Cumulative Step-size Adaptation on Linear Functions: Technical Report

Alexandre Chotard, Anne Auger, Nikolaus Hansen

Cumulative Step-size Adaptation on Linear Functions: Technical Report

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

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

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

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

1 Introduction

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

□

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

Lemma 2 ([12]). *Let* g *be a real valued function on* $\mathbb{R}$. *For* $\lambda \geq 2$,

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

Proof. of Lemma 2

This method can be found with more details in [12].

Let $\chi_i = g\left(\xi_i\right)$, and $\chi_{i:\lambda} = g\left(\xi_{i:\lambda}\right)$. Note that in general $\chi_{1:\lambda} \neq \min_{i\in[[1,\lambda]]}\chi_i$. The sorting is made on (ξ_i), not on (χ_i).

We will also note $\chi_{i:\lambda}^{\{j\}}$ the i^th order statistic after that the variable χ_j has been taken away. $\chi_{i:\lambda}^{[j]}$ will be i^th order statistic after $\chi_{j:\lambda}$ has been taken away : if $i \neq 1$ then we have $\chi_{1:\lambda}^{[i]} = \chi_{1:\lambda}$, and for $i = 1$ $\chi_{1:\lambda}^{[i]} = \chi_{2:\lambda}$.

Then we have $\mathbb{E}\left(\chi_{1:\lambda}^{\{i\}}\right) = \chi_{1:\lambda-1}$,

And $\sum_{i=1}^{\lambda} \chi_{1:\lambda}^{\{i\}} = \sum_{i=1}^{\lambda} \chi_{1:\lambda}^{[i]}$ (2).

From the first equation we deduce that $\lambda\mathbb{E}\left(\chi_{1:\lambda-1}\right) = \lambda\mathbb{E}(\chi_{1:\lambda}^{\{i\}}) = \sum_{i=1}^{\lambda} \mathbb{E}(\chi_{1:\lambda}^{\{i\}}) = \mathbb{E}(\sum_{i=1}^{\lambda} \chi_{1:\lambda}^{\{i\}})$.

With the second equation, we get that $\mathbb{E}(\sum_{i=1}^{\lambda} \chi_{1:\lambda}^{\{i\}}) = \mathbb{E}(\sum_{i=1}^{\lambda} \chi_{1\lambda}^{[i]}) = \mathbb{E}(\chi_{2:\lambda}) + (\lambda - 1)\mathbb{E}(\chi_{1:\lambda})$.

By combining both, we get the final equation:

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

□

We are now ready to prove the following result.

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

Proof. of Lemma 3 The strict monotony of $\mathbb{E}(\mathcal{N}_{1:\lambda}^2)$ in λ from the previous proposition is equivalent to show that $\mathbb{E}(\mathcal{N}_{1:\lambda}^2) > \mathbb{E}\left(\mathcal{N}_{2:\lambda}^2\right)$ for $\lambda \geq 3$. Indeed $\mathbb{E}(\mathcal{N}_{1:\lambda}^2) - \mathbb{E}(\mathcal{N}_{1:\lambda-1}^2) = \mathbb{E}(\mathcal{N}_{1:\lambda}^2 - \mathcal{N}_{2:\lambda}^2)/\lambda$ which follows from Lemma 2 taking g as the square function.

Let $E_1 = \{\omega \in \Omega | \mathcal{N}_{1:\lambda}^2(\omega) < \mathcal{N}_{2:\lambda}^2(\omega)\}$, where $\Omega = \mathbb{R}^\lambda$ and $P(\omega) = \exp(-\|\omega\|^2/2)/\sqrt{2\pi}$. For $\omega \in \Omega$, let us note $\omega_{i:\lambda}$ the i^{th} order statistic of $([\omega]_j)_{j\in[[1,\lambda]]}$. Let g be a function that maps $\omega \in \Omega$ to $\tilde{\omega} \in \Omega$, where $\tilde{\omega}_{1:\lambda} = -\omega_{2:\lambda}$, $\tilde{\omega}_{2:\lambda} = \omega_{1:\lambda}$ and for $i \geq 3$, $\tilde{\omega}_{i:\lambda} = \omega_{i:\lambda}$. The function g is bijective between E_1 and its image by g, E_2. Let us note that for $\omega \in E_1$, $\mathcal{N}_{2:\lambda}^2(\omega) - \mathcal{N}_{1:\lambda}^2(\omega) = \mathcal{N}_{1:\lambda}^2(g(\omega)) - \mathcal{N}_{2:\lambda}^2(g(\omega))$, and $P(\omega) = P(g(\omega))$ since the standard normal distribution is symmetric. That is $\int_{E_1}(\mathcal{N}_{2:\lambda}^2(\omega) - \mathcal{N}_{1:\lambda}^2(\omega))P(\omega)d\omega = \int_{E_2}(\mathcal{N}_{1:\lambda}^2(\tilde{\omega}) - \mathcal{N}_{2:\lambda}^2(\tilde{\omega}))P(\tilde{\omega})d\tilde{\omega}$ by a change of variables $\tilde{omega} = g(\omega)$. As according to the definition of E_1, for all $\omega \in \Omega \backslash E_1$ $\mathcal{N}_{1:\lambda}^2(\omega) \geq \mathcal{N}_{2:\lambda}^2(\omega)$, and that E_1 is properly counterweighted by E_2 in the expected value of $\mathcal{N}_{1:\lambda}^2 - \mathcal{N}_{2:\lambda}^2$, we do have $\mathbb{E}(\mathcal{N}_{1:\lambda}^2) \geq \mathbb{E}(\mathcal{N}_{2:\lambda}^2)$ for all $\lambda \geq 2$.

For $\lambda \geq 3$, let $E_3 = \{\omega \in \Omega | \omega_{3:\lambda} \in]-|\omega_{1:\lambda}|, |\omega_{1:\lambda}|[\text{ and } \omega_{1:\lambda} < \omega_{2:\lambda}\}$. Then, for $\omega \in E_3$ we also have $\omega_{2:\lambda} \in]-|\omega_{1:\lambda}|, |\omega_{1:\lambda}|[$, so $\mathcal{N}_{1:\lambda}^2(\omega) > \mathcal{N}_{2:\lambda}^2(\omega)$ which means $\omega \notin E_1$, or $E_1 \cap E_3 = \emptyset$. For $\omega \in E_1$, as $\omega_{1:\lambda}^2 < \omega_{2:\lambda}^2$ and $\omega_{1:\lambda} < \omega 2 : \lambda$, $\omega_{2:\lambda} > 0$, so $\omega_{3:\lambda} \notin [-\omega 2 : \lambda, \omega 2 : \lambda]$. Hence, as $g(\omega)_{3:\lambda} = \omega 3 : \lambda$ and $[-\omega 2 : \lambda, \omega 2 : \lambda] = [-|g(\omega)_{1:\lambda})|, |g(\omega)_{1:\lambda})|]$, $g(\omega) \notin E_3$. That is $E_2 \cap E_3 = \emptyset$. So E_3 is disjoint with E_1 and E_2. Furthermore, for every $\omega \in \Omega$, except when $\omega_{1:\lambda} \neq 0$ which is a negligible subset of events, there exists a non negligible set of $(\omega_{i:\lambda})_{i\in[[1,\lambda]]}$ such that $\omega_{3:\lambda} \in]-|\omega_{1:\lambda}|, |\omega_{1:\lambda}|[\text{ and } \omega_{1:\lambda} < \omega_{2:\lambda}\}$. So E_3 is a non negligible subset of Ω, where $\mathcal{N}_{1:\lambda}^2(\omega) > \mathcal{N}_{2:\lambda}^2(\omega)$. Hence $\mathbb{E}(\mathcal{N}_{1:\lambda}^2(\omega)) > \mathbb{E}(\mathcal{N}_{2:\lambda}^2(\omega))$, which is the monotony of Lemma 3.

For $\lambda = 1$, $\mathcal{N}_{1:1} \sim \mathcal{N}(0,1)$ so $\mathbb{E}(\mathcal{N}_{1:1}^2) = 1$. For $\lambda = 2$ we have $\mathbb{E}(\mathcal{N}_{1:2}^2 + \mathcal{N}_{2:2}^2) = 2\mathbb{E}(\mathcal{N}(0,1)^2) = 2$, and since the normal distribution is symmetric $\mathbb{E}(\mathcal{N}_{1:2}^2) = \mathbb{E}(\mathcal{N}_{2:2}^2)$, hence $\mathbb{E}(\mathcal{N}_{1:2}^2) = 1$. □

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

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

$$\frac{1}{t} \ln \left(\frac{\sigma_t}{\sigma_0} \right) \xrightarrow[t \to \infty]{a.s.} \mathbb{E} \left(\ln \left(\frac{\sigma_{t+1}}{\sigma_t} \right) \right) = \frac{1}{2d_\sigma n} \left(\mathbb{E} \left(\mathcal{N}_{1:\lambda}^2 \right) - 1 \right) \quad . \tag{7}$$

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

Proof. For $\lambda > 2$, from Lemma 3 we know that $\mathbb{E}(\mathcal{N}_{1:\lambda}^2) > \mathbb{E}(\mathcal{N}_{1:2}^2) = 1$. Therefore $\mathbb{E}(\mathcal{N}_{1:\lambda}^2) - 1 > 0$, hence Eq. (7) is strictly positive, and with Proposition 1 we get that the step-size diverges geometrically almost surely at the rate $1/(2d_\sigma)(\mathbb{E}(\mathcal{N}_{1:\lambda}^2) - 1)$.

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

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

As the selection occurs only on the first dimension, if there is geometric divergence for $\boldsymbol{X}_t$, it is on $[\boldsymbol{X}_t]_1$. From Eq (1)

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

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

$$\frac{1}{t} \ln \left| \frac{[\boldsymbol{X}_t]_1}{[\boldsymbol{X}_0]_1} \right| = \frac{1}{t} \sum_{k=0}^{t-1} \ln \left| 1 + \frac{\sigma_k}{[\boldsymbol{X}_k]_1} [\boldsymbol{\xi}_t^\star]_1 \right| \quad . \tag{8}$$

Although it is not obvious at first sight, it is important to take the logarithm, as we intuitively know that the speed of σ_t and the speed of $\boldsymbol{X}_t$ are connected. The divergence rate of σ_t being log-linear, so should be the one of $\boldsymbol{X}_t$. Let $Z_{-1} = 0$, and $Z_t = \frac{[\boldsymbol{X}_{t+1}]_1 - [\boldsymbol{X}_0]_1}{\sigma_t}$ for $t \geq 0$.

$$Z_{t+1} = \frac{[\boldsymbol{X}_{t+2}]_1 - [\boldsymbol{X}_0]_1}{\sigma_{t+1}} = \frac{[\boldsymbol{X}_{t+1}]_1 - [\boldsymbol{X}_0]_1 + \sigma_{t+1} \left[\boldsymbol{\xi}_{t+1}^\star \right]_1}{\sigma_{t+1}}$$

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

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

By introducing $\mathcal{Z}$ in Eq (8), we obtain:

$$\begin{aligned}
\frac{1}{t}\ln\left|\frac{[\boldsymbol{X}_t]_1}{[\boldsymbol{X}_0]_1}\right| &= \frac{1}{t}\sum_{k=0}^{t-1}\ln\left|1+\frac{\sigma_{k-1}\eta^\star_{k-1}}{[\boldsymbol{X}_k]_1}[\boldsymbol{\xi}^\star_k]_1\right| \\
&= \frac{1}{t}\sum_{k=0}^{t-1}\ln\left|1+\frac{\eta^\star_{k-1}}{Z_{k-1}}[\boldsymbol{\xi}^\star_k]_1\right| \\
&= \frac{1}{t}\sum_{k=0}^{t-1}\ln\left|\frac{\frac{Z_{k-1}}{\eta^\star_{k-1}}+\boldsymbol{\xi}^\star_k}{\frac{Z_{k-1}}{\eta^\star_{k-1}}}\right| \\
&= \frac{1}{t}\sum_{k=0}^{t-1}\left(\ln|Z_k|-\ln|Z_{k-1}|+\ln\left|\eta^\star_{k-1}\right|\right) \qquad (9)
\end{aligned}$$

The right hand side of this equation reminds us again of the law of large numbers. There is no independence over time, but $\mathcal{Z}$ being a Markov chain, if it follows some specific stability properties of Markov chains, then a law of large numbers may apply.

Study of the Markov chain $\mathcal{Z}$ To apply a law of large numbers to a Markov chain, it has to satisfies some stability properties: in particular, the Markov chain $\boldsymbol{\mathcal{P}}$ has to be φ-irreducible, that is, there exists a measure φ such that every Borel set A of $\mathbb{R}^n$ with $\varphi(A) > 0$ has a positive probability to be reached in a finite number of steps by $\boldsymbol{\mathcal{P}}$ starting from any $\boldsymbol{p}_0 \in \mathbb{R}^n$. In addition, the chain $\boldsymbol{\mathcal{P}}$ needs to be (i) positive, that is the chain admits an invariant probability measure π, i.e., for any borelian A, $\pi(A) = \int_{\mathbb{R}^n} P(x, A)\pi(dx)$ with $P(x, A)$ being the probability to transition in one time step from x into A, and (ii) Harris recurrent which means for any borelian A such that $\varphi(A) > 0$, the chain $\boldsymbol{\mathcal{P}}$ visits A an infinite number of times with probability one. Under those conditions, $\boldsymbol{\mathcal{P}}$ satisfies a law of large numbers, more precisely:

Lemma 4. *[11, 17.0.1] Suppose that $\boldsymbol{\Phi}$ is a positive Harris chain with stationary measure π, and let g be a π-integrable function that is such that $\pi(|g|) = \int_{\mathbb{R}^n} |g(x)|\pi(dx) < \infty$. Then*

$$1/t\sum_{k=1}^{t} g(\boldsymbol{\Phi}_k) \xrightarrow[t\to\infty]{a.s} \pi(g) \ . \qquad (10)$$

To show that a Markov defined in a space X is positive Harris recurrent, we generally show that the chain follows a so-called drift condition over a small set, that is for a function V, an inequality over the drift operator $\Delta V : x \mapsto \int_X V(y)P(x, dy) - V(x)$. A small set is a borel set such that there exists a $m \in \mathbb{N}^*$ and a non-trivial measure ν_m on $\beta(X)$ such that for all $x \in C$, $B \in \beta(X)$, $P^m(x, B) \geq \nu_m(B)$. The set C is then called a ν_m-small set. The chain also needs to be aperiodic, that is there is no d-cycle, that is disjoint Borel sets $(D_i)_{i\in[[1,d]]}$, such that for $x \in D_i$, $P(x, D_{i+1}) = 1$ for $i = 0 \cdots d-1(modd)$, and $[\cup_{i=1}^d]^c$ is φ-negligible. If there exists a ν_1-small-set A such that $\nu_1(A) > 0$, then the chain is strongly aperiodic (and therefore aperiodic). We then have the following lemma.

Lemma 5. *[11, 14.0.1] Suppose that the chain $\boldsymbol{\Phi}$ is φ-irreductible and aperiodic, and $f \geq 1$ a function on X. Let us assume that there exists V some extended-valued non-negative function finite for some $x_0 \in X$, a small set C and $b \in \mathbb{R}$ such that*

$$\Delta V(x) \leq -f(x) + b\mathbf{1}_C(x) \ \ , x \in X. \tag{11}$$

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

$$\pi(f) = \int_X \pi(dx) f(x) < \infty \ . \tag{12}$$

To prove the irreducibility, aperiodicity and to exhibit the small sets of the Markov chain $\mathcal{Z}$ through its transition kernel would be difficult. Instead, it can be done by showing some properties of its underlying control model. In our case, the model associated to $\mathcal{Z}$ is called a non-linear state space model. We will, in the following, define this non-linear state space model and some of its properties.

Suppose $\boldsymbol{X} = \{\boldsymbol{X}_k\}$, $\boldsymbol{X}_k \in \mathcal{X}$. If there is a smooth function (C^∞) F such that $\boldsymbol{X}_{k+1} = F(\boldsymbol{X}_k, \boldsymbol{W}_{k+1})$ with $(\boldsymbol{W}_i)_{i\in\mathbb{N}}$ being a sequence of i.i.d. random variables, whose marginal distribution Γ possesses a semi lower-continuous density γ_w which is supported on an open set O_w; then $\boldsymbol{X}$ is called a non-linear state space model driven by F or NSS(F) model, with control set O_w.

We define its associated control model CM(F) the deterministic system $x_k = F_k(x_0, u_1, \cdots, u_k)$, where F_k is given by $F_k(x_0, u_1, \cdots, u_k) = F(F_{k-1}(x_0, u_1, \cdots, u_{k-1}), u_k)$, and $F_0(x_0) = x_0$, provided that $(u_i)_{i\in\mathbb{N}}$ lies in the control set O_w.

For a point $\boldsymbol{x} \in \mathcal{X}$, and $k \in \mathbb{N}$ we define $A_+^k(\boldsymbol{x}) = \{F_k(\boldsymbol{x}, u_1, \cdots, u_k) | u_i \in O_w \ \ \forall i \in \mathbb{N}\}$, the set of points reachable from $\boldsymbol{x}$ after k steps of time. And $A_+(\boldsymbol{x}) = \bigcup_{i\in\mathbb{N}} A_+^i(\boldsymbol{x})$.

The associated control model CM(F) is called forward accessible if for each $\boldsymbol{x}_0 \in \mathcal{X}$, the set $A_+(\boldsymbol{x_0})$ has non empty-interior.

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

Lemma 6. *[11, 7.2.4, 7.3.5] Let $M \subset \mathcal{X}$ be a minimal set for CM(F). If CM(F) is forward accessible then the NSS(F) model restricted to M is an open set irreducible T-chain.*

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

We can now prove the following lemma:

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

Proof. This is exactly the result of Theorem 6 when all conditions are fulfilled. We then have to show the right properties of the underlying control model.

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

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

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

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

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

$$\mathbb{E}\left(\eta_\cdot^{\star -\alpha}\right) - 1 < 0 \ . \tag{13}$$

Proof. Using the Taylor series of the exponential function we have

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

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

We are now ready to prove the following lemma:

Lemma 9. *The Markov chain* $\mathcal{Z}$ *is Harris recurrent positive, and admits a unique invariant measure* μ *such that for* $f\colon x \mapsto |x|^\alpha \in \mathbb{R}$, $\mu(f) = \int_{\mathbb{R}} \mu(dx) f(x) < \infty$, *with* α *such that Eq.* (13) *holds true.*

Proof. By using Lemma 7 and Lemma 5, we just need the drift condition (11) to prove Lemma 9. Let V be such that for $x \in \mathbb{R}$, $V(x) = |x|^\alpha + 1$.

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

We take α such that Eq. (13) holds true (as according to Lemma 8, there exists such a α). As $\mathbb{E}(\eta_\cdot^{\star -\alpha} - 1) < 0$, there exists $\epsilon > 0$ and $M > 0$ such that for all $|x| \geq M$, $\Delta V/V(x) \leq -\epsilon$. Let b be equal to $\mathbb{E}([\boldsymbol{\xi}_t^\star]_1^2) + \epsilon V(M)$. Then for all $|x| \leq M$, $\Delta V(x) \leq -\epsilon V(x) + b$. Therefore, if we note $C = [-M, M]$, which is according to Lemma 7 a small-set, we do have $\Delta V(x) \leq -\epsilon V(x) + b\mathbf{1}_C(x)$ which is Eq. (11) with $f = \epsilon V$. Therefore from Lemma 5 the chain $\mathcal{Z}$ is positive Harris recurrent with invariant probability measure μ, and ϵV is μ-integrable. As $\int_{\mathbb{R}} \mu(dx)|x|^\alpha = 1/\epsilon \int_{\mathbb{R}} \mu(dx)\epsilon V(x) - 1 < \infty$, the function $x \mapsto |x|^\alpha$ is also μ-integrable.

□

In order to use Lemma 4 on $\mathcal{Z}$ with the function $g : x \mapsto \mathbb{E}\left(\ln |x/x - [\boldsymbol{\xi}_\cdot^\star]_1|\right)$, we must prove that this function is μ-integrable, that is $\int_{\mathbb{R}} g(u)\mu(du) < \infty$. To do so we will need the following lemma on the existence of moments for stationary Markov chains:

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

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

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

.

We may now prove the following theorem:

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

$$\frac{1}{t} \ln \left| \frac{[\boldsymbol{X}_t]_1}{[\boldsymbol{X}_0]_1} \right| \xrightarrow[t\to\infty]{a.s} \frac{1}{2d_\sigma n} \left(\mathbb{E}\left(\mathcal{N}^2_{1:\lambda}\right) - 1 \right) \quad . \tag{14}$$

Proof. We will first prove here that the function $g : x \mapsto \ln |x|$ is μ-integrable. From Lemma 9 we know that the function $f : x \mapsto |x|^\alpha$ is μ-integrable, and as for any $M > 0$, and any $x \in [-M, M]^c$ there exists $K > 0$ such that $K|x|^\alpha > |\ln |x||$, then $g\mathbf{1}_{A^c}$ is μ-integrable, with $A = [-M, M]$. So what is left is to prove that $g\mathbf{1}_A$ is also μ-integrable. We will now check the conditions to use Lemma 10.

According to Lemma 7 the chain $\mathcal{Z}$ is open-set irreducible, so $\mu(A^c) > 0$. For $C > 0$, if we take $h : z \mapsto C/\sqrt{|z|}$, with M small enough we do have for all $z \in A^c$, $h(z) \geq |g(z)|$. Furthermore, if we study the inequality

$$\int_{A^c} P(z, dy)h(y) \leq h(z) - |g(z)|$$

$$\int_S P\left(\frac{z}{\eta^\star_\cdot} + [\boldsymbol{\xi}^\star_\cdot]_1 \in dy\right) \mathbf{1}_{A^c}(y) \frac{C}{\sqrt{|y|}} \leq \frac{C}{\sqrt{|z|}} - |ln|z||$$

$$\mathbb{E}\left(\frac{1}{\sqrt{\left|\frac{z}{\eta^\star_\cdot} + [\boldsymbol{\xi}^\star_\cdot]_1\right|}} \mathbf{1}_{A^c}\left(\frac{z}{\eta^\star_\cdot} + [\boldsymbol{\xi}^\star_\cdot]_1\right) \right) \leq \frac{1}{\sqrt{|z|}} - \frac{|\ln |z||}{C}$$

We can increase C up until $|\ln |z||/C$ is negligible compared to $1/\sqrt{|z|}$, and we can decrease M to make $\mathbb{E}\left(1/\sqrt{|z/\eta^\star_\cdot + [\boldsymbol{\xi}^\star_\cdot]_1|}\mathbf{1}_{A^c}\left(z/\eta^\star_\cdot + [\boldsymbol{\xi}^\star_\cdot]_1\right)\right)$ as small as we would like it to be, as it decreases the size of A^c, so the inequality holds if we choose M and $1/C$ small enough. The second inequality for Lemma 10 holds as well:

$$\int_{A^c} P(u, dv)h(v) \leq \int_{A^c} \frac{C}{\sqrt{|v|}} dv = 4C\sqrt{M} < \infty$$

Finally, according to Lemma 9, the chain $\mathcal{Z}$ is Harris recurrent. So Lemma 10 shows that g is μ-integrable. This allows us to apply Lemma 4 to the function g: $1/t \sum_{k=1}^t g(z_k) \xrightarrow[t\to\infty]{a.s} \mu(g)$.

With Lemma 1 we can apply a strong law of large numbers upon $1/t \sum_{k=0}^{t-1} \ln |\eta^\star_{k-1}| = 1/t \sum_{k=0}^{t-1} 1/(2d_\sigma)(\boldsymbol{\xi}^\star_{k-1}/n - 1)$, to get as in the proof of Proposition 1 $1/(2d_\sigma n)(\mathbb{E}(\mathcal{N}^2_{1:\lambda}) - 1)$.

By inserting these results into Eq. (9), we get that $1/t \ln |\,[\boldsymbol{X}_t]_1 / [\boldsymbol{X}_0]_1\,| \xrightarrow[t\to\infty]{a.s} \mu(g) - \mu(g) + 1/(2d_\sigma n)(\mathbb{E}(\mathcal{N}^2_{1:\lambda}) - 1)$, which with Lemma 3 is strictly positive for $\lambda \geq 3$. □

4 Divergence rate of CSA-ES with cumulation

We are now investigating the $(1, \lambda)$-CSA-ES with cumulation, i.e. $0 < c < 1$.

According to Lemma 1, the random variables $(\boldsymbol{\xi}_t^\star)_{t\in\mathbb{N}}$ are i.i.d., hence the path $\boldsymbol{\mathcal{P}} = (\boldsymbol{p}_t)_{t\in\mathbb{N}}$ is a Markov chain. By a recurrence on Eq. (2) we see that the path follows the following equation

$$\boldsymbol{p}_t = (1-c)^t \boldsymbol{p}_0 + \sqrt{c(2-c)} \sum_{k=0}^{t-1} (1-c)^k \underbrace{\boldsymbol{\xi}_{t-1-k}^\star}_{\text{i.i.d.}} \quad . \tag{15}$$

For $i \neq 1$, $[\boldsymbol{\xi}_t^\star]_i \sim \mathcal{N}(0,1)$ and, as also $[\boldsymbol{p}_0]_i \sim \mathcal{N}(0,1)$, by recurrence $[\boldsymbol{p_t}]_i \sim \mathcal{N}(0,1)$ for all $t \in \mathbb{N}$. For $i = 1$ with cumulation ($c < 1$), the influence of $[\boldsymbol{p}_0]_1$ vanishes with $(1-c)^t$. Furthermore, as from Lemma 1 the sequence $([\boldsymbol{\xi}_t^\star]_1])_{t\in\mathbb{N}}$ is independent, we get by applying the Kolgomorov's three series theorem that the series $\sum_{k=0}^{t-1}(1-c)^k \left[\boldsymbol{\xi}_{t-1-k}^\star\right]_1$ converges almost surely. Therefore, the first component of the path becomes distributed as the random variable $[\boldsymbol{p}_\infty]_1 = \sqrt{c(2-c)}\sum_{k=0}^{\infty}(1-c)^k[\boldsymbol{\xi}_k^\star]_1$ (by re-indexing the variable $\boldsymbol{\xi}_{t-1-k}^\star$ in $\boldsymbol{\xi}_k^\star$, as the sequence $(\boldsymbol{\xi}_t^\star)_{t\in\mathbb{N}}$ is i.i.d.).

As in Subsection 3.1 we will show that $\boldsymbol{\mathcal{P}}$ has the right stability properties to apply a law of large numbers to it. First we will extract from $\boldsymbol{\mathcal{P}}$ the part of interest as stated in the following lemma.

Lemma 11. *On linear functions, for any λ the step-size of the $(1,\lambda)$-CSA-ES follows almost surely*

$$\frac{1}{t}\ln\left(\frac{\sigma_t}{\sigma_0}\right) - \frac{c}{2d_\sigma n}\left(\frac{1}{t}\sum_{i=1}^{t}([\boldsymbol{p}_i]_1^2 - 1\right) \xrightarrow[t\to\infty]{a.s} 0 \quad , \tag{16}$$

and in expectancy

$$\mathbb{E}\left(\ln\left(\frac{\sigma_{t+1}}{\sigma_t}\right)\right) = \frac{c}{2d_\sigma n}\left(\mathbb{E}\left([\boldsymbol{p}_{t+1}]_1^2\right) - 1\right) \tag{17}$$

Proof. We separate Eq. (5) over the dimensions, which gives us that $1/t\ln(\sigma_t/\sigma_0) = c/(2d_\sigma n)(\sum_{i=1}^n 1/t\sum_{j=1}^t[\boldsymbol{p}_j]_i^2 - n)$, so $1/t\ln(\sigma_t/\sigma_0) - c/(2d_\sigma n)(1/t\sum_{j=1}^t[\boldsymbol{p}_j]_1^2 - 1) = \sum_{i=2}^n c/(2d_\sigma n)(1/t\sum_{j=1}^t[\boldsymbol{p}_j]_i^2 - 1)$. As for $i \neq 1$, $[\boldsymbol{p}_0]_i\boldsymbol{\mathcal{N}}(\mathbf{0}, I_n)$ and $[\boldsymbol{\xi}_0^\star]_i$ has no selection pressure, then $[\boldsymbol{p}_1]_i\boldsymbol{\mathcal{N}}(\mathbf{0}, I_n)$, and per recurrence $[\boldsymbol{p}_k]_i\boldsymbol{\mathcal{N}}(\mathbf{0}, I_n)$ for any $k \in \mathbb{N}$. Therefore, we can apply the strong law of large numbers and $1/t\sum_{j=1}^t[\boldsymbol{p}_j]_i^2 \xrightarrow[t\to\infty]{a.s} 1$, which gives us Eq. (16).

The same reasoning over Eq. (4) gives Eq. (17). □

The part of $\boldsymbol{\mathcal{P}}$ left to analyse is its first dimension $[\boldsymbol{\mathcal{P}}]_1 = ([\boldsymbol{p}_i]_1)_{i\in\mathbb{N}}$. We start the study of $[\boldsymbol{\mathcal{P}}]_1$ with the following lemma.

Lemma 12. *The Markov chain $[\boldsymbol{\mathcal{P}}]_1$ is φ-irreducible, aperiodic, and compacts of $\mathbb{R}$ are small-sets.*

Proof. We have the following transition kernel:

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

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

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

As $P\left(\mathcal{N}_{1:\lambda} = x\right) > 0$ for all $x \in \mathbb{R}$, for all A non-μ_{Leb}-negligible we have $P(p, A) > 0$, thus the chain $[\boldsymbol{\mathcal{P}}]_1$ is μ_{Leb}-irreducible.

Furthermore, if we take C a non-μ_{Leb}-negligible compact of $\mathbb{R}$, and ν_C a measure such that for A a borel set of $\mathbb{R}$,
$\nu_C(A) = \frac{1}{\sqrt{(2-c)c}} \int_{\mathbb{R}} \mathbf{1}_A\left(\tilde{u}\right) \min_{p \in C} P\left(\boldsymbol{\xi}_t^\star = \left(\tilde{u} - (1-c)p\right) / \left(\sqrt{(2-c)c}\right)\right) d\tilde{u}$, we see that $P(p, A) \geq \nu_C(A)$ for all $p \in \mathbb{R}$, while ν_C is not a trivial measure (indeed, $P(\mathcal{N}_{1:\lambda} = x) > k > 0$ for all $x \in C$); Therefore compact sets of $\mathbb{R}$ are small sets for $[\boldsymbol{\mathcal{P}}]_1$. Finally, $\nu_C(C) > 0$, so the chain $[\boldsymbol{\mathcal{P}}]_1$ is strongly aperiodic. □

We use this new lemma with Lemma 5 to prove what is needed to apply the law of large numbers on $[\boldsymbol{\mathcal{P}}]_1$.

Lemma 13. *The chain $[\boldsymbol{\mathcal{P}}]_1$ is Harris recurrent positive with invariant measure μ_{path}, and the function $x \mapsto x^2$ is μ_{path}-integrable.*

Proof. We now have to get the right drift condition for the chain. Let $V : x \mapsto x^2 + 1$.

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

As $0 < c \leq 1$, $c(2-c)$ is strictly positive and therefore, for $\epsilon > 0$ there exists $C = [-M, M]$ with $M > 0$ such that for all $x \in C^c$, $\Delta V(x)/V(x) \leq -\epsilon$. If we take $b = \epsilon V(M) + 2M\sqrt{c(2-c)}\mathbb{E}(|\left[\boldsymbol{\xi}_\cdot^\star\right]_1|) + c(2-c)\mathbb{E}(\left[\boldsymbol{\xi}_t^\star\right]_1{}^2)$, then for all $x \in C$ we have $\Delta V(x) \leq b$. Hence the drift condition $\Delta V(x) \leq -\epsilon V(x) + b\mathbf{1}_C$ is satisfied for all $x \in \mathbb{R}$.

According to Lemma 12 the chain $[\boldsymbol{\mathcal{P}}]_1$ is φ-irreducible and aperiodic, so with Lemma 5 it is positive Harris recurrent, with invariant measure μ_{path}, and V is μ_{path}-integrable. Therefore the function $x \mapsto x^2$ is also μ_{path}-integrable.

To obtain an equality between the rate we get through almost sure divergence, and the rate in expectation, we need to define the f-norm, which for a signed measure ν and a function $f \geq 1$ is equal to $\|\nu\|_f = \sup_{g:|g| \leq f} |\nu(g)|$, and we need the following lemma.

Lemma 14. *[11, 14.3.5] Suppose $\boldsymbol{\Phi}$ is an aperiodic positive Harris chain on a space $\mathcal{X}$ with stationary measure π, and that there exists some non-negative function V, a function $f \geq 1$, a small-set C and $b \in \mathbb{R}$ such that for all $x \in \mathcal{X}$, $\Delta V(x) \leq -f(x) + b\mathbf{1}_C(x)$. Then for all initial probability distribution ν, $\|\nu P^n - \pi\|_f \underset{t\to\infty}{\longrightarrow} 0$.*

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

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

$$\frac{1}{t}\ln\left(\frac{\sigma_t}{\sigma_0}\right) \underset{t\to\infty}{\longrightarrow} \frac{1}{2d_\sigma n} \underbrace{\left(2(1-c)\,\mathbb{E}\left(\mathcal{N}_{1:\lambda}\right)^2 + c\left(\mathbb{E}\left(\mathcal{N}_{1:\lambda}^2\right) - 1\right)\right)}_{>0 \text{ for } \lambda\geq 3 \text{ and for } \lambda=2 \text{ and } c<1} . \tag{18}$$

Proof. We will start by the convergence in expectation. From Eq. (17) we see that the part to develop is $\mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^2)$. By recurrence $\left[\boldsymbol{p}_{t+1}\right]_1 = (1-c)^{t+1}[\boldsymbol{p}_0]_1 + \sqrt{c(2-c)}\sum_{i=0}^{t}(1-c)^i\left[\boldsymbol{\xi}_{t-i}^\star\right]_1$. When t goes to infinity, the influence of $[\boldsymbol{p}_0]_1$ in this equation goes to 0 with $(1-c)^{t+1}$, so we can remove it when taking the limit:

$$\lim_{t\to\infty}\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^2\right) = \lim_{t\to\infty}\mathbb{E}\left(\left(\sqrt{c(2-c)}\sum_{i=0}^{t}(1-c)^i\left[\boldsymbol{\xi}_{t-i}^\star\right]_1\right)^2\right) \tag{19}$$

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

$$\lim_{t\to\infty} c(2-c)\left(2\sum_{i=0}^{t}\sum_{j=i+1}^{t}(1-c)^{i+j}\underbrace{\mathbb{E}\left(\left[\boldsymbol{\xi}_{t-i}^\star\right]_1\left[\boldsymbol{\xi}_{t-j}^\star\right]_1\right)}_{=\mathbb{E}\left[\boldsymbol{\xi}_{t-i}^\star\right]_1\mathbb{E}\left[\boldsymbol{\xi}_{t-j}^\star\right]_1=\mathbb{E}[\mathcal{N}_{1:\lambda}]^2} + \sum_{i=0}^{t}(1-c)^{2i}\underbrace{\mathbb{E}\left(\left[\boldsymbol{\xi}_{t-i}^\star\right]_1^2\right)}_{=\mathbb{E}[\mathcal{N}_{1:\lambda}^2]}\right) \tag{20}$$

Now the expected value does not depend on i or j, so what is left is to calculate $\sum_{i=0}^t\sum_{j=i+1}^t(1-c)^{i+j}$ and $\sum_{i=0}^t(1-c)^{2i}$. We have $\sum_{i=0}^t\sum_{j=i+1}^t(1-c)^{i+j} = \sum_{i=0}^t(1-c)^{2i+1}\frac{1-(1-c)^{t-i}}{1-(1-c)}$ and when we separates this sum in two, the right hand side goes to 0 for $t\to\infty$. Therefore, the left hand side converges to $\lim_{t\to\infty}\sum_{i=0}^t(1-c)^{2i+1}/c$, which is equal to $\lim_{t\to\infty}(1-c)/c\sum_{i=0}^t(1-c)^{2i}$. And $\sum_{i=0}^t(1-c)^{2i}$ is equal to $(1-(1-c)^{2t+2})/(1-(1-c)^2)$, which converges to $1/(c(2-c))$. So, by inserting this in Eq. (20) we get that $\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^2\right) \underset{t\to\infty}{\longrightarrow} 2\frac{1-c}{c}\mathbb{E}\left(\mathcal{N}_{1:\lambda}\right)^2 + \mathbb{E}\left(\mathcal{N}_{1:\lambda}^2\right)$, which gives us the right hand side of Eq. (18).

By summing $\mathbb{E}(\ln(\sigma_{i+1}/\sigma_i))$ for $i = 0,\ldots,t-1$ and dividing by t we have the Cesaro mean $1/t\mathbb{E}(\ln(\sigma_t/\sigma_0))$ that converges to the same value that $\mathbb{E}(\ln(\sigma_{t+1}/\sigma_t))$ converges to when t goes to infinity. Therefore we have in expectation Eq. (18).

We will now focus on the almost sure convergence. From Lemma 13, we see that we have the right conditions to apply Lemma 4 to the chain $[\boldsymbol{\mathcal{P}}]_1$ with the μ_{path}-integrable function $g : x \mapsto x^2$. So $1/t \sum_{k=1}^{t} [\boldsymbol{p}_k]_1^2 \xrightarrow[t\to\infty]{a.s} \mu_{path}(g)$. With Eq. (16) we obtain that $1/t \ln(\sigma_t/\sigma_0) \xrightarrow[t\to\infty]{a.s} c/(2d_\sigma n)(\mu_{path}(g) - 1)$.

We will now prove that $\mu_{path}(g) = \lim_{t\to\infty} \mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^2)$. Let ν be the initial distribution of $[\boldsymbol{p}_0]1$, so we have $|\mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^2) - \mu_{path}(g)| \leq \|\nu P^{t+1} - \mu_{path}\|_h$, with $h : x \mapsto 1 + x^2$. From the proof of Lemma 13 and from Lemma 12 we have all conditions for Lemma 14. Therefore $\|\nu P^{t+1} - \mu_{path}\|_h \xrightarrow[t\to\infty]{} 0$, which shows that $\mu_{path}(g) = \lim_{t\to\infty} \mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^2) = (2-2c)/c\mathbb{E}(\mathcal{N}_{1:\lambda})^2 + \mathbb{E}(\mathcal{N}_{1:\lambda}^2)$.

According to Lemma 3, for $\lambda = 2$, $\mathbb{E}(\mathcal{N}_{1:2}^2) = 1$, so the RHS of Eq. (18) is equal to $(1-c)/(d_\sigma n)\mathbb{E}(\mathcal{N}_{1:2})^2$. The expected value of $\mathcal{N}_{1:2}$ is strictly negative, so the previous expression is strictly positive. Furthermore, according to Lemma 3, $\mathbb{E}(\mathcal{N}_{1:\lambda}^2)$ increases strictly with λ, as does $\mathbb{E}(\mathcal{N}_{1:2})^2$. Therefore we have geometric divergence for $\lambda \geq 2$ if $c < 1$, and for $\lambda \geq 3$.

□

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

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

The proof of Theorem 3 shows that the step size increase converges to the right hand side of Eq. (18), for $t \to \infty$. When the dimension increases this increment goes to zero, which also suggests that it becomes more likely that σ_{t+1} is smaller than σ_t. To analyze this behavior, we study the variance of $\ln(\sigma_{t+1}/\sigma_t)$ as a function of c and the dimension.

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

$$\mathrm{Var}\left(\ln\left(\frac{\sigma_{t+1}}{\sigma_t}\right)\right) = \frac{c^2}{4d_\sigma^2 n^2}\left(\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^4\right) - \mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^2\right)^2 + 2(n-1)\right) \quad . \quad (21)$$

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

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

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

Proof.

$$\mathrm{Var}\left(\ln\left(\frac{\sigma_{t+1}}{\sigma_t}\right)\right) = \mathrm{Var}\left(\frac{c}{2d_\sigma}\left(\frac{\|\boldsymbol{p}_{t+1}\|^2}{n} - 1\right)\right) = \frac{c^2}{4d_\sigma^2 n^2} \underbrace{\mathrm{Var}\left(\|\boldsymbol{p}_{t+1}\|^2\right)}_{\mathbb{E}\left(\|\boldsymbol{p}_{t+1}\|^4\right) - \mathbb{E}\left(\|\boldsymbol{p}_{t+1}\|^2\right)^2} \tag{23}$$

The first part of $\mathrm{Var}(\|\boldsymbol{p}_{t+1}\|^2)$, $\mathbb{E}(\|\boldsymbol{p}_{t+1}\|^4)$, is equal to $\mathbb{E}((\sum_{i=1}^n \left[\boldsymbol{p}_{t+1}\right]_i^2)^2)$. We develop it along the dimensions such that we can use the independence of $[\boldsymbol{p}_{t+1}]_i$ with $[\boldsymbol{p}_{t+1}]_j$ for $i \neq j$, to get $\mathbb{E}(2\sum_{i=1}^n \sum_{j=i+1}^n \left[\boldsymbol{p}_{t+1}\right]_i^2 \left[\boldsymbol{p}_{t+1}\right]_j^2 + \sum_{i=1}^n \left[\boldsymbol{p}_{t+1}\right]_i^4)$. For $i \neq 1$ $\left[\boldsymbol{p}_{t+1}\right]_i$ is distributed according to a standard normal distribution, so $\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_i^2\right) = 1$ and $\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_i^4\right) = 3$.

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

The other part left is $\mathbb{E}(\|\boldsymbol{p}_{t+1}\|^2)^2$, which we develop along the dimensions to get $\mathbb{E}(\sum_{i=1}^n \left[\boldsymbol{p}_{t+1}\right]_i^2)^2 = (\mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^2) + (n-1))^2$, which equals to $\mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^2)^2 + 2(n-1)\mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^2) + (n-1)^2$. So by subtracting both parts we get $\mathbb{E}(\|\boldsymbol{p}_{t+1}\|^4) - \mathbb{E}(\|\boldsymbol{p}_{t+1}\|^2)^2 = \mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^4) - \mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^2)^2 + 2(n-1)$, which we insert into Eq. (23) to get Eq. (21).

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

$$\lim_{t\to\infty} \mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^4\right) = \lim_{t\to\infty} \mathbb{E}\left(c^2(2-c)^2\left(\sum_{i=0}^t (1-c)^i \left[\boldsymbol{\xi}_{t-i}^\star\right]_1\right)^4\right) . \tag{24}$$

To develop the RHS of Eq.(24) we use the following formula: for $(a_i)_{i\in[[1,m]]}$

$$\left(\sum_{i=1}^{m} a_i\right)^4 = \sum_{i=1}^{m} a_i^4 + 4\sum_{i=1}^{m}\sum_{\substack{j=1\\ j\neq i}}^{m} a_i^3 a_j + 6\sum_{i=1}^{m}\sum_{j=i+1}^{m} a_i^2 a_j^2 + 12\sum_{i=1}^{m}\sum_{\substack{j=1\\ j\neq i}}^{m}\sum_{\substack{k=j+1\\ k\neq i}}^{m} a_i^2 a_j a_k + 24\sum_{i=1}^{m}\sum_{j=i+1}^{m}\sum_{k=j+1}^{m}\sum_{l=k+1}^{m} a_i a_j a_k a_l \ . \tag{25}$$

This formula will allow us to use the independence over time of $[\boldsymbol{\xi}_t^\star]_1$ from Lemma 1, so that $\mathbb{E}([\boldsymbol{\xi}_i^\star]_1^3 [\boldsymbol{\xi}_j^\star]_1) = \mathbb{E}([\boldsymbol{\xi}_i^\star]_1^3)\mathbb{E}([\boldsymbol{\xi}_j^\star]_1) = \mathbb{E}(\mathcal{N}_{1:\lambda}^3)\mathbb{E}(\mathcal{N}_{1:\lambda})$ for $i \neq j$, and so on. We apply Eq (25) on Eq (22), with $a = 1 - c$.

$$\begin{aligned}\lim_{t\to\infty} \frac{\mathbb{E}\left([\boldsymbol{p}_{t+1}]_1^4\right)}{c^2(2-c)^2} = \lim_{t\to\infty} &\sum_{i=0}^{t} a^{4i}\mathbb{E}\left(\mathcal{N}_{1:\lambda}^4\right) + 4\sum_{i=0}^{t}\sum_{\substack{j=0\\ j\neq i}}^{t} a^{3i+j}\mathbb{E}\left(\mathcal{N}_{1:\lambda}^3\right)\mathbb{E}\left(\mathcal{N}_{1:\lambda}\right) \\ &+ 6\sum_{i=0}^{t}\sum_{j=i+1}^{t} a^{2i+2j}\mathbb{E}\left(\mathcal{N}_{1:\lambda}^2\right)^2 \\ &+ 12\sum_{i=0}^{t}\sum_{\substack{j=0\\ j\neq i}}^{t}\sum_{\substack{k=j+1\\ k\neq i}}^{t} a^{2i+j+k}\mathbb{E}\left(\mathcal{N}_{1:\lambda}^2\right)\mathbb{E}\left(\mathcal{N}_{1:\lambda}\right)^2 \\ &+ 24\sum_{i=0}^{t}\sum_{j=i+1}^{t}\sum_{k=j+1}^{t}\sum_{l=k+1}^{t} a^{i+j+k+l}\mathbb{E}\left(\mathcal{N}_{1:\lambda}\right)^4 \end{aligned} \tag{26}$$

We now have to develop each term of Eq. (26).

$$\begin{aligned}\sum_{i=0}^{t} a^{4i} &= \frac{1-a^{4(t+1)}}{1-a^4} \\ \lim_{t\to\infty}\sum_{i=0}^{t} a^{4i} &= \frac{1}{1-a^4}\end{aligned} \tag{27}$$

$$\sum_{i=0}^{t}\sum_{\substack{j=0\\ j\neq i}}^{t} a^{3i+j} = \sum_{i=0}^{t-1}\sum_{j=i+1}^{t} a^{3i+j} + \sum_{i=1}^{t}\sum_{j=0}^{i-1} a^{3i+j} \tag{28}$$

$$\sum_{i=0}^{t-1}\sum_{j=i+1}^{t} a^{3i+j} = \sum_{i=0}^{t-1} a^{4i+1}\frac{1-a^{t-i}}{1-a}$$
$$\lim_{t\to\infty}\sum_{i=0}^{t-1}\sum_{j=i+1}^{t} a^{3i+j} = \lim_{t\to\infty}\frac{a}{1-a}\sum_{i=0}^{t-1} a^{4i}$$
$$= \frac{a}{(1-a)(1-a^4)} \tag{29}$$

$$\sum_{i=1}^{t}\sum_{j=0}^{i-1} a^{3i+j} = \sum_{i=1}^{t} a^{3i}\frac{1-a^i}{1-a}$$
$$= \frac{1}{1-a}\left(a^3\frac{1-a^{3t}}{1-a^3} - a^4\frac{1-a^{4t}}{1-a^4}\right)$$
$$\lim_{t\to\infty}\sum_{i=1}^{t}\sum_{j=0}^{i-1} a^{3i+j} = \frac{1}{1-a}\left(\frac{a^3}{1-a^3} - \frac{a^4}{1-a^4}\right)$$
$$= \frac{a^3(1-a^4)-a^4(1-a^3)}{(1-a)(1-a^3)(1-a^4)}$$
$$= \frac{a^3-a^4}{(1-a)(1-a^3)(1-a^4)} \tag{30}$$

By combining Eq (29) with Eq (30) to Eq (28) we get

$$\lim_{t\to\infty}\sum_{i=0}^{t}\sum_{\substack{j=0\\ j\neq i}}^{t} a^{3i+j} = \frac{a(1-a^3)+a^3-a^4}{(1-a)(1-a^3)(1-a^4)} = \frac{a(1+a^2-2a^3)}{(1-a)(1-a^3)(1-a^4)}$$
$$= \frac{a(1-a)(1+a+2a^2))}{(1-a)(1-a^3)(1-a^4)} = \frac{a(1+a+2a^2))}{(1-a^3)(1-a^4)} \tag{31}$$

$$\sum_{i=0}^{t-1}\sum_{j=i+1}^{t} a^{2i+2j} = \sum_{i=0}^{t-1} a^{4i+2}\frac{1-a^{2(t-i)}}{1-a^2}$$
$$\lim_{t\to\infty}\sum_{i=0}^{t-1}\sum_{j=i+1}^{t} a^{2i+2j} = \frac{a^2}{1-a^2}\sum_{i=0}^{t-1} a^{4i}$$
$$= \frac{a^2}{(1-a^2)(1-a^4)} \tag{32}$$

$$\sum_{i=0}^{t}\sum_{\substack{j=0\\ j\neq i}}^{t-1}\sum_{\substack{k=j+1\\ k\neq i}}^{t} a^{2i+j+k} = \sum_{i=2}^{t}\sum_{j=0}^{i-2}\sum_{k=j+1}^{i-1} a^{2i+j+k} + \sum_{i=1}^{t-1}\sum_{j=0}^{i-1}\sum_{k=i+1}^{t} a^{2i+j+k}$$
$$+ \sum_{i=0}^{t-2}\sum_{j=i+1}^{t-1}\sum_{k=j+1}^{t} a^{2i+j+k} \tag{33}$$

$$
\begin{aligned}
\sum_{i=2}^{t}\sum_{j=0}^{i-2}\sum_{k=j+1}^{i-1} a^{2i+j+k} &= \sum_{i=2}^{t}\sum_{j=0}^{i-2} a^{2i+2j+1}\frac{1-a^{i-j-1}}{1-a} \\
&= \frac{1}{1-a}\sum_{i=2}^{t} a^{2i+1}\frac{1-a^{2(i-1)}}{1-a^2} - a^{3i}\frac{1-a^{i-1}}{1-a} \\
&= \frac{1}{1-a}\Bigg(\frac{a^5}{1-a^2}\frac{1-a^{2(t-1)}}{1-a^2} - \frac{a^7}{(1-a^2)}\frac{1-a^{4(t-1)}}{1-a^4} \\
&\quad - \frac{a^6}{1-a}\frac{1-a^{3(t+1)}}{1-a^3} + \frac{a^7}{1-a}\frac{1-a^{4(t+1)}}{1-a^4}\Bigg) \\
&\xrightarrow[t\to\infty]{} \frac{a^5}{1-a}\Bigg(\frac{1}{(1-a^2)^2} - \frac{a^2}{(1-a^2)(1-a^4)} \\
&\quad - \frac{a}{(1-a)(1-a^3)} + \frac{a^2(1+a)}{(1+a)(1-a)(1-a^4)}\Bigg) \\
&\xrightarrow[t\to\infty]{} \frac{a^5}{1-a}\Bigg(\frac{(1+a^2)}{(1-a^2)^2(1+a^2)} + \frac{a^3}{(1-a^2)(1-a^4)} \\
&\quad - \frac{a}{(1-a)(1-a^3)}\Bigg) \\
&\xrightarrow[t\to\infty]{} \frac{a^5}{1-a}\Bigg(\frac{1+a^2+a^3}{(1+a)(1-a)(1-a^4)} - \frac{a}{(1-a)(1-a^3)}\Bigg) \\
&\xrightarrow[t\to\infty]{} \frac{a^5}{(1-a)^2}\frac{(1+a^2+a^3)(1-a^3)-a(1+a)(1-a^4)))}{(1+a)(1-a^3)(1-a^4)} \\
&\xrightarrow[t\to\infty]{} a^5\frac{1+a^2-a^5-a^6-(a+a^2-a^5-a^6)}{(1-a)(1-a^2)(1-a^3)(1-a^4)} \\
&\xrightarrow[t\to\infty]{} \frac{a^5}{(1-a^2)(1-a^3)(1-a^4)}
\end{aligned} \tag{34}
$$

$$
\begin{aligned}
\sum_{i=1}^{t-1}\sum_{j=0}^{i-1}\sum_{k=i+1}^{t} a^{2i+j+k} &= \sum_{i=1}^{t-1}\sum_{j=0}^{i-1} a^{3i+j+1}\frac{1-a^{t-i}}{1-a} \\
&\xrightarrow[t\to\infty]{} \lim_{t\to\infty}\frac{a}{1-a}\sum_{i=1}^{t-1} a^{3i}\frac{1-a^{i}}{1-a} \\
&\xrightarrow[t\to\infty]{} \lim_{t\to\infty}\frac{a}{(1-a)^2}\left(a^3\frac{1-a^{3t}}{1-a^3} - a^4\frac{1-a^{4t}}{1-a^4}\right) \\
&\xrightarrow[t\to\infty]{} \frac{a}{(1-a)^2}\left(\frac{a^3(1-a^4)-a^4(1-a^3)}{(1-a^3)(1-a^4)}\right) \\
&\xrightarrow[t\to\infty]{} \frac{a^4-a^5}{(1-a)^2(1-a^3)(1-a^4)} = \frac{a^4}{(1-a)(1-a^3)(1-a^4)}
\end{aligned} \tag{35}
$$

$$\begin{aligned}
\sum_{i=0}^{t-2}\sum_{j=i+1}^{t-1}\sum_{k=j+1}^{t} a^{2i+j+k} &= \sum_{i=0}^{t-2}\sum_{j=i+1}^{t-1} a^{2i+2j+1}\frac{1-a^{t-j}}{1-a} \\
&\xrightarrow[t\to\infty]{} \lim_{t\to\infty}\frac{a}{1-a}\sum_{i=0}^{t-2} a^{4i+2}\frac{1-a^{2(t-i-1)}}{1-a^2} \\
&\xrightarrow[t\to\infty]{} \lim_{t\to\infty}\frac{a^3}{(1-a)(1-a^2)}\frac{1-a^{4(t-1)}}{1-a^4} \\
&\xrightarrow[t\to\infty]{} \frac{a^3}{(1-a)(1-a^2)(1-a^4)}
\end{aligned} \tag{36}$$

We now combine Eq (34), Eq. (35) and Eq. (34) in Eq. (33).

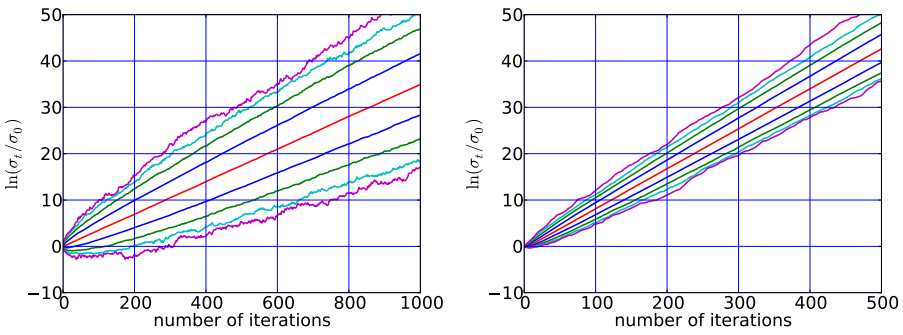
$$\begin{aligned}
\sum_{i=0}^{t}\sum_{\substack{j=0\\ j\neq i}}^{t-1}\sum_{\substack{k=j+1\\ k\neq i}}^{t} a^{2i+j+k} &\xrightarrow[t\to\infty]{} \frac{a^5(1-a)+a^4(1-a^2)+a^3(1-a^3)}{(1-a)(1-a^2)(1-a^3)(1-a^4)} \\
&\xrightarrow[t\to\infty]{} \frac{a^3+a^4+a^5-3a^6}{(1-a)(1-a^2)(1-a^3)(1-a^4)} \\
&\xrightarrow[t\to\infty]{}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

$$\begin{aligned}
\sum_{i=0}^{t-3}\sum_{j=i+1}^{t-2}\sum_{k=j+1}^{t-1}\sum_{l=k+1}^{t} a^{i+j+k+l} &= \sum_{i=0}^{t-3}\sum_{j=i+1}^{t-2}\sum_{k=j+1}^{t-1} a^{i+j+2k+1}\frac{1-a^{t-k}}{1-a} \\
&\xrightarrow[t\to\infty]{} \lim_{t\to\infty}\frac{a}{1-a}\sum_{i=0}^{t-3}\sum_{j=i+1}^{t-2} a^{i+3j+2}\frac{1-a^{2(t-1-j)}}{1-a^2} \\
&\xrightarrow[t\to\infty]{} \lim_{t\to\infty}\frac{a^3}{(1-a)(1-a^2)}\sum_{i=0}^{t-3} a^{4i+3}\frac{1-a^{3(t-2-i)}}{1-a^3} \\
&\xrightarrow[t\to\infty]{} \lim_{t\to\infty}\frac{a^6}{(1-a)(1-a^2)(1-a^3)}\frac{1-a^{4(t-2)}}{1-a^4} \\
&\xrightarrow[t\to\infty]{} \frac{a^6}{(1-a)(1-a^2)(1-a^3)(1-a^4)}
\end{aligned} \tag{38}$$

By factorising Eq. (27), Eq. (31), Eq. (32), Eq. (37) and Eq. (38) by $\frac{1}{1-a^4}$ we get the coefficients of Theorem 4. □

Figure 1 shows the time evolution of $\ln(\sigma_t/\sigma_0)$ for 5001 runs and $c = 1$ (left) and $c = 1/\sqrt{n}$ (right). By comparing Figure 1a and Figure 1b we observe smaller variations of $\ln(\sigma_t/\sigma_0)$ with the smaller value of c.

Figure 2 shows the relative standard deviation of $\ln(\sigma_{t+1}/\sigma_t)$ (i.e. the standard deviation divided by its expected value). Lowering c, as shown in the left, decreases



(a) Without cumulation ($c = 1$) (b) With cumulation ($c = 1/\sqrt{20}$)

Fig. 1: $\ln(\sigma_t/\sigma_0)$ against t. The different curves represent the quantiles of a set of $5.10^3 + 1$ samples, more precisely the 10^i-quantile and the $1 - 10^{-i}$-quantile for i from 1 to 4; and the median. We have $n = 20$ and $\lambda = 8$.

the relative standard deviation. To get a value below one, c must be smaller for larger dimension. In agreement with Theorem 4, In Figure 2, right, the relative standard deviation increases like $\sqrt{n}$ with the dimension for constant c (three increasing curves). A careful study [8] of the variance equation of Theorem 4 shows that for the choice of $c = 1/(1 + n^\alpha)$, if $\alpha > 1/3$ the relative standard deviation converges to 0 with $\sqrt{(n^{2\alpha} + n)/n^{3\alpha}}$. Taking $\alpha = 1/3$ is a critical value where the relative standard deviation converges to $1/(\sqrt{2}\mathbb{E}(\mathcal{N}_{1:\lambda})^2)$. On the other hand, lower values of α makes the relative standard deviation diverge with $n^{(1-3\alpha)/2}$.

6 Summary

We investigate throughout this paper the $(1, \lambda)$-CSA-ES on affine linear functions composed with strictly increasing transformations. We find, in Theorem 3, the limit distribution for $\ln(\sigma_t/\sigma_0)/t$ and rigorously prove the desired behaviour of σ with $\lambda \geq 3$ for any c, and with $\lambda = 2$ and cumulation ($0 < c < 1$): the step-size diverges geometrically fast. In contrast, without cumulation ($c = 1$) and with $\lambda = 2$, a random walk on $\ln(\sigma)$ occurs, like for the $(1, 2)$-σSA-ES [9] (and also for the same symmetry reason). We derive an expression for the variance of the step-size increment. On linear functions when $c = 1/n^\alpha$, for $\alpha \geq 0$ ($\alpha = 0$ meaning c constant) and for $n \to \infty$ the standard deviation is about $\sqrt{(n^{2\alpha} + n)/n^{3\alpha}}$ times larger than the step-size increment. From this follows that keeping $c < 1/n^{1/3}$ ensures that the standard deviation of $\ln(\sigma_{t+1}/\sigma_t)$ becomes negligible compared to $\ln(\sigma_{t+1}/\sigma_t)$ when the dimensions goes to infinity. That means, the signal to noise ratio goes to zero, giving the algorithm strong stability. The result confirms that even the largest default cumulation parameter $c = 1/\sqrt{n}$ is a stable choice.

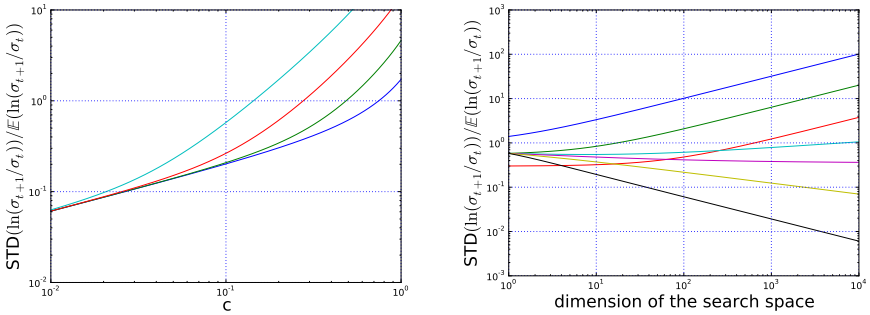


Fig. 2: Standard deviation of $\ln(\sigma_{t+1}/\sigma_t)$ relatively to its expectation. Here $\lambda = 8$. The curves were plotted using Eq. (21) and Eq. (22). On the left, curves for (right to left) $n = 2$, 20, 200 and 2000. On the right, different curves for (top to bottom) $c = 1$, 0.5, 0.2, $1/(1+n^{1/4})$, $1/(1+n^{1/3})$, $1/(1+n^{1/2})$ and $1/(1+n)$.

Acknowledgments

This work was partially supported by the ANR-2010-COSI-002 grant (SIMINOLE) of the French National Research Agency and the ANR COSINUS project ANR-08-COSI-007-12.

References

1. D. V. Arnold and H.-G. Beyer. Performance analysis of evolutionary optimization with cumulative step length adaptation. *IEEE Transactions on Automatic Control*, 49(4):617–622, 2004.
2. D. V. Arnold and H.-G. Beyer. On the behaviour of evolution strategies optimising cigar functions. *Evolutionary Computation*, 18(4):661–682, 2010.
3. D.V. Arnold. Cumulative step length adaptation on ridge functions. In *Parallel Problem Solving from Nature PPSN IX*, pages 11–20. Springer, 2006.
4. D.V. Arnold. On the behaviour of the (1,λ)-es for a simple constrained problem. In *Foundations of Genetic Algorithms FOGA 11*, pages 15–24. ACM, 2011.
5. D.V. Arnold and H.G. Beyer. Random dynamics optimum tracking with evolution strategies. In *Parallel Problem Solving from Nature PPSN VII*, pages 3–12. Springer, 2002.
6. D.V. Arnold and H.G. Beyer. Optimum tracking with evolution strategies. *Evolutionary Computation*, 14(3):291–308, 2006.
7. D.V. Arnold and H.G. Beyer. Evolution strategies with cumulative step length adaptation on the noisy parabolic ridge. *Natural Computing*, 7(4):555–587, 2008.
8. A. Chotard, A. Auger, and N. Hansen. Cumulative step-size adaptation on linear functions: Technical report. 2012. http://hal.inria.fr/hal-00704903.
9. N. Hansen. An analysis of mutative σ-self-adaptation on linear fitness functions. *Evolutionary Computation*, 14(3):255–275, 2006.
10. N. Hansen and A. Ostermeier. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In *International Conference on Evolutionary Computation*, pages 312–317, 1996.

11. S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, second edition, 1993.
12. A. Ostermeier N. Hansen, A. Gawelczyk. Sizing the population with respect to the local progress in (1,lambda)-evolution strategies - a theoretical analysis. *1995 IEEE Internation Conference on Evolutionary Computation Proceedings*, pages 80 – 85, 1995.
13. A. Ostermeier, A. Gawelczyk, and N. Hansen. Step-size adaptation based on non-local use of selection information. In *Proceedings of Parallel Problem Solving from Nature — PPSN III*, volume 866 of *Lecture Notes in Computer Science*, pages 189–198. Springer, 1994.