}_t^i.\boldsymbol{e}_n)$, which shows Eq. (10). Eq. (11) is a direct consequence of Eq. (10) and the fact that $\boldsymbol{M}_t^\star = \underset{\boldsymbol{G}\in\{\mathcal{G}(\delta,\boldsymbol{v}_i)|i\in[1..\lambda]\}}{\operatorname{argmax}} f(\boldsymbol{G})$ (which holds as f is linear). □

We may now use these results to show the divergence of the algorithm when the step-size is constant, using the theory of Markov chains [15].

4 Divergence of the $(1, \lambda)$-ES with constant step-size

Following the first part of [8], we restrict our attention to the constant step size in the remainder of the paper, that is for all $t \in \mathbb{N}$ we take $\sigma_t = \sigma \in \mathbb{R}_+^*$.

From Eq. (4), by recurrence and dividing by t, we see that

$$\frac{[\boldsymbol{X}_t - \boldsymbol{X}_0]_1}{t} = \frac{\sigma}{t} \sum_{i=0}^{t-1} \boldsymbol{M}_i^\star \ . \tag{12}$$

The latter term suggests the use of a Law of Large Numbers to show the convergence of the LHS (Left Hand Side) to a constant that we call the divergence rate. The random vectors $(\boldsymbol{M}_t^\star)_{t\in\mathbb{N}}$ are not i.i.d. so in order to apply a Law of Large Numbers on the RHS (Right Hand Side) of the previous equation we use Markov chain theory, more precisely the fact that $(\boldsymbol{M}_t^\star)_{t\in\mathbb{N}}$ is a function of a $(\delta_t, (\boldsymbol{M}_t^{i,j})_{i\in[1..\lambda],j\in\mathbb{N}})_{t\in\mathbb{N}}$ which is a geometrically ergodic Markov chain. As $(\boldsymbol{M}_t^{i,j})_{i\in[1..\lambda],j\in\mathbb{N},t\in\mathbb{N}}$ is a i.i.d. sequence, it is a Markov chain, and the sequence $(\delta_t)_{t\in\mathbb{N}}$ is also a Markov chain as stated in the following proposition.

Proposition 1. *Let a $(1,\lambda)$-ES with constant step-size optimize problem (2), handling the constraint through resampling, and take δ_t as defined in (5). Then no matter what distribution the i.i.d. sequence $(\boldsymbol{M}_t^{i,j})_{i\in[1..\lambda],(j,t)\in\mathbb{N}^2}$ have, $(\delta_t)_{t\in\mathbb{N}}$ is a homogeneous Markov chain and*

$$\delta_{t+1} = \delta_t - g(\boldsymbol{M}_t^\star) = \delta_t - \cos\theta[\boldsymbol{M}_t^\star]_1 - \sin\theta[\boldsymbol{M}_t^\star]_2 \ . \tag{13}$$

Proof. By definition in (5) and since for all t, $\sigma_t = \sigma$,

$$\begin{aligned}\delta_{t+1} &= -\frac{g(\boldsymbol{X}_{t+1})}{\sigma_{t+1}} \\ &= -\frac{g(\boldsymbol{X}_t) + \sigma g(\boldsymbol{M}_t^\star)}{\sigma} \\ &= \delta_t - g(\boldsymbol{M}_t^\star) \ ,\end{aligned}$$

and as shown in (7) the density of $\boldsymbol{M}_t^\star$ is determined by δ_t. So the distribution of δ_{t+1} is determined by δ_t, hence $(\delta_t)_{t\in\mathbb{N}}$ is a time-homogeneous Markov chain. □

We now show ergodicity of the Markov chain $(\delta_t)_{t\in\mathbb{N}}$, which implies that the t-steps transition kernel (the function $A \mapsto \Pr(\delta_t \in A | \delta_0 = \delta)$ for $A \in \mathcal{B}(\mathbb{R}_+)$) converges towards a stationary measure π, generalizing Propositions 3 and 4 of [8].

Proposition 2. *Let a $(1,\lambda)$-ES with constant step-size optimize problem (2), handling the constraint through resampling. We assume that the distribution of $\boldsymbol{M}_t^{i,j}$ is absolutely continuous with probability density function h, and that h is continuous and strictly positive on $\mathbb{R}^n$. Denote μ_+ the Lebesgue measure on $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$, and for $\alpha > 0$ take the functions $V : \delta \mapsto \delta$, $V_\alpha : \delta \mapsto \exp(\alpha\delta)$ and $r_1 : \delta \mapsto 1$. Then $(\delta_t)_{t\in\mathbb{N}}$ is μ_+-irreducible, aperiodic and compact sets are small sets for the Markov chain.*

If the following two additional conditions are fulfilled

$$\mathbf{E}(|g(\boldsymbol{M}_t^{i,j})| \ | \ \delta_t = \delta) < \infty \text{ for all } \delta \in \mathbb{R}_+ \ , \text{ and} \tag{14}$$

$$\lim_{\delta\to+\infty} \mathbf{E}(g(\boldsymbol{M}_t^\star)|\delta_t = \delta) \in \mathbb{R}_+^* \ , \tag{15}$$

then $(\delta_t)_{t\in\mathbb{N}}$ is r_1-ergodic and positive Harris recurrent with some invariant measure π.

Furthermore, if

$$\mathbf{E}(\exp(g(\boldsymbol{M}_t^{i,j}))|\delta_t = \delta) < \infty \text{ for all } \delta \in \mathbb{R}_+ \ , \tag{16}$$

then for $\alpha > 0$ small enough, $(\delta_t)_{t\in\mathbb{N}}$ is also $V_\alpha-$geometrically ergodic.

Proof. The probability transition kernel of $(\delta_t)_{t\in\mathbb{N}}$ writes

$$\begin{aligned}
P(\delta, A) &= \int_{\mathbb{R}^n} 1_A(\delta - g(\boldsymbol{x}))\tilde{h}^\star_\delta(\boldsymbol{x})d\boldsymbol{x} \\
&= \int_{\mathbb{R}^n} 1_A(\delta - g(\boldsymbol{x}))\lambda \frac{h(\boldsymbol{x})1_{L_\delta}(\boldsymbol{x})H((-\infty, [\boldsymbol{x}]_1)\times\mathbb{R}^{n-1}\cap L_\delta)^{\lambda-1}}{H(L_\delta)^\lambda} \\
&= \frac{\lambda}{H(L_\delta)^\lambda}\int_{g^{-1}(A)} h\begin{pmatrix}\delta - [\boldsymbol{u}]_1 \\ -[\boldsymbol{u}]_2 \\ \vdots \\ -[\boldsymbol{u}]_n\end{pmatrix} H((-\infty, \delta - [\boldsymbol{u}]_1)\times\mathbb{R}^{n-1}\cap L_\delta)^{\lambda-1}d\boldsymbol{u} \ ,
\end{aligned}$$

with the substitution of variables $[\boldsymbol{u}]_1 = \delta - [\boldsymbol{x}]_1$ and $[\boldsymbol{u}]_i = -[\boldsymbol{x}]_i$ for $i \in [2..n]$. Denote $L^\star_{\delta,v} := (-\infty, v)\times\mathbb{R}^{n-1}\cap L_\delta$ and $t_\delta : \boldsymbol{u} \mapsto (\delta - [\boldsymbol{u}]_1, -[\boldsymbol{u}]_2, \ldots, -[\boldsymbol{u}]_n)$, take C a compact of $\mathbb{R}_+$, and define ν_C such that for $A \in \mathcal{B}(\mathbb{R}_+)$

$$\nu_C(A) := \lambda\int_{g^{-1}(A)} \inf_{\delta\in C} \frac{h(t_\delta(\boldsymbol{u}))H(L^\star_{\delta,[\boldsymbol{u}]_1})^{\lambda-1}}{H(L_\delta)^\lambda}d\boldsymbol{u} \ .$$

As the density h is supposed to be strictly positive on $\mathbb{R}^n$, for all $\delta \in \mathbb{R}_+$ we have $H(L_\delta) \geq H(L_0) > 0$. Using the fact that H is a finite measure, and is absolutely continuous, applying the dominated convergence theorem shows that the functions $\delta \mapsto H(L_\delta)$ and $\delta \mapsto H((-\infty, \delta - [\boldsymbol{u}]_1)\times\mathbb{R}^{n-1}\cap L_\delta)$ are continuous. Therefore the function $\delta \mapsto h(t_\delta(\boldsymbol{u}))H(L^\star_{\delta,[\boldsymbol{u}]_1})^{\lambda-1}/H(L_\delta)^\lambda$ is continuous and C being a compact, the infimum of this function is reached on C is reached on C. Since this function is strictly positive, if $g^{-1}(A)$ has strictly positive Lebesgue measure then $\nu_C(A) > 0$ which proves that this measure is not trivial. By construction $P(\delta, A) \geq \nu_C(A)$ for all $\delta \in C$, so C is a small set which shows that compact sets are small. Since if $\mu_+(A) > 0$ we have $P(\delta, A) \geq \nu_C(A) > 0$, the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is μ_+-irreducible. Finally, if we take C a compact set of $\mathbb{R}_+$ with strictly positive Lebesgue measure, then it is a small set and $\nu_C(C) > 0$ which means the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is strongly aperiodic.

The function ΔV is defined as $\delta mapsto \mathbf{E}(V(\delta_{t+1})|\delta_t = \delta) - V(\delta)$. We want to show a drift condition (see [15]) on V. Using Eq. (13)

$$\begin{aligned}
\Delta V(\delta) &= \mathbf{E}(\delta - g(\boldsymbol{M}^\star_t)|\delta_t = \delta) - \delta) \\
&= -\mathbf{E}(g(\boldsymbol{M}^\star_t)) \ .
\end{aligned}$$

Therefore using the condition (15), we have that there exists a $\epsilon > 0$ and a $M \in \mathbb{R}_+$ such that $\forall\delta \in (M, +\infty)$, $\Delta V(\delta) \leq -\epsilon$. With condtion (14) implies that the function $\Delta V + \epsilon$ is bounded on the compact $[0, M]$ by a constant $b \in \mathbb{R}$. Hence for all $\delta \in \mathbb{R}_+$

$$\frac{\Delta V(\delta)}{\epsilon} \leq -1 + \frac{b}{\epsilon}1_{[0,M]}(\delta) \ . \tag{17}$$

For all $x \in \mathbb{R}$ the level set $C_{V,x}$ of the function V, $\{y \in \mathbb{R}_+|V(y) \leq x\}$, is equal to $[0, x]$ which is a compact set, hence a small set according to what we proved

earlier (and hence petite [15, Proposition 5.5.3]). Therefore V is unbounded off small sets and with (17) and Theorem 9.1.8 of [15], the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is Harris recurrent. The set $[0, M]$ is compact and therefore small and petite, so with (17), if we denote r_1 the constant function $\delta \in \mathbb{R}_+ \mapsto 1$ then with Theorem 14.0.1 of [15] the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is positive and is r_1-ergodic.

We now want to show a drift condition (see [15]) on V_α.

$$\begin{aligned}
\Delta V_\alpha(\delta) &= \mathbf{E}\left(\exp\left(\alpha\delta - \alpha g\left(\boldsymbol{M}_t^\star\right)\right)|\delta_t = \delta\right) - \exp\left(\alpha\delta\right) \\
\frac{\Delta V_\alpha}{V_\alpha}(\delta) &= \mathbf{E}\left(\exp\left(-\alpha g\left(\boldsymbol{M}_t^\star\right)\right)|\delta_t = \delta\right) - 1 \\
&= \int_{\mathbb{R}^n} \lim_{t\to+\infty} \sum_{k=0}^{t} \frac{(-\alpha g(\boldsymbol{x}))^k}{k!} \tilde{h}_\delta^\star(\boldsymbol{x}) d\boldsymbol{x} - 1 \ .
\end{aligned}$$

With Eq. (7) we see that $\tilde{h}_\delta^\star(\boldsymbol{x}) \leq \lambda h(\boldsymbol{x})/H(L_0)^\lambda$, so with our assumption that $\mathbf{E}(\exp\alpha|g(\boldsymbol{M}_t^{i,j})||\delta_t = \delta) < \infty$ for $\alpha > 0$ small enough we have that the function $\delta \mapsto \mathbf{E}(\exp(\alpha|g(\boldsymbol{M}_t^\star)||\delta_t = \delta)$ is bounded for the same α. As $\sum_{k=0}^{t}(-\alpha g(\boldsymbol{x}))^k/k!\tilde{h}_\delta^\star(\boldsymbol{x}) \leq \exp(\alpha|g(\boldsymbol{x})|)\tilde{h}_\delta^\star(\boldsymbol{x})$ which, with condition (16), is integrable so we may apply the theorem of dominated convergence to invert limit and integral:

$$\begin{aligned}
\frac{\Delta V_\alpha}{V_\alpha}(\delta) &= \lim_{t\to+\infty} \sum_{k=0}^{t} \int_{\mathbb{R}^n} \frac{(-\alpha g(\boldsymbol{x}))^k}{k!} \tilde{h}_\delta^\star(\boldsymbol{x}) d\boldsymbol{x} - 1 \\
&= \sum_{k\in\mathbb{N}} (-\alpha)^k \frac{\mathbf{E}\left(g\left(\boldsymbol{M}_t^\star\right)^k |\delta_t = \delta\right)}{k!} - 1
\end{aligned}$$

Since $\tilde{h}_\delta^\star(\boldsymbol{x}) \leq \lambda h(\boldsymbol{x})/H(L_0)^2$, $(-\alpha)^k\mathbf{E}(g(\boldsymbol{M}_t^\star)^k|\delta_t = \delta)/k! \leq (-\alpha)^k\mathbf{E}(g(\boldsymbol{M}_t^{i,j})^k)/k!$ which is integrable with respect to the counting measure so we may apply the dominated convergence theorem with the counting measure to invert limit and serie.

$$\begin{aligned}
\lim_{\delta\to+\infty} \frac{\Delta V_\alpha}{V_\alpha}(\delta) &= \sum_{k\in\mathbb{N}} \lim_{\delta\to+\infty} (-\alpha)^k \frac{\mathbf{E}\left(g\left(\boldsymbol{M}_t^\star\right)^k |\delta_t = \delta\right)}{k!} - 1 \\
&= -\alpha \lim_{\delta\to+\infty} \mathbf{E}\left(g\left(\boldsymbol{M}_t^\star\right)|\delta_t = \delta\right) + o\left(\alpha\right) \ .
\end{aligned}$$

With condition (17) we supposed that $\lim_{\delta\to+\infty}\mathbf{E}(g(\boldsymbol{M}_t^\star)|\delta_t = \delta) > 0$ this implies that for $\alpha > 0$ and small enough, $lim_{\delta\to+\infty}\Delta V_\alpha(\delta)/V_\alpha(\delta) < 0$, hence there exists $M \in \mathbb{R}_+$ and $epsilon > 0$ such that $\forall\delta > M$, $\Delta V_\alpha(\delta) < -\epsilon V_\alpha(\delta)$. Finally as $\Delta V_\alpha - V_\alpha$ is bounded on $[0, M]$ there exists $b \in \mathbb{R}$ such that

$$\Delta V_\alpha(\delta) \leq -\epsilon V_\alpha(\delta) + b1_{[0,M]}(\delta) \ .$$

According to what we did before in this proof, the compact set $[0, M]$ is small, and hence is petite ([15, Proposition 5.5.3]). So the μ_+-irreducible Markov chain

$(\delta_t)_{t\in\mathbb{N}}$ satisfies the conditions of Theorem 15.0.1 of [15] which with Theorem 14.0.1 of [15] proves that the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is V_α-geometrically ergodic. □

We now use a law of large numbers ([15] Theorem 17.0.1) on the Markov chain $(\delta_t, (\boldsymbol{M}_t^{i,j})_{i\in[1..\lambda],j\in\mathbb{N}})_{t\in\mathbb{N}}$ to obtain an almost sure divergence of the algorithm.

Proposition 3. *Let a $(1,\lambda)$-ES optimize problem (2), handling the constraint through resampling. Assume that the distribution H of the random step $\boldsymbol{M}_t^{i,j}$ is absolutely continuous with continuous and strictly positive density h, that conditions (16) and (15) of Proposition 2 hold, and denote π and μ_M the stationary distribution of respectively $(\delta_t)_{t\in\mathbb{N}}$ and $(\boldsymbol{M}_t^{i,j})_{i\in[1..\lambda],(j,t)\in\mathbb{N}^2}$. Then*

$$\frac{[\boldsymbol{X}_t - \boldsymbol{X}_0]_1}{t} \xrightarrow[t\to+\infty]{a.s.} \sigma\mathbf{E}_{\pi\times\mu_M}([\boldsymbol{M}_t^\star]_1) \ . \tag{18}$$

Furthermore if $\mathbf{E}([\boldsymbol{M}_t^\star]_2) < 0$, then the right hand side of Eq. (18) is strictly positive.

Proof. According to Proposition 2 the sequence $(\delta_t)_{t\in\mathbb{N}}$ is a Harris recurrent positive Markov chain with invariant measure π. As $(\boldsymbol{M}_t^{i,j})_{i\in[1..\lambda],(j,t)\in\mathbb{N}^2}$ is a i.i.d. sequence with distribution μ_M, $(\delta_t, (\boldsymbol{M}_t^{i,j})_{i\in[1..\lambda],j\in\mathbb{N}})_{t\in\mathbb{N}}$ is also a Harris recurrent positive Markov chain. As $[\boldsymbol{M}_t^\star]_1$ is a function of δ_t and $(\boldsymbol{M}_t^{i,j})_{i\in[1..\lambda],j\in\mathbb{N}}$, if $\mathbf{E}_{\pi\times\mu_M}(|[\boldsymbol{M}_t^\star]_1|) < \infty$, according to Theorem 17.0.1 of [15], we may apply a law of large numbers on the right hand side of Eq. (12) to obtain (18).

Using Fubini-Tonelli's theorem $\mathbf{E}_{\pi\times\mu_M}(|[\boldsymbol{M}_t^\star]_1|) = \mathbf{E}_\pi(\mathbf{E}_{\mu_M}(|[\boldsymbol{M}_t^\star]_1||\delta_t = \delta))$. From Eq. (7) for all $\boldsymbol{x}\in\mathbb{R}^n$, $\tilde{h}_\delta^\star(\boldsymbol{x}) \le \lambda h(\boldsymbol{x})/H(L_0)^2$, so the condition in (16) implies that for all $\delta\in\mathbb{R}_+$, $\mathbf{E}_{\mu_M}(|[\boldsymbol{M}_t^\star]_1||\delta_t = \delta)$ is finite. Furthermore, with condition (15), the function $\delta\in\mathbb{R}_+ \mapsto \mathbf{E}_{\mu_M}(|[\boldsymbol{M}_t^\star]_1||\delta_t = \delta)$ is bounded by some $M\in\mathbb{R}$. Therefore as π is a probability measure, $\mathbf{E}_\pi(\mathbf{E}_{\mu_M}(|[\boldsymbol{M}_t^\star]_1||\delta_t = \delta)) \le M < \infty$ so we may apply the law of large numbers of Theorem 17.0.1 of [15].

Using the fact that π is an invariant measure, we have $\mathbf{E}_\pi(\delta_t) = \mathbf{E}_\pi(\delta_{t+1})$, so $\mathbf{E}_\pi(\delta_t) = \mathbf{E}_\pi(\delta_t - \sigma g(\boldsymbol{M}_t^\star))$ and hence $\cos\theta\mathbf{E}_\pi([\boldsymbol{M}_t^\star]_1) = -\sin\theta\mathbf{E}_\pi([\boldsymbol{M}_t^\star]_2)$. So using the assumption that $\mathbf{E}([\boldsymbol{M}_t^{i,j}]_2) \le 0$ then we get the strict positivity of $\mathbf{E}_{\pi\times\mu_M}([\boldsymbol{M}_t^{i,j}]_1)$. □

5 Application to More Specific Distributions

Throughout this section we give cases where the assumptions on the distribution of the random steps H used in Proposition 2 or Proposition 3 are verified.

The following lemma shows an equivalence between a non-identity covariance matrix for H and a different norm and constraint angle θ.

Lemma 3. *Let a $(1,\lambda)$-ES optimize problem (2), handling the constraint with resampling. Assume that the distribution H of the random step $\boldsymbol{M}_t^{i,j}$ has positive definite covariance matrix $\mathbf{C}$ with eigenvalues $(\alpha_i^2)_{i\in[1..n]}$ and take $\boldsymbol{B} =$*

$(b_{i,j})_{(i,j)\in[1..n]^2}$ *such that* $\boldsymbol{B}\mathbf{C}\boldsymbol{B}^{-1}$ *is diagonal. Denote* $\mathcal{A}_{H,g,\boldsymbol{X}_0}$ *the sequence of parent points* $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ *of the algorithm with distribution* H *for the random steps* $\boldsymbol{M}_t^{i,j}$*, constraint angle* θ *and initial parent* $\boldsymbol{X}_0$*. Then for all* $k \in [1..n]$

$$\beta_k \left[\mathcal{A}_{H,\theta,\boldsymbol{X}_0}\right]_k \overset{(d)}{=} \left[\mathcal{A}_{\mathbf{C}^{-1/2}H,\theta',\boldsymbol{X}'_0}\right]_k \ , \tag{19}$$

where $\beta_k = \sqrt{\sum_{j=1}^n \frac{b_{j,i}^2}{\alpha_i^2}}$, $\theta' = \arccos(\frac{\beta_1 cos\theta}{\beta_g})$ *with* $\beta_g = \sqrt{\beta_1^2\cos^2\theta + \beta_2^2\sin^2\theta}$*, and* $[\boldsymbol{X}'_0]_k = \beta_k[\boldsymbol{X}_0]_k$ *for all* $k \in [1..n]$.

Proof. Take $(\bar{\boldsymbol{e}}_k)_{k\in[1..n]}$ the image of $(\boldsymbol{e}_k)_{k\in[1..n]}$ by $\boldsymbol{B}^{-1}$. We define a new norm $\|\cdot\|_-$ such that $\|\bar{\boldsymbol{e}}_k\|_- = 1/\alpha_k$. We define two orthonormal basis $(\boldsymbol{e}'_k)_{k\in[1..n]}$ and $(\bar{\boldsymbol{e}}'_k)_{k\in[1..n]}$ for $(\mathbb{R}^n, \|\cdot\|_-)$ by taking $\boldsymbol{e}'_k = \boldsymbol{e}_k/\|\boldsymbol{e}_k\|_-$ and $\bar{\boldsymbol{e}}'_k = \bar{\boldsymbol{e}}_k/\|\bar{\boldsymbol{e}}_k\|_- = \alpha_k\bar{\boldsymbol{e}}_k$. As $\mathrm{Var}(\boldsymbol{M}_t^{i,j}.\bar{\boldsymbol{e}}_k) = \alpha_k^2$, $\mathrm{Var}(\boldsymbol{M}_t^{i,j}.\bar{\boldsymbol{e}}'_k) = 1$ so in $(\mathbb{R}^n, \|\cdot\|_-)$ the covariance matrix of $\boldsymbol{M}_t^{i,j}$ is the identity.

Take h the function that to $\boldsymbol{x} \in \mathbb{R}^n$ maps its image in the new orthonormal basis $(\boldsymbol{e}'_k)_{k\in[1..n]}$. As $\boldsymbol{e}'_k = \boldsymbol{e}_k/\|\boldsymbol{e}_k\|_-$, $h(\boldsymbol{x}) = (\|\boldsymbol{e}_k\|_-[\boldsymbol{x}]_k)_{k\in[1..n]}$, where $\|\boldsymbol{e}_k\|_- = \|\sum_{i=1}^n b_{i,k}\bar{\boldsymbol{e}}_k\|_- = \sqrt{\sum_{i=1}^n b_{i,k}^2/\alpha_k^2} = \beta_k$. As we changed the norm, the angle between ∇f and ∇g is also different in the new space. Indeed $\cos\theta' = h(\nabla g).h(\nabla f)/(\|h(\nabla g)\|_-\|h(\nabla f)\|_-) = \beta_1^2\cos\theta/(\sqrt{\beta_1^2\cos^2\theta + \beta_2^2\sin^2\theta}\beta_1) = \beta_1\cos\theta/\beta_g$.

If we take $\boldsymbol{N}_t^{i,j} \sim \mathbf{C}^{-1/2}H$ then it has the same distribution as $h(\boldsymbol{M}_t^{i,j})$. Take $\boldsymbol{X}'_t = h(\boldsymbol{X}_t)$ then for a constraint angle $\theta' = \arccos(\beta_1\cos\theta/\beta_g)$ and a normalized distance to the constraint $\delta_t = \boldsymbol{X}'_t.h(\nabla g)/\sigma_t$ the ressampling is the same for $\boldsymbol{N}_t^{i,j}$ and $h(\boldsymbol{M}_t^{i,j})$ so $\boldsymbol{N}_t^i \overset{(d)}{=} h(\boldsymbol{M}_t^i)$. Finally the rankings induced by ∇f or $h(\nabla f)$ are the same so the selection in the same, hence $\boldsymbol{N}_t^\star \overset{(d)}{=} h(\boldsymbol{M}_t^\star)$, and therefore $\boldsymbol{X}'_{t+1} \overset{(d)}{=} h(\boldsymbol{X}_{t+1})$. □

Although Eq. (18) shows divergence of the algorithm, it is important that it diverges in the right direction, i.e. that the right hand side of Eq. (18) has a positive sign. This is achieved when the distribution of the random steps is isotropic, as stated in the following proposition.

Proposition 4. *Let a* $(1,\lambda)$*-ES optimize problem* (2) *with constant step-size, handling the constraint with resampling. Suppose that the Markov chain* $(\delta_t)_{t\in\mathbb{N}}$ *is positive Harris, that the distribution* H *of the random step* $\boldsymbol{M}_t^{i,j}$ *is absolutely continuous with strictly positive density* h*, and take* $\mathbf{C}$ *its covariance matrix. If the distribution* $\mathbf{C}^{-1/2}H$ *is isotropic then* $\mathbf{E}_{\pi\times\mu_M}([\boldsymbol{M}_t^\star]_1) > 0$.

Proof. First if $\mathbf{C} = \mathbf{I}_n$, using the same method than in the proof of Lemma 1

$$h^\star_{\delta,2}(y) = \lambda\int_{\mathbb{R}}\dots\int_{\mathbb{R}}\tilde{h}_\delta(u_1, y, u_3, \dots, u_n)\Pr(u_1 \geq [\boldsymbol{M}_t^i]_1)^{\lambda-1}du_1\prod_{k=3}^n du_k \ .$$

Using Eq.(6) and the fact that the condition $\boldsymbol{x} \in L_\delta$ is equivalent to $[\boldsymbol{x}]_1 \leq (\delta - [\boldsymbol{x}]_2 \sin\theta)/\cos\theta$ we obtain

$$h^\star_{\delta,2}(y) = \lambda \int_{\mathbb{R}} \dots \int_{-\infty}^{\frac{\delta - y\sin\theta}{\cos\theta}} \frac{h(u_1, y, u_3, \dots, u_n)}{H(L_\delta)} \Pr(u_1 \geq [\boldsymbol{M}^i_t]_1)^{\lambda-1} du_1 \prod_{k=3}^{n} du_k \ .$$

If the distribution of the random steps steps is isotropic then $h(u_1, y, u_3, \dots, u_n) = h(u_1, -y, u_3, \dots, u_n)$, and as the density h is supposed strictly positive, for $y > 0$ and all $\delta \in$, $h^\star_{\delta,2}(y) - h^\star_{\delta,2}(-y) < 0$ so $\mathbf{E}([\boldsymbol{M}^\star_t]_2 | \delta_t = \delta) < 0$. If the Markov chain is Harris recurrent and positive then this imply that $\mathbf{E}_\pi([\boldsymbol{M}^\star_t]_2) < 0$ and using the reasoning in the proof of Proposition 3 $\mathbf{E}_\pi([\boldsymbol{M}^\star_t]_1) > 0$.

For any covariance matrix $\mathbf{C}$ this result is generalized with the use of Lemma 3. □

Lemma 3 and Proposition 4 imply the following result to hold for multivariate normal distributions.

Proposition 5. *Let a $(1, \lambda)$-ES optimize problem (2) with constant step-size, handling the constraint with resampling. If H is a multivariate normal distribution with mean $\mathbf{0}$, then $(\delta_t)_{t\in\mathbb{N}}$ is a geometrically ergodic positive Harris Markov chain, Eq. (18) holds and its right hand side is strictly positive.*

Proof. Suppose $\boldsymbol{M}^{i,j}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$. Then H is absolutely continuous and h is strictly positive. The function $\boldsymbol{x} \mapsto \exp(g(\boldsymbol{x}))\exp(-\|\boldsymbol{x}\|^2/2)/\sqrt{2\pi}$ is integrable, so Eq. (16) is satisfied. Furthermore, when $\delta \to +\infty$ the constraint disappear so $\boldsymbol{M}^{i,j}_t$ behaves like $(\mathcal{N}_{\lambda:\lambda}, \mathcal{N}(0,1), \dots, \mathcal{N}(0,1))$ where $\mathcal{N}_{\lambda:\lambda}$ is the last order statistic of λ i.i.d. standard normal variables, so using that $\mathbf{E}(\mathcal{N}_{\lambda:\lambda}) > 0$ and $\mathbf{E}(\mathcal{N}(0,1)) = 0$, with multiple uses of the dominated convergence theorem we obtain condition (15) so with Proposition 2 the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is geometrically ergodic and positive Harris.

Finally H being isotropic the conditions of Proposition 4 are fulfilled, and therefore so are every condition of Proposition 3 which shows what we wanted. □

To obtain sufficient conditions for the density of the random steps to be strictly positive, it is advantageous to decompose that distribution into its marginals and the copula combining them. We pay a particular attention to *Archimedean copulas,* i.e., copulas defined

$$(\forall \boldsymbol{u} \in [0,1]^n)\ C_\psi(\boldsymbol{u}) = \psi(\psi^{-1}([\boldsymbol{u}]_1) + \dots + \psi^{-1}([\boldsymbol{u}]_n)), \tag{20}$$

where $\psi : [0, +\infty] \to [0, 1]$ is an *Archimedean generator*, i.e., $\psi(0) = 1, \psi(+\infty) = \lim_{t\to+\infty} \psi(t) = 0$, ψ is continuous and strictly decreasing on $[0, \inf\{t : \psi(t) = 0\})$, and ψ^{-1} denotes the generalized inverse of ψ,

$$(\forall u \in [0,1])\ \psi^{-1}(u) = \inf\{t \in [0, +\infty] : \psi(t) = u\}. \tag{21}$$

The reason for our interest is that Archimedean copulas are invariant with respect to permutations of variables, i.e.,

$$(\forall \boldsymbol{u} \in [0,1]^n)\ C_\psi(\mathbf{Q}\boldsymbol{u}) = C_\psi(\boldsymbol{u}). \tag{22}$$

holds for any permutation matrix $\mathbf{Q} \in \mathbb{R}^{n,n}$. This can be seen as a weak form of isotropy because in the case of isotropy, (20) holds for any rotation matrix, and a permutation matrix is a specific rotation matrix.

Proposition 6. *Let H be the distribution of the two first dimensions of the random step $\boldsymbol{M}_t^{i,j}$, H_1 and H_2 be its marginals, and C be the copula relating H to H_1 and H_2. Then the following holds:*

1. *Sufficient for H to have a continuous strictly positive density is the simultaneous validity of the following two conditions.*
   (i) *H_1 and H_2 have continuous strictly positive densities h_1 and h_2, respectively.*
   (ii) *C has a continuous strictly positive density c.*
   Moreover, if (i) and (ii) are valid, then

$$(\forall \boldsymbol{x} \in \mathbb{R}^2)\ h(\boldsymbol{x}) = c(H_1([\boldsymbol{x}]_1), H_2([\boldsymbol{x}]_2))h_1([\boldsymbol{x}]_1)h_2([\boldsymbol{x}]_2). \tag{23}$$

2. *If C is Archimedean with generator ψ, then it is sufficient to replace (ii) with*
   (ii') ψ is at least 4-monotone, i.e., ψ is continuous on $[0,+\infty]$, ψ'' is decreasing and convex on $\mathbb{R}_+$, and $(\forall t \in \mathbb{R}_+)\ (-1)^k\psi^{(k)}(t) \geq 0, k = 0,1,2$.
   In this case, if (i) and (ii') are valid, then

$$(\forall \boldsymbol{x} \in \mathbb{R}^2)\ h(\boldsymbol{x}) = \frac{\psi''(\psi^{-1}(H_1([\boldsymbol{x}]_1)) + \psi^{-1}(H_2([\boldsymbol{x}]_2)))}{\psi'(\psi^{-1}(H_1([\boldsymbol{x}]_1)) + \psi^{-1}(H_2([\boldsymbol{x}]_2)))} h_1([\boldsymbol{x}]_1)h_2([\boldsymbol{x}]_2). \tag{24}$$

6 Discussion

The paper presents a generalization of recent results of the first author [8] concerning linear optimization by a $(1,\lambda)$-ES in the constant step size case. The generalization consists in replacing the assumption of normality of random steps involved in the evolution strategy by substantially more general distributional assumptions. This generalization shows that isotropic distributions solve the linear problem. Also, although the conditions for the ergodicity of the studied Markov chain accept some heavy-tail distributions, an expnentially vanishing tail allow for geometric ergodicity, which imply a faster convergence to its stationary distribution, and faster convergence of Monte Carlo simulations. In our opinion, these conditions increase the insight into the role that different kinds of distributions play in evolutionary computation, and enlarges the spectrum of possibilities for designing evolutionary algorithms with solid theoretical fundamentals. At the same time, applying the decomposition of a multidimensional distribution into its marginals and the copula combining them, the paper attempts to bring a

small contribution to the research into applicability of copulas in evolutionary computation, complementing the more common application of copulas to the Estimation of Distribution Algorithms [12, 14, 13].

Needless to say, more realistic than the constant step size case, but also more difficult to investigate, is the varying step size case. The most important results in [8] actually concern that case. A generalization of those results for non-Gaussian distributions of random steps for cumulative step-size adaptation ([9]) is especially difficult as the evolution path is tailored for Gaussian steps, and some careful tweaking would have to be applied. The σ self-adaptation evolution strategy ([16]), studied in [6] for the same problem, appears easier, and would be our direction for future research.

Acknowledgment

The research reported in this paper has been supported by grant ANR-2010-COSI-002 (SIMINOLE) of the French National Research Agency, and Czech Science Foundation (GAČR) grant 13-17187S.

References

1. X. Yao and Y. Liu, "Fast evolution strategies," in *Evolutionary Programming VI*, pp. 149–161, Springer, 1997.
2. T. Schaul, "Benchmarking Separable Natural Evolution Strategies on the Noiseless and Noisy Black-box Optimization Testbeds," in *Black-box Optimization Benchmarking Workshop, Genetic and Evolutionary Computation Conference*, (Philadelphia, PA), 2012.
3. T. Schaul, T. Glasmachers, and J. Schmidhuber, "High dimensions and heavy tails for natural evolution strategies," in *Genetic and Evolutionary Computation Conference (GECCO)*, 2011.
4. N. Hansen, F. Gemperle, A. Auger, and P. Koumoutsakos, "When do heavy-tail distributions help?," in *Parallel Problem Solving from Nature PPSN IX* (T. P. Runarsson *et al.*, eds.), vol. 4193 of *Lecture Notes in Computer Science*, pp. 62–71, Springer, 2006.
5. D. Arnold, "On the behaviour of the (1,λ)-ES for a simple constrained problem," in *Foundations of Genetic Algorithms - FOGA 11*, pp. 15–24, ACM, 2011.
6. D. Arnold, "On the behaviour of the $(1, \lambda)$-σSA-ES for a constrained linear problem," in *Parallel Problem Solving from Nature - PPSN XII*, pp. 82–91, Springer, 2012.
7. A. Chotard, A. Auger, and N. Hansen, "Cumulative step-size adaptation on linear functions," in *Parallel Problem Solving from Nature - PPSN XII*, pp. 72–81, Springer, september 2012.
8. A. Chotard, A. Auger, and N. Hansen, "Markov chain analysis of evolution strategies on a linear constraint optimization problem," in *2014 IEEE Congress on Evolutionary Computation (CEC)*.
9. N. Hansen and A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies," *Evolutionary Computation*, vol. 9, no. 2, pp. 159–195, 2001.

10. C. A. Coello Coello, "Constraint-handling techniques used with evolutionary algorithms," in *Proceedings of the 2008 GECCO conference companion on Genetic and evolutionary computation*, GECCO '08, (New York, NY, USA), pp. 2445–2466, ACM, 2008.
11. D. Arnold and D. Brauer, "On the behaviour of the $(1 + 1)$-ES for a simple constrained problem," in *Parallel Problem Solving from Nature - PPSN X* (I. G. R. et al., ed.), pp. 1–10, Springer, 2008.
12. A. Cuesta-Infante, R. Santana, J. Hidalgo, C. Bielza, and P. Larrañaga, "Bivariate empirical and n-variate archimedean copulas in estimation of distribution algorithms," in *IEEE Congress on Evolutionary Computation*, pp. 1–8, 2010.
13. L. Wang, X. Guo, J. Zeng, and Y. Hong, "Copula estimation of distribution algorithms based on exchangeable archimedean copula," *International Journal of Computer Applications in Technology*, vol. 43, pp. 13–20, 2012.
14. R. Salinas-Gutierrez, A. Hernández Aguirre, and E. Villa Diharce, "Using copulas in estimation of distribution algorithms," in *MICAI 2009: Advances in Artificial Intelligence*, pp. 658–668, 2009.
15. S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*. Cambridge University Press, second ed., 1993.
16. H.-G. Beyer, "Toward a theory of evolution strategies: Self-adaptation," *Evolutionary Computation*, vol. 3, no. 3, pp. 311–347, 1995.