Cumulative Step-Size Adaptation on Linear Functions

Alexandre Chotard, Anne Auger, and Nikolaus Hansen

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,

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

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) [12,10]. 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 $\mathcal{N}(\mathbf{0}, I_n)$, the ith 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 ith dimension, that is $[x]_i = x_i$. A random variable X distributed according to a law $\mathcal{L}$ will be denoted $X \sim \mathcal{L}$.

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

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

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

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

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

with $0 < c \leq 1$. The constant $1/c$ represents the life span of the information contained in $\boldsymbol{p}_t$, as after $1/c$ generations $\boldsymbol{p}_t$ is multiplied by a factor that approaches $1/e \approx 0.37$ for $c \to 0$ from below (indeed $(1-c)^{1/c} \leq \exp(-1)$). The typical value for c is between $1/\sqrt{n}$ and $1/n$. We will consider that $\boldsymbol{p}_0 \sim \mathcal{N}(\mathbf{0}, I_n)$ 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 $\mathcal{N}(\mathbf{0}, I_n)$ then also $\boldsymbol{p}_{t+1}$ follows $\mathcal{N}(\mathbf{0}, I_n)$. Hence the length of the path can be compared to the expected length of $\|\mathcal{N}(\mathbf{0}, I_n)\|$ 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:

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

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.

Preliminary results on linear functions. Selection on the linear function, $f(\boldsymbol{x}) = [\boldsymbol{x}]_1$, is determined by $[\boldsymbol{X}_t]_1 + \sigma_t [\boldsymbol{\xi}_t^\star]_1 \le [\boldsymbol{X}_t]_1 + \sigma_t [\boldsymbol{\xi}_{t,i}]_1$ for all i which is equivalent to $[\boldsymbol{\xi}_t^\star]_1 \le [\boldsymbol{\xi}_{t,i}]_1$ for all i where by definition $[\boldsymbol{\xi}_{t,i}]_1$ 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 $i \ge 2$.*

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

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

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

$$\frac{1}{t} \ln \left(\frac{\sigma_t}{\sigma_0} \right) = \frac{1}{2d_\sigma} \left(\frac{1}{t} \sum_{k=0}^{t-1} \frac{\|\boldsymbol{\xi}_k^\star\|^2}{n} - 1 \right)$$

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

□

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.

Lemma 2. *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. (see [8] for the full proof) The idea of the proof is to use the symmetry of the normal distribution to show that for two random variables $U \sim \Psi_{1:\lambda+1}$ and $V \sim \Psi_{1:\lambda}$, for every event E_1 where $U^2 < V^2$, there exists another event E_2 counterbalancing the effect of E_1, i.e $\int_{E_2} (u^2 - v^2) f_{U,V}(u, v) \, \mathrm{d}u \, \mathrm{d}v = \int_{E_1} (v^2 - u^2) f_{U,V}(u, v) \, \mathrm{d}u \, \mathrm{d}v$, with $f_{U,V}$ the joint density of the couple (U, V). We then have $\mathbb{E} \left(\mathcal{N}_{1:\lambda+1}^2 \right) \geq \mathbb{E} \left(\mathcal{N}_{1:\lambda}^2 \right)$. As there is a non-negligible set of events E_3, distinct of E_1 and E_2, where $U^2 > V^2$, we have $\mathbb{E}(\mathcal{N}_{1:\lambda+1}^2) > \mathbb{E}(\mathcal{N}_{1:\lambda}^2)$.
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 2 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.*

$$\frac{1}{t} \ln \left(\frac{\sigma_t}{\sigma_0} \right) \xrightarrow[t \to \infty]{a.s.} \mathbb{E} \left(\ln \left(\frac{\sigma_{t+1}}{\sigma_t} \right) \right) = \frac{1}{2d_\sigma n} \left(\mathbb{E} \left(\mathcal{N}_{1:\lambda}^2 \right) - 1 \right) . \tag{6}$$

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 2 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. (6) 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 2, $\mathbb{E}(W_t) = 0$. 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 $\mathcal{N}_{1:1} = \mathcal{N}(0,1)$, $\|\boldsymbol{\xi}\|^2 = \chi_n^2$. $\qquad\square$

In [8] we extend this result on the step-size to $|[\boldsymbol{X}_t]_1|$, which diverges geometrically almost surely at the same rate.

4   Divergence Rate of $(1, \lambda)$-CSA-ES with Cumulation

We are now investigating the $(1, \lambda)$-CSA-ES with cumulation, i.e. $0 < c < 1$. The path $\mathcal{P}$ is then a Markov chain and contrary to the case where $c = 1$ we cannot apply a LLN for independent variables to Eq. (5) in order to prove the almost sure geometric divergence. However LLN for Markov chains exist as well, provided the Markov chain satisfies some stability properties: in particular, if the Markov chain $\mathcal{P}$ is 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 $\mathcal{P}$ starting from any $\boldsymbol{p}_0 \in \mathbb{R}^n$. In addition, the chain $\mathcal{P}$ needs to be (i) positive, that is the chain admits an invariant probability measure π, i.e., for any borelian A, $\pi(A) = \int_{\mathbb{R}^n} P(x, A)\pi(A)$ 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 $\mathcal{P}$ visits A an infinite number of times with probability one. Under those conditions, $\mathcal{P}$ satisfies a LLN, more precisely:

Lemma 3. *[11, 17.0.1] Suppose that $\mathcal{P}$ is a positive Harris chain with invariant probability measure π, and let g be a π-integrable function such that $\pi(|g|) = \int_{\mathbb{R}^n} |g(x)|\pi(dx) < \infty$. Then $1/t \sum_{k=1}^{t} g(\boldsymbol{p}_k) \xrightarrow[t \to \infty]{a.s} \pi(g)$.*

The path $\mathcal{P}$ satisfies the conditions of Lemma 3 and exhibits an invariant measure [8]. By a recurrence on Eq. (2) we see that the path follows the following equation

$$\boldsymbol{p}_t = (1-c)^t \boldsymbol{p}_0 + \sqrt{c(2-c)} \sum_{k=0}^{t-1} (1-c)^k \underbrace{\boldsymbol{\xi}_{t-1-k}^\star}_{\text{i.i.d.}} \quad . \tag{7}$$

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 $\sum_{k=0}^{t-1} (1-c)^k [\boldsymbol{\xi}_{t-1-k}^\star]_1$ converges almost surely. Therefore, the first component of the path becomes distributed as the random variable $[\boldsymbol{p}_\infty]_1 = \sqrt{c(2-c)} \sum_{k=0}^\infty (1-c)^k [\boldsymbol{\xi}_k^\star]_1$ (by re-indexing the variable $\boldsymbol{\xi}_{t-1-k}^\star$ in $\boldsymbol{\xi}_k^\star$, as the sequence $(\boldsymbol{\xi}_t^\star)_{t \in \mathbb{N}}$ is i.i.d.).

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

Theorem 2. *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* $0 < c \leq 1$,

$$\frac{1}{t} \ln\left(\frac{\sigma_t}{\sigma_0}\right) \xrightarrow[t \to \infty]{} \frac{1}{2d_\sigma n} \underbrace{\left(2(1-c) \mathbb{E}(\mathcal{N}_{1:\lambda})^2 + c\left(\mathbb{E}(\mathcal{N}_{1:\lambda}^2) - 1\right)\right)}_{>0 \text{ for } \lambda \geq 3 \text{ and for } \lambda = 2 \text{ and } c < 1} \quad . \tag{8}$$

Proof. For proving almost sure convergence of $\ln(\sigma_t/\sigma_0)/t$ we need to use the LLN for Markov chain. We refer to [8] for the proof that $\mathcal{P}$ satisfies the right assumptions. We now focus on the convergence in expectation. From Eq. (4) we have $\mathbb{E}(\ln(\sigma_{t+1}/\sigma_t)) = c/(2d_\sigma)(\mathbb{E}(\|\boldsymbol{p}_{t+1}\|^2)/n - 1)$, so $\mathbb{E}(\|\boldsymbol{p}_{t+1}\|^2) = \mathbb{E}(\sum_{i=1}^n [\boldsymbol{p}_{t+1}]_i^2)$ is the term we have to analyse. From Eq. (7) and its conclusions we get that for $\jmath \neq 1$ $[\boldsymbol{p}_t]_\jmath \sim \mathcal{N}(0,1)$, so $\mathbb{E}(\sum_{j=1}^n [\boldsymbol{p}_{t+1}]_j^2) = \mathbb{E}([\boldsymbol{p}_{t+1}]_1^2) + (n-1)$. 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:

$$\lim_{t \to \infty} \mathbb{E}\left([\boldsymbol{p}_{t+1}]_1^2\right) = \lim_{t \to \infty} \mathbb{E}\left(\left(\sqrt{c(2-c)} \sum_{i=0}^t (1-c)^i [\boldsymbol{\xi}_{t-i}^\star]_1\right)^2\right) \tag{9}$$

We will now develop the sum with the square, such that we have either a product $[\boldsymbol{\xi}_{t-i}^\star]_1 [\boldsymbol{\xi}_{t-j}^\star]_1$ with $i \neq j$, or $[\boldsymbol{\xi}_{t-j}^\star]_1^2$. 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 $\mathbb{E}([\boldsymbol{p}_{t+1}]_1^2)$ and find that it is equal to

$$\lim_{t \to \infty} c(2-c) \left(2 \sum_{i=0}^t \sum_{j=i+1}^t (1-c)^{i+j} \underbrace{\mathbb{E}\left([\boldsymbol{\xi}_{t-i}^\star]_1 [\boldsymbol{\xi}_{t-j}^\star]_1\right)}_{=\mathbb{E}[\boldsymbol{\xi}_{t-i}^\star]_1 \mathbb{E}[\boldsymbol{\xi}_{t-j}^\star]_1 = \mathbb{E}[\mathcal{N}_{1:\lambda}]^2} + \sum_{i=0}^t (1-c)^{2i} \underbrace{\mathbb{E}\left([\boldsymbol{\xi}_{t-i}^\star]_1^2\right)}_{=\mathbb{E}[\mathcal{N}_{1:\lambda}^2]}\right) \tag{10}$$

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. (10) 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. (8).

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. (8).

According to Lemma 2, for $\lambda=2$, $\mathbb{E}(\mathcal{N}_{1:2}^2)=1$, so the RHS of Eq. (8) 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 2, $\mathbb{E}(\mathcal{N}_{1:\lambda}^2)$ increases with λ, as does $\mathbb{E}(\mathcal{N}_{1:2})^2$. Therefore we have geometric divergence for $\lambda\geq2$. $\qquad\square$

From Eq. (1) we see that the behavior 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 2, means that also the movements in search space and the improvements on affine linear functions f increase geometrically fast. Therefore, as we showed in Theorem 2 geometric divergence for the step-size when $\lambda\geq2$ and $c<1$, or when $\lambda\geq3$, we expect geometric divergence on the first dimension of $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ (the first dimension being the only dimension with selection pressure). Analyzing $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ with cumulation requires to study a double Markov chain, which is left to possible future research.

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

The proof of Theorem 2 shows that the step size increase converges to the right hand side of Eq. (8), 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 $\ln\left(\sigma_{t+1}/\sigma_t\right)$ as a function of c and the dimension.

Theorem 3. *The variance of* $\ln\left(\sigma_{t+1}/\sigma_t\right)$ *equals to*

$$\mathrm{Var}\left(\ln\left(\frac{\sigma_{t+1}}{\sigma_t}\right)\right)=\frac{c^2}{4d_\sigma^2n^2}\left(\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^4\right)-\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^2\right)^2+2(n-1)\right).\quad(11)$$

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

$$\lim_{t\to\infty}\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^4\right)=\frac{(1-a^2)^2}{1-a^4}\left(k_4+k_{31}+k_{22}+k_{211}+k_{1111}\right),\quad(12)$$

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

Proof.

$$\text{Var}\left(\ln\left(\frac{\sigma_{t+1}}{\sigma_t}\right)\right) = \text{Var}\left(\frac{c}{2d_\sigma}\left(\frac{\|\boldsymbol{p}_{t+1}\|^2}{n}-1\right)\right) = \frac{c^2}{4d_\sigma^2 n^2}\underbrace{\text{Var}\left(\|\boldsymbol{p}_{t+1}\|^2\right)}_{\mathbb{E}\left(\|\boldsymbol{p}_{t+1}\|^4\right)-\mathbb{E}\left(\|\boldsymbol{p}_{t+1}\|^2\right)^2}$$
(13)

The first part of $\text{Var}(\|\boldsymbol{p}_{t+1}\|^2)$, $\mathbb{E}(\|\boldsymbol{p}_{t+1}\|^4)$, is equal to $\mathbb{E}((\sum_{i=1}^n \left[\boldsymbol{p}_{t+1}\right]_i^2)^2)$. We develop it along the dimensions such that we can use the independence of $\left[\boldsymbol{p}_{t+1}\right]_i$ with $\left[\boldsymbol{p}_{t+1}\right]_j$ 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 $\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_i^4\right) = 3$.

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

$$= \left(2\sum_{i=2}^n \sum_{j=i+1}^n 1\right) + 2\sum_{j=2}^n \mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^2\right) + \left(\sum_{i=2}^n 3\right) + \mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^4\right)$$

$$= \left(2\sum_{i=2}^n (n-i)\right) + 2(n-1)\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^2\right) + 3(n-1) + \mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^4\right)$$

$$= \mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^4\right) + 2(n-1)\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^2\right) + (n-1)(n+1)$$

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. (13) to get Eq. (11).

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 2. We refer to [8] for the development of $\mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^4)$, since limits of space in the paper prevents us to present it here. □

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 the relative standard deviation. To get a value below one, c must be smaller for larger dimension. In agreement with Theorem 3, 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 3 shows that for the choice of $c = 1/(1 + n^\alpha)$, if $\alpha > 1/3$ the relative standard deviation converges to 0 with
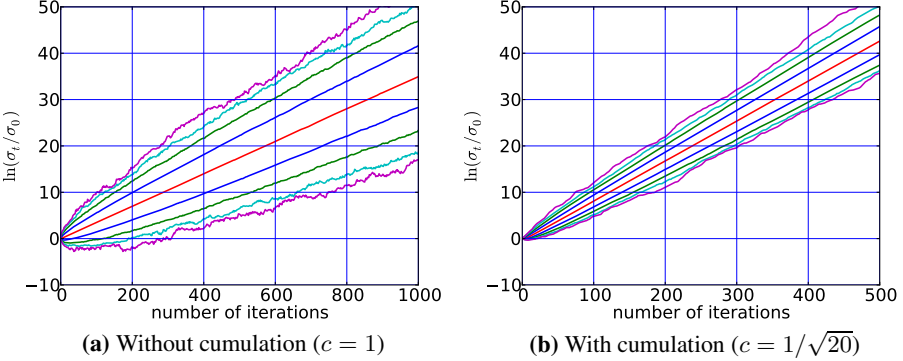
(a) Without cumulation ($c = 1$)   **(b)** With cumulation ($c = 1/\sqrt{20}$)

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 $\lambda = 8$.
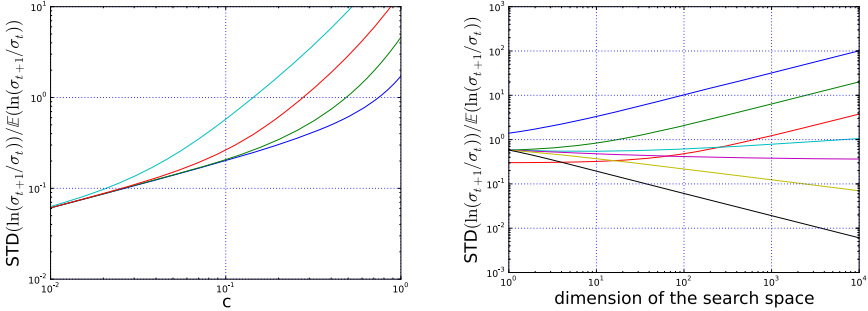

Fig. 2. Standard deviation of $\ln(\sigma_{t+1}/\sigma_t)$ relatively to its expectation. Here $\lambda = 8$. The curves were plotted using Eq. (11) and Eq. (12). On the left, curves for (right to left) $n = 2, 20, 200$ and 2000. On the right, different curves for (top to bottom) $c = 1, 0.5, 0.2, 1/(1 + n^{1/4})$, $1/(1 + n^{1/3})$, $1/(1 + n^{1/2})$ and $1/(1 + n)$.

$\sqrt{(n^{2\alpha} + n)/n^{3\alpha}}$. 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 2, 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.

References

1. Arnold, D.V., Beyer, H.-G.: Performance analysis of evolutionary optimization with cumulative step length adaptation. IEEE Transactions on Automatic Control 49(4), 617–622 (2004)
2. Arnold, D.V., Beyer, H.-G.: On the behaviour of evolution strategies optimising cigar functions. Evolutionary Computation 18(4), 661–682 (2010)
3. Arnold, D.V.: Cumulative Step Length Adaptation on Ridge Functions. In: Runarsson, T.P., Beyer, H.-G., Burke, E.K., Merelo-Guervós, J.J., Whitley, L.D., Yao, X. (eds.) PPSN IX. LNCS, vol. 4193, pp. 11–20. Springer, Heidelberg (2006)
4. Arnold, D.V.: On the behaviour of the (1,λ)-es for a simple constrained problem. In: Foundations of Genetic Algorithms FOGA 2011, pp. 15–24. ACM (2011)
5. Arnold, D.V., Beyer, H.-G.: Random Dynamics Optimum Tracking with Evolution Strategies. In: Guervós, J.J.M., Adamidis, P.A., Beyer, H.-G., Fernández-Villacañas, J.-L., Schwefel, H.-P. (eds.) PPSN VII. LNCS, vol. 2439, pp. 3–12. Springer, Heidelberg (2002)
6. Arnold, D.V., Beyer, H.G.: Optimum tracking with evolution strategies. Evolutionary Computation 14(3), 291–308 (2006)
7. Arnold, D.V., Beyer, H.G.: Evolution strategies with cumulative step length adaptation on the noisy parabolic ridge. Natural Computing 7(4), 555–587 (2008)
8. Chotard, A., Auger, A., Hansen, N.: Cumulative step-size adaptation on linear functions: Technical report (2012), `http://hal.inria.fr/hal-00704903`
9. Hansen, N.: An analysis of mutative σ-self-adaptation on linear fitness functions. Evolutionary Computation 14(3), 255–275 (2006)
10. Hansen, N., Ostermeier, A.: Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In: International Conference on Evolutionary Computation, pp. 312–317 (1996)
11. Meyn, S.P., Tweedie, R.L.: Markov chains and stochastic stability, 2nd edn. Cambridge University Press (1993)
12. Ostermeier, A., Gawelczyk, A., Hansen, N.: Step-size Adaptation Based on Non-local Use of Selection Information. In: Davidor, Y., Männer, R., Schwefel, H.-P. (eds.) PPSN III. LNCS, vol. 866, pp. 189–198. Springer, Heidelberg (1994)