Investigating Adaptive Population Control Strategies with Cumulative Step-Size Adaptation

Amir Omeradzic and Hans-Georg Beyer

***Abstract*—Three state-of-the-art adaptive population control strategies (PCS) are theoretically and empirically investigated for a multi-recombinative, cumulative step-size adaptation Evolution Strategy $(\mu/\mu_I, \lambda)$-CSA-ES. First, scaling properties for the generation number and mutation strength rescaling are derived on the sphere in the limit of large population sizes. Then, the adaptation properties of three standard CSA variants are studied for varying population size and dimensionality, and compared to the predicted scaling results. Thereafter, three PCS are implemented along the CSA-ES and studied on a test bed of sphere, random, and Rastrigin functions. The CSA properties significantly influence the performance and stability of PCS, which is shown in greater detail. Given the test bed, well-performing parameter sets (in terms of scaling, efficiency, and success rate) for both the CSA- and PCS-subroutines are identified.**

***Index Terms*—Evolution Strategy, Population Size Adaptation, CSA-ES, Benchmark**

I. Introduction

Evolution Strategies (ES) have been shown to be well-suited for the optimization of objective functions subject to strong noise or high multimodality. The crucial strategy parameter in both cases is the population size. On noisy functions, such as the noisy sphere, large populations enable to improve the signal-to-noise ratio of the ES and reduce the expected residual distance to the optimizer [1]. For well-structured highly multimodal problems, e.g., the Rastrigin function, they help to facilitate a global search, enabling to locate the global optimizer among exponentially many (in the dimensionality) local optima [2], [3]. While models for the population sizing of a multi-recombinative ES on the Rastrigin function are derived in [2], [3], the results of [4] demonstrate the importance of tuning the population size for the highest efficiency on Rastrigin.

Adaptive population control strategies (abbreviated as PCS) are able to adapt to both strong noise [5], [6], [7] and high multimodality [7], [8] by increasing the population size when insufficient algorithm performance is detected. While benchmarks on the well-known Black-Box Optimization Benchmarking (BBOB) test bed [9] are available in many cases, an in-depth analysis of adaptive population control is still pending. Furthermore, PCS are often tuned for the underlying test bed, such that direct comparison between different methods turns out to be difficult in some cases. In principle, adaptive population control can be implemented into any ES. State-of-the-art ES use covariance matrix adaptation (CMA) [10] for the search space distribution together with cumulative step-size adaptation (CSA) [11], [12] for the global mutation strength. As it turns out, the PCS performance varies notably depending on the parameter settings of CMA and CSA, including cumulation and damping factors. This also holds among standard implementations of the CSA-ES, which will be shown. Population control cannot be studied isolated from the underlying mutation strength adaptation. A change of the population size usually results in changes in the mutation strength, which in turn influences the performance measure of the PCS. Hence, three state-of-the-art PCS routines [6], [7], [8] will be selected that are suitable for both noisy and multimodal objective functions. The PCS will incorporate performance measures both in search space and in fitness space, respectively. However, they will be investigated for isotropic mutations by using only the CSA-ES (without CMA) on a simple test bed. The idea is to reduce the problem complexity and provide theoretical reasoning for the parameter choices by using results from progress rate theory. To this end, adaptation time scales and mutation strength rescaling are analyzed for CSA and PCS both theoretically and experimentally. The obtained parametrization is then applied to all PCS, enabling a direct comparison of the methods for the first time. Furthermore, the influence of different CSA variants on adaptive population control is studied.

In Sec. II, the ES incorporating CSA and PCS is introduced. In Sec. III, progress rate theory on the sphere function is applied to derive laws for the generation number scaling and mutation strength rescaling. In Sec. IV, a simple population update schedule is applied to test the stability of the CSA. Then, the adaptive PCS are introduced and discussed in Sec. V. Furthermore, their performances are compared and the influence of the CSA is discussed. Finally, conclusions are drawn in Sec. VI.

Manuscript received DD MMMMMM YYYY; revised DD MMMMMM, YYYY; accepted DD MMMMMM YYYY. Date of publication DD MMMMMM YYYY; date of current version DD MMMMMM YYYY. This research was funded by the Austrian Science Fund (FWF) under grant P33702-N. The authors gratefully thank Lisa Schönenberger for providing valuable feedback.

The authors are with the Research Center Business Informatics, Vorarlberg University of Applied Sciences, 6850 Dornbirn, Austria, with the Institute of Theoretical Computer Science, Ulm University, 89069 Ulm, Germany, and also with the Institute of Medical Systems Biology, Ulm University, 89069 Ulm, Germany (e-mail: amir.omeradzic@hotmail.com, hans-georg.beyer@fhv.at).

This article has supplementary downloadable material and color versions of the figures available online at https://doi.org/xxxxxx/TEVC.xxxx.xxxxxxx.

Digital Object Identifier xxxxxxx/TEVC.xxxx.xxxxxxx

II. CSA-ES WITH ADAPTIVE POPULATION CONTROL

The multi-recombinative $(\mu/\mu_I, \lambda)$-CSA-ES with adaptive population control (PCS) is presented in Alg. 1. The parent and offspring population is denoted by μ and λ, respectively, with the truncation ratio $\vartheta := \mu/\lambda = 1/2$. The ES operates with isotropic mutations of strength σ. Algorithm 1 consists of three main parts. The initialization is performed between Lines 1-5. The standard CSA implementation is given in Lines 7-17 and is discussed below. PCS methods to be used are based on [8] (APOP), [6] (pcCSA), and [7] (PSA). Each method measures the ES performance (denoted by $\mathcal{P}$) by its own means (Lines 18-29), see Sec. V. Lines 21-29 control which single PCS is used throughout an optimization. The details of the respective population control subroutines are presented later in Alg. 2. The population is changed within the bounds $[\mu_{\min}, \mu_{\max}]$ via a factor $\alpha_\mu > 1$ using a simple schedule (Lines 31-37, 39). A generational idle (wait) time $\Delta_g > 0$ can be set (Lines 5, 30, 45). Rescaling of the mutation strength is performed in Line 41 and will be discussed in Sec. III. Before investigating PCS in Sec. V, adaptation and stability properties of the underlying CSA are studied in Secs. III and IV. The results will provide reasoning for the choice of adaptation time scales and mutation strength rescaling of PCS in Sec. V.

The cumulation path $\mathbf{s}$ of CSA is given in terms of cumulation constant c_σ and recombined mutation direction $\langle \mathbf{z} \rangle$

$$\mathbf{s}^{(g+1)} = (1-c_\sigma)\mathbf{s}^{(g)} + \sqrt{c_\sigma(2-c_\sigma)\mu}\langle \mathbf{z} \rangle^{(g+1)}. \tag{1}$$

Then, the path length $\left\|\mathbf{s}^{(g+1)}\right\|$ is measured and compared to its expected result under random selection. The first update rule for the σ-change is chosen according to [13]

$$\sigma^{(g+1)} = \sigma^{(g)} \exp\left[\frac{1}{D}\left(\left\|\mathbf{s}^{(g+1)}\right\|/E_\chi - 1\right)\right], \tag{2}$$

where D is a damping factor. Alternatively, a slightly different update rule [10, (44)] with damping d_σ yields

$$\sigma^{(g+1)} = \sigma^{(g)} \exp\left[\frac{c_\sigma}{d_\sigma}\left(\left\|\mathbf{s}^{(g+1)}\right\|/E_\chi - 1\right)\right], \tag{3}$$

which is usually chosen as the default CSA within the CMA-ES [10]. E_χ is the expected value of a chi-distributed random variate $\chi \sim \|\mathcal{N}(\mathbf{0},\mathbf{1})\|$ with N degrees of freedom. One usually approximates $E_\chi \simeq \sqrt{N}\big(1 - 1/4N + 1/21N^2\big)$ for large N [10]. The CSA variants under investigation are

$$\text{Update (2)} \quad \text{with} \quad c_\sigma = 1/\sqrt{N}, \quad D = c_\sigma^{-1} \tag{4a}$$

$$\text{Update (2)} \quad \text{with} \quad c_\sigma = 1/N, \quad D = c_\sigma^{-1} \tag{4b}$$

$$\begin{aligned} &\text{Update (3)} \quad \text{with} \quad c_\sigma = \frac{\mu+2}{N+\mu+5}, \quad \text{and} \\ &d_\sigma = 1 + c_\sigma + 2\max\left(0, \sqrt{(\mu-1)/(N+1)} - 1\right). \end{aligned} \tag{4c}$$

CSA implementations (4a) and (4b) were investigated in more detail in [11] and [13]. [13] derives the inverse proportionality $D = c_\sigma^{-1}$ with $c_\sigma = N^{-a}$ for $\frac{1}{2} \le a \le 1$ based on theoretical and experimental investigations on the sphere ($\mu \ll N$, see also (19)). CSA (4c) was defined to improve step-size adaptation properties with active CMA. In the next section, steady-state adaptation properties of the CSA variants are discussed on the sphere function.

Algorithm 1 Population Size Control via $(\mu/\mu_I, \lambda)$-CSA-ES

1: $g \leftarrow 0$
2: initialize_CSA$(\mathbf{y}^{(0)}, \mathbf{s}^{(0)}, \sigma^{(0)}, c_\sigma, d_\sigma, D, E_\chi)$
3: initialize_PCS_generic$(\mu^{(0)}, \mu_{\min}, \mu_{\max}, \vartheta, \alpha_\mu, \Delta_g)$
4: initialize_PCS_specific$(\beta, L, f_{\mathrm{rec}}^{(0)}, f_{\mathrm{med}}^{(0)}, \mathbf{p}_m^{(0)}, \mathbf{p}_c^{(0)})$
5: $w \leftarrow \Delta_g$
6: **repeat**

Standard $(\mu/\mu_I, \lambda)$-CSA-ES

7: **for** $l = 1, ..., \lambda$ **do**
8: $\tilde{\mathbf{z}}_l \leftarrow [\mathcal{N}(0,1), ..., \mathcal{N}(0,1)]$
9: $\tilde{\mathbf{y}}_l \leftarrow \mathbf{y}^{(g)} + \sigma^{(g)}\tilde{\mathbf{z}}_l$
10: $\tilde{f}_l \leftarrow f(\tilde{\mathbf{y}}_l)$
11: **end for**
12: $(\tilde{f}_{1;\lambda}, ..., \tilde{f}_{m;\lambda}, ..., \tilde{f}_{\mu;\lambda}) \leftarrow \mathrm{sort}(\tilde{f}_1, ..., \tilde{f}_\lambda)$
13: $\mathbf{y}^{(g+1)} \leftarrow \frac{1}{\mu}\sum_{m=1}^{\mu} \tilde{\mathbf{y}}_{m;\lambda}$
14: $\langle \mathbf{z} \rangle^{(g+1)} \leftarrow \frac{1}{\mu}\sum_{m=1}^{\mu} \mathbf{z}_{m;\lambda}$
15: $\mathbf{s}^{(g+1)} \leftarrow (1-c_\sigma)\mathbf{s}^{(g)} + \sqrt{\mu^{(g)} c_\sigma(2-c_\sigma)}\langle \mathbf{z} \rangle^{(g+1)}$
16: CSA (2): $\sigma^{(g+1)} \leftarrow \sigma^{(g)} \exp\left[\frac{1}{D}\left(\|\mathbf{s}^{(g+1)}\|/E_\chi - 1\right)\right]$
17: CSA (3): $\sigma^{(g+1)} \leftarrow \sigma^{(g)} \exp\left[\frac{c_\sigma}{d_\sigma}\left(\|\mathbf{s}^{(g+1)}\|/E_\chi - 1\right)\right]$

Population Control Strategy (PCS)

18: $f_{\mathrm{rec}}^{(g+1)} \leftarrow f(\mathbf{y}^{(g+1)})$
19: $f_{\mathrm{med}}^{(g+1)} \leftarrow \mathrm{median}(\tilde{f}_{1;\lambda}, ..., \tilde{f}_{m;\lambda}, ..., \tilde{f}_{\mu;\lambda})$
20: $g_0 \leftarrow g - L + 1$
21: **if** PCS = APOP **and** $g_0 \ge 0$ **then** ▷ Evaluate PCS
22: $\mathcal{P} \leftarrow$ get_apop()
23: **else if** PCS = pcCSA **and** $g_0 \ge 0$ **then**
24: $\mathcal{P} \leftarrow$ get_pccsa()
25: **else if** PCS = PSA **then**
26: $\mathcal{P} \leftarrow$ get_psa()
27: **else**
28: $\mathcal{P} \leftarrow 0$
29: **end if**
30: **if** $w = 0$ **then** ▷ No wait: update population size
31: **if** $\mathcal{P} < 0$ **then**
32: $\mu^{(g+1)} \leftarrow \lceil \alpha_\mu \mu^{(g)} \rceil$ ▷ Increase
33: **else if** $\mathcal{P} > 0$ **then**
34: $\mu^{(g+1)} \leftarrow \lfloor \mu^{(g)}/\alpha_\mu \rfloor$ ▷ Decrease
35: **else**
36: $\mu^{(g+1)} \leftarrow \mu^{(g)}$ ▷ No change
37: **end if**
38: **if** $\mu^{(g)} \ne \mu^{(g+1)}$ **then** ▷ Parameter updates
39: $\mu^{(g+1)} \leftarrow \min\left(\max(\mu^{(g+1)}, \mu_{\min}), \mu_{\max}\right)$
40: $w \leftarrow \Delta_g$
41: $\sigma^{(g+1)} \leftarrow \sigma^{(g)} r_\sigma(\mu^{(g)}, \mu^{(g+1)})$
42: $\mu \leftarrow \mu^{(g+1)}, \quad \lambda \leftarrow \mathrm{round}(\mu^{(g+1)}/\vartheta)$
43: $c_\sigma \leftarrow c_\sigma(\mu), \ \ D \leftarrow D(\mu), \ \ d_\sigma \leftarrow d_\sigma(\mu)$
44: **end if**
45: **else** ▷ Wait: no update
46: $w \leftarrow w - 1$
47: **end if**
48: $g \leftarrow g + 1$
49: **until** termination criterion

III. CSA-ES FOR LARGE POPULATION SIZES

A. Scaling of Generation Number and Mutation Strength

In this section, progress rate theory on the sphere function $f(R) = R^2$ with $R = ||\mathbf{y}||$ and $\mathbf{y} \in \mathbb{R}^N$ is introduced to derive laws for the generation number scaling and mutation strength rescaling. To this end, the scale invariance of the sphere will be utilized through the introduction of appropriate normalizations.

Furthermore, large population sizes $\mu \gg \sqrt{N}$ will be assumed to simplify the progress rate and enable closed-form solutions. The subsequent derivation of the generation number as a function of the progress rate was already given in [14, Sec. 2.4]. The progress rate φ is defined as the expected residual distance change between two generations g and $g+1$ as

$$\varphi^{(g)} := R^{(g)} - \mathrm{E}\left[R^{(g+1)}\right]. \tag{5}$$

Due to the scale invariance, one can define the normalized (denoted by *) progress rate and mutation strength as

$$\varphi^* = \varphi N/R, \quad \sigma^* = \sigma N/R. \tag{6}$$

A properly working σ-adaptation, such as the CSA, attains constant σ^* and φ^* in expectation on the sphere, giving linear convergence order [14]. Assuming constant φ^*, one can derive the generation number $G := g - g_0$ for a given relative change $R^{(g)}/R^{(g_0)}$ under the assumption $\varphi^*/N \ll 1$ as [14, Sec. 2.4]

$$G = (N/\varphi^*)\ln\left(R^{(g_0)}/R^{(g)}\right). \tag{7}$$

The next step is to derive φ^* for large populations μ in relation to N. One can start with the normalized progress rate of the sphere derived in [14, (6.54)]

$$\begin{aligned}\varphi^* = {} & \frac{c_{\mu/\mu,\lambda}\sigma^*(1+\sigma^{*2}/2\mu N)}{\sqrt{1+\sigma^{*2}/\mu N}\sqrt{1+\sigma^{*2}/2N}} \\ & - N\left(\sqrt{1+\sigma^{*2}/\mu N} - 1\right) + O\left(N^{-1/2}\right).\end{aligned} \tag{8}$$

Coefficient $c_{\mu/\mu,\lambda}$ in (8) is obtained by numerical solving of

$$c_{\mu/\mu,\lambda} = \frac{\lambda-\mu}{2\pi}\binom{\lambda}{\mu}\int_{-\infty}^{\infty} \mathrm{e}^{-t^2}\left[\Phi(t)\right]^{\lambda-\mu-1}\left[1-\Phi(t)\right]^{\mu-1}\mathrm{d}t\,, \tag{9}$$

with Φ denoting the normal cumulative distribution function. The numerically obtained non-trivial zero of (8) will be denoted as σ_0^*. An approximation thereof is derived now. For sufficiently large $\mu \gtrapprox 20$, $c_{\mu/\mu,\lambda} \simeq c_\vartheta$ is only a function of the truncation ratio ϑ with [14, (6.113)]

$$c_\vartheta = \mathrm{e}^{-\frac{1}{2}\left[\Phi^{-1}(\vartheta)\right]^2}/\sqrt{2\pi}\vartheta. \tag{10}$$

Given progress rate (8), an approximation for large population sizes is derived. As a first step, one assumes $\mu N \gg \sigma^{*2}$, such that a Taylor-expansion in (8) yields $\sqrt{1+\sigma^{*2}/\mu N} = 1+\sigma^{*2}/2\mu N + O((\sigma^{*2}/\mu N)^2)$. Neglecting higher order terms, one gets a simplified expression

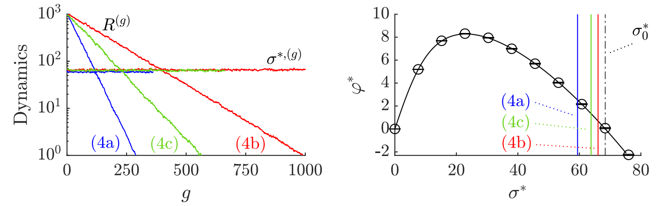
$$\varphi^* \simeq c_\vartheta\sigma^*/\sqrt{1+\sigma^{*2}/2N} - \sigma^{*2}/2\mu. \tag{11}$$

For increasing μ at fixed N, one observes increasingly higher levels of the normalized mutation strength σ^* on the sphere. Hence, we simplify (11) by assuming $\sigma^{*2}/2N \gg 1$, such that "1" is neglected within the square-root. One gets

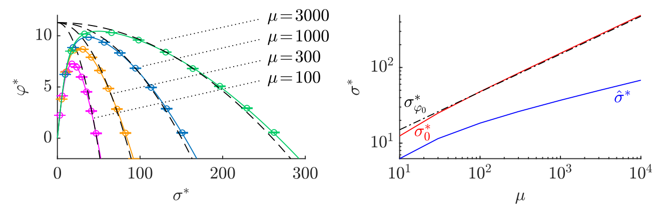
$$\varphi^* \simeq \sqrt{2N}c_\vartheta - \sigma^{*2}/2\mu. \tag{12}$$

The zero of approximation (12) is easily obtained as

$$\sigma^*_{\varphi_0} = (8N)^{1/4}(c_\vartheta\mu)^{1/2}. \tag{13}$$



(a) On the left, the median dynamics of $(200/200_I, 400)$-CSA-ES for $N = 100$ is shown (10 trials) for (4a), (4b) and (4c). On the right, $\varphi^*(\sigma^*)$ is shown (data points: (5) simulated over 10^4 trials, normalized; black curve: (8)). Furthermore, the solid vertical lines mark measured median values from the left, showing $\sigma^*_{\mathrm{ss}} \approx 59.4, 63.8, 66.1$ with $\gamma = \sigma^*_{\mathrm{ss}}/\sigma^*_0 \approx 0.87, 0.93, 0.97$, from left to right, with numerically obtained $\sigma^*_0 \approx 68.4$ (dash-dotted black). One can determine $\hat{\sigma}^* = \arg\max\varphi^*(\sigma^*) \approx 23$ numerically from (8).



(b) On the left, $\varphi^*(\sigma^*)$ at $N = 100$ for varying μ ($\mu/\lambda = 1/2$) is displayed. The solid lines show (8) and the corresponding data points (5) averaged over 10^4 trials and normalized using $\varphi^* = \varphi N/R$. The dashed lines display (12). On the right, (8) is used to numerically calculate the second zero σ^*_0 (solid red) and $\hat{\sigma}^*$ (solid blue). The black dash-dotted line shows approximation (13).

Fig. 1: Progress rate φ^* and scale-invariant σ^* on the sphere.

Given (13), positive progress $\varphi^* > 0$ on the sphere is expected to occur for $\sigma^* \in (0, \sigma^*_{\varphi_0})$. Setting $\sigma^* = \sigma^*_{\varphi_0}$ for the applied conditions $\mu N \gg \sigma^{*2}_{\varphi_0}$ and $\sigma^{*2}_{\varphi_0} \gg 2N$, respectively, one gets

$$\sqrt{N} \gg \sqrt{8}c_\vartheta \quad \text{and} \quad \mu \gg \sqrt{N/2c_\vartheta^2}. \tag{14}$$

Since c_ϑ is of the order of one, sufficiently large populations are demanded with $\mu \gg \sqrt{N}$ and $N \gg 8$.

At this point, exemplary experiments regarding φ^* and σ^* are performed. In Fig. 1a, exemplary dynamics of $R^{(g)}$ and $\sigma^{*,(g)}$ are displayed on the left. Linear convergence order of R is observed, while a steady-state $\sigma^* = \sigma^*_{\mathrm{ss}}$ is achieved in expectation. On the right, the respective progress rate (8) is shown together with one-generation experiments thereof. One observes excellent agreement of (8) with one-generation experiments of (5). Furthermore, the measured σ^*_{ss} are marked. The CSAs achieve σ^*_{ss}-levels which are close to the (second) progress rate zero. While the theoretically optimal value $\hat{\sigma}^* = \arg\max\varphi^*(\sigma^*)$ can only be determined numerically from (8), an analytic approximation for the second zero exists in (13). Figure 1b shows (8) and (12) on the left, while second zero and optimal $\hat{\sigma}^*$ are shown on the right. Note that approximation (12) improves with increasing μ, which was expected from the underlying assumptions. For approximation (13), one observes good agreement of $\sigma^*_{\varphi_0}$ with σ^*_0 for increasing μ. Since the CSA operate sufficiently slow, i.e., in the vicinity $\sigma^*_{\mathrm{ss}} \lessapprox \sigma^*_{\varphi_0}$, the CSA steady-state σ^*_{ss} will be modeled in terms of a scaling parameter $0 < \gamma < 1$ (slow-adaptation: $\gamma \lessapprox 1$), such that

$$\sigma^*_{\mathrm{ss}} = \gamma\sigma^*_{\varphi_0} = \gamma(8N)^{1/4}(c_\vartheta\mu)^{1/2}. \tag{15}$$

There are two main reasons for modeling σ^*_{ss} in terms of the second zero $\sigma^*_{\varphi_0}$. First, there is no closed-form solution for $\hat{\sigma}^*$ from (8), while (12) yields the optimal value $\hat{\sigma}^* = 0$, which is useless. Second, $\hat{\sigma}^*$ decreases in relation to $\sigma^*_{\varphi_0}$ with increasing μ, see Fig. 1b. For optimization under multimodality, relatively high σ^*-levels are desirable to improve the success rate for global convergence, see [3] for an analysis of the Rastrigin function.

In general, the steady-state γ varies to some extent with μ, N, and the CSA parameters c_σ and D (or d_σ). The derivation of G in (17) and σ-rescaling in (18) require the assumption that γ is independent of both μ and N. This simplification is supported by experimental results in Sec. III-B, where $\gamma(\mu, N)$ is investigated in greater detail for the given CSA variants. Inserting σ^*_{ss} from (15) into (12), one obtains

$$\varphi^* \simeq c_\vartheta (2N)^{1/2}(1-\gamma^2). \tag{16}$$

Thereafter, one can insert (16) into (7), which yields

$$G \simeq \sqrt{N}\frac{\ln\big(R^{(g_0)}/R^{(g)}\big)}{\sqrt{2}c_\vartheta(1-\gamma^2)}. \tag{17}$$

For constant γ and sufficiently large $\mu \gg \sqrt{N}$, the generation number approaches a $\sqrt{N}$-law independent of μ, given a relative change in R. This is a remarkable result which is tested later in experiments. It is also useful when defining a time scale for the adaptation of PCS in Secs. IV and V.

Various PCS employ additional corrections of σ after μ has been updated [7], [8], see also Line 41 of Alg. 1. ES with different population sizes operate at varying mutation strength levels. On the sphere, one observes distinct σ^*_{ss}-levels depending on μ. This observation suggests systematic rescaling of σ when μ is changed to faster reach the new steady-state. However, these additional changes in σ can have effects on the stability of the CSA, see also Sec. IV. Given the results on the sphere in (15) with $\sigma^* = \sigma N/R$, we derive a $\sqrt{\mu}$-law for the σ-rescaling $r_\sigma := \sigma^{(g+1)}/\sigma^{(g)}$. Since the distance to the optimizer does not change during σ-rescaling, one has $R^{(g+1)} = R^{(g)}$. Assuming a constant γ in (15) with $K := \gamma(8N)^{1/4}c_\vartheta^{1/2}$, one has $\sigma^{(g)} = \sigma^{*,(g)}_{\text{ss}} R^{(g)}/N = K\sqrt{\mu^{(g)}}R^{(g)}/N$ and $\sigma^{(g+1)} = \sigma^{*,(g+1)}_{\text{ss}} R^{(g+1)}/N = K\sqrt{\mu^{(g+1)}}R^{(g+1)}/N$. Thus, one gets $r_\sigma = \sqrt{\mu^{(g+1)}/\mu^{(g)}}$ and

$$\sigma^{(g+1)} = \sigma^{(g)}\sqrt{\mu^{(g+1)}/\mu^{(g)}}. \tag{18}$$

Alternatively, in the case of $N \to \infty$ and $\mu \ll N$, (8) can be used to derive the well-known progress rate

$$\varphi^* = c_{\mu/\mu,\lambda}\sigma^* - \sigma^{*2}/2\mu. \tag{19}$$

In this case, the optimal value and the second zero of (19) are $\hat{\sigma}^* = c_{\mu/\mu,\lambda}\mu$ and $\sigma^*_{\varphi_0} = 2c_{\mu/\mu,\lambda}\mu$, respectively. Based on these results, one obtains mutatis mutandis a linear scaling law for the σ-rescaling as

$$\sigma^{(g+1)} = \sigma^{(g)}\mu^{(g+1)}/\mu^{(g)}. \tag{20}$$

While (20) is not expected to yield better results (since $\mu \ll N$), it will still be tested as it was applied in [7] (see supplementary material A-A).

B. Analysis of the CSA-ES

In this section, the steady-state analysis of (4a), (4b), and (4c) is recapitulated for the sphere. The CSA-ES has already been investigated in [15] for the three CSA variants. Hence, only the main results will be stated. The goal is to first investigate γ from (15) and later apply the results to obtain G from (17).

As a first step, basic adaptation properties of the three variants are identified. Note that c_σ and D of (4a) and (4b) are independent of μ. Changing the population size should not have a large influence on the adaptation characteristics. For (4c), c_σ and d_σ require further analysis under the assumption $\mu \gg N$. c_σ yields simply

$$c_\sigma = \frac{\mu(1+2/\mu)}{\mu(1+N/\mu+5/\mu)} \overset{\mu\to\infty}{\simeq} 1. \tag{21}$$

The cumulation constant c_σ approaches "1" as μ is increased, which results in a faster cumulation in (1). Now, the factor c_σ/d_σ is brought into a similar form as $1/D$ in (2). One finds in [15] that c_σ/d_σ yields a damping factor $D = 1 + 1/c_\sigma + g(N,\mu)/c_\sigma$ with $g(N,\mu) := 2\max\left(0, \sqrt{\frac{\mu-1}{N+1}} - 1\right)$. Hence, the damping term $g(N,\mu)/c_\sigma$ yields the scaling

$$g(N,\mu)/c_\sigma \simeq 2\sqrt{\mu/N} \quad (\mu\to\infty). \tag{22}$$

The resulting damping D of (4c) scales with $\sqrt{\mu}$ according to (22) for fixed N and c_σ approaches one with (21). Hence, CSA (4c) employs a μ-dependent damping which is in contrast to (4a) and (4b). If μ is changed during adaptive population control, the CSA adaptation characteristic in terms of γ will be affected, which is shown later in Fig. 2.

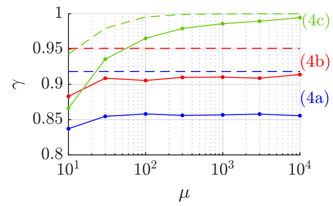
In [15], the three CSA variants are investigated and the CSA update equations are expressed in the sphere steady-state, where σ^* and φ^* are constant in expectation (see Fig. 1a). By applying certain approximations, i.e., by assuming sufficiently slow progress $\varphi^*/N \ll 1$ and large dimensionality N, this approach enables a closed-form solution of the CSA sphere steady-state as a function of the given cumulation constant and damping. By including this dependency into a single constant b [15, (58)] with

$$b := \left(c_\sigma D/(1-c_\sigma) + \sqrt{2}c_\vartheta D/\sqrt{N}\right)^{-1}, \tag{23}$$

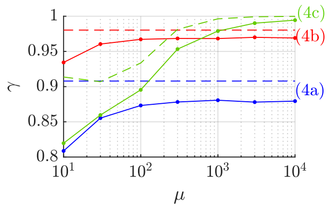
one can derive the respective γ from (15) as a function of b (assuming $1/\sqrt{2} < \gamma < 1$) as [15, (57)]

$$\gamma = \sqrt{\frac{1}{2}\left(\sqrt{1+b^2} - b + 1\right)}. \tag{24}$$

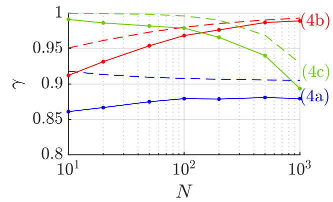
The result (24) is interesting as it relates the CSA parameters and N to a scale factor γ on the left side. Demanding γ to be constant and independent of μ and N, one must choose $D \propto \sqrt{N}$ and $c_\sigma = D^{-1}$ within b (this holds with $O(1/\sqrt{N})$). This is only fulfilled by CSA (4a). In this case, the analytic solution of $\gamma \approx 0.90$ ($\mu/\lambda = 1/2$) was derived as an approximation [15, (63)]. For (4b) one has $D \propto N$, and for (4c) D scales with $\sqrt{\mu/N}$ via (22). The limit $D \to \infty$ ($0 < c_\sigma < 1$) yields $b \to 0$ and $\gamma \to 1$. In this limit, $\sigma^*_{\text{ss}} \to \sigma^*_{\varphi_0}$ in (15) and the progress rate vanishes $\varphi^* \to 0$ by approaching its second zero.



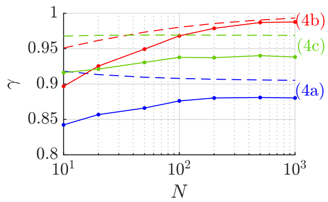
(a) $N = 10$.



(b) $N = 100$



(c) $\mu = 1000$.



(d) $\mu = 2N$.

Fig. 2: Steady-state ratio γ on the sphere function for $\vartheta = 1/2$. Measured ratio $\sigma^*_{\text{ss}}/\sigma^*_0$ (solid, with dots) compared to γ from (24) (dashed) for the CSA variants (4a) (blue), (4b) (red), and (4c) (green).

Hence, (4b) becomes increasingly slow for large N, while (4c) becomes slower for increasing ratio $\sqrt{\mu/N}$.

In Fig. 2, experiments on the sphere are shown to compare the CSA variants with the predicted adaptation behavior in terms of γ. To this end, the dimensionality N and population size μ are varied. The measured steady-state σ^*_{ss} is averaged over at least 10 trials and the median is taken over the measured $\sigma^{*,(g)}$ (the median is necessary due to a slightly skewed distribution of σ^* at small N). The measured σ^*_{ss} is normalized by σ^*_0 (numerically obtained second zero of (8)[1]), yieding a reference value for γ. In Figs. 2a and 2b, the measured γ remains relatively constant for sufficiently large μ for (4a) and (4b). CSA (4c) shows a significant increase of γ as μ increases, which was expected from its damping D. Deviations between the predicted γ (dashed) and measurement (solid) are expected to occur due to the underlying approximations applied in [15]. Hence, best agreement is obtained for vanishing $\varphi^* \to 0$ in the limit $\gamma \to 1$. In Fig. 2c, the dimensionality is varied. CSA (4a) remains relatively constant, while γ of (4b) increases for larger N, which was expected from its damping $D \propto N$. CSA (4c) shows a decreasing γ due to $D \propto \sqrt{\mu/N}$. In Fig. 2d, both μ and N are varied together by maintaining $\mu = 2N$. CSA (4a) and (4c) remain approximately constant, while for (4b) $\gamma \to 1$. As expected, only (4a) maintains an approximately constant ratio with $\gamma \lessapprox 0.9$, which agrees satisfactory with the prediction (24). Furthermore, it realizes lower γ-levels, which lead to higher progress rates φ^* (cf. Fig. 1a). Note that all three CSA yield relatively large γ-values between 0.8 and 1. This means they achieve comparably low progress rates. However, under high multimodality, slower adaptation is usually beneficial to achieve higher success rates [2], [4].

[1]For $N \gtrapprox 100$ one may numerically calculate σ^*_0 using (8). For $N < 100$ the progress rate zero is calculated from one-generation experiments of (5) to achieve higher accuracy. Slow adaptation yields σ^*_{ss} very close to σ^*_0 such that the accuracy of (8) is reduced due to missing $O\big(N^{-1/2}\big)$-terms.

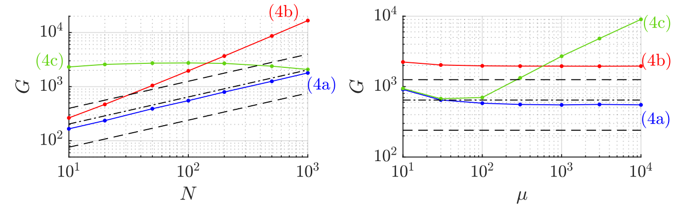


Fig. 3: Number of generations G, see (17), to reach the target $R^{(g)}/R^{(g_0)} = 10^{-6}$ using a $(1000/1000_I, 2000)$-ES (left) and $N = 100$ (right). The average over 10 runs is taken. The solid lines show (4a) in blue, (4b) in red, and (4c) in green. G from (17) at $\gamma = 0.9$ is displayed in dash-dotted black. The lower and upper dashed lines show $\gamma = 0.8, 0.95$, respectively.

Now that the results for γ have been investigated in Fig. 2, one can study the generation number G from (17). Figure 3 shows G evaluated for varying N (left) and varying μ (right). CSA (4a) yields very good agreement with (17) due to the relatively constant γ. On the left, one observes a $\sqrt{N}$-law for G. On the right, G remains asymptotically constant for increasing μ. CSA (4b) shows a faster increase of G than $\sqrt{N}$ (left), which can be attributed to the increased damping ($\gamma \to 1$) at large N. On the right, it also remains asymptotically constant for large μ, requiring more generations due to slower adaptation. CSA (4c) shows a different characteristic, being comparably slow at high ratios μ/N and fast at low ratios due to its damping D.

The G-asymptotic for $\mu \to \infty$ is a notable result. From a modeling perspective, one could argue that CSA (4a) shows important (desired) properties for PCS on sphere-like functions. The adaptation is comparably fast and one observes a generational speedup of G as μ is increased from small initial values. At large μ, asymptotic behavior is observed. CSA (4c) slows down its adaptation as μ is increased, which can be detrimental in terms of efficiency. This is discussed later in Fig. 9.

It is important to emphasize that the global step-size adaptation properties are different with CMA. CSA (4c) is the standard implementation for the default CMA-ES [10]. An exemplary evaluation is displayed in the supplementary material A-B. If $\mu \gg N$ with CMA, one observes that CMA contributes significantly to the reduction of the overall global step-size compared to CSA. This results in significantly faster convergence for $\mu \gg N$ compared to Fig. 3. As μ is decreased, the contribution of CMA to the global step-size reduces accordingly and σ-decrease via CSA becomes more pronounced. The analysis of this work does not consider the effects of CMA. It relies solely on progress rate theory on the sphere where a constant scale-invariant σ^* is achieved. Hence, a more detailed discussion of CMA will be left for future research.

Note that the present work provides a steady-state analysis and does not study initialization effects of the CSA variants. Under certain conditions, especially with small initial $\sigma^{(0)}$ and very large μ, CSA (4a) and (4b) may exhibit an undesired large σ-increase during the initial optimization phase [16]. Therefore, in real-world applications, it is advisable to test the optimization robustness using multiple initializations.

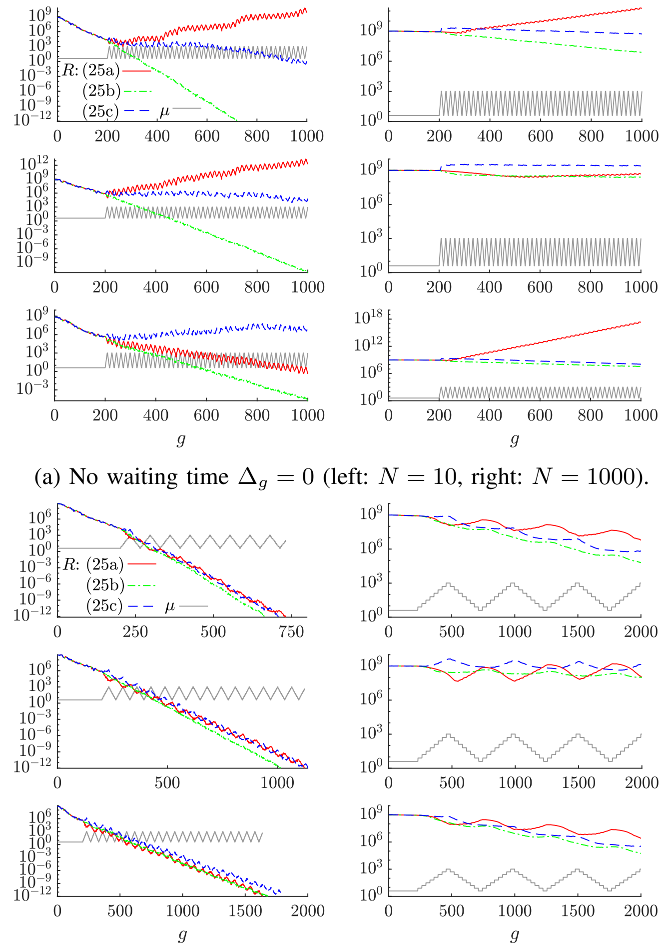
IV. Population Control Schedule

Before adaptive population control is investigated further, the CSA stability w.r.t. changes of μ is investigated for all three variants (4a), (4b), and (4c). The PCS will later introduce additional method-specific feedback in the ES dynamics. Having effects of σ- and μ-adaptation combined makes the analysis more difficult. Hence, we introduce a simple $\mu^{(g)}$-schedule on the sphere function, generating artificial changes in μ. The schedule is defined as being constant within the first 200 generations. Then, μ is increased using a factor $\alpha_\mu > 1$ as $\mu^{(g+1)} = \lceil \mu^{(g)}\alpha_\mu \rceil$ until $\mu_{\max}$ is reached. After reaching $\mu_{\max}$, it is decreased as $\mu^{(g+1)} = \lfloor \mu^{(g)}/\alpha_\mu \rfloor$ until $\mu^{(0)}$ is reached. This oscillation is repeated multiple times to study the stability of CSA. Furthermore, the effect of σ-rescaling (via $\sigma^{(g+1)} = r_\sigma \sigma^{(g)}$) on the stability is investigated. Given results (18) and (20), one can summarize the three tested rescaling methods as

$$r_\sigma = 1 \quad \text{(no rescaling)} \tag{25a}$$
$$r_\sigma = (\mu^{(g+1)}/\mu^{(g)})^{1/2} \tag{25b}$$
$$r_\sigma = \mu^{(g+1)}/\mu^{(g)}. \tag{25c}$$

The experiments in Figs. 4a and 4b display the convergence dynamics on the sphere function. Two different values for N are tested based on Fig. 2, where large differences for $\gamma(N)$ could be observed. In both experiments, $\alpha_\mu = 2$ is chosen with $\mu^{(0)} = \mu_{\min} = 4$ and $\mu_{\max} = 1024$. The dynamics are initialized at $\mathbf{s}^{(0)} = \mathbf{1}$ and $\sigma^{(0)} = \sigma^*_{\varphi_0} R/N$ using (13), which reduces undesired initialization effects. In Fig. 4b, a waiting time $\Delta_g = \lceil \sqrt{N} \rceil$ is applied after a population change based on the result $G \propto \sqrt{N}$ from (17). This introduces additional time for the CSA to adapt to the μ-change. Note that no waiting is implemented in Fig. 4a.

In Fig. 4a, one observes that no rescaling ($r_\sigma = 1$) leads to instabilities (divergence) of the CSA in most cases, while $r_\sigma = (\mu^{(g+1)}/\mu^{(g)})^{1/2}$ shows very good results in terms of convergence speed for all configurations. It also exhibits the least amount of $R^{(g)}$-oscillations along its convergence. Scaling $r_\sigma = \mu^{(g+1)}/\mu^{(g)}$ shows mixed results. It helps to stabilize the CSA in some cases. However, it would be more suitable for μ-control with $\mu \ll N$ due to progress rate (19). The instabilities are related to the cumulation constant c_σ, the damping D, and the applied r_σ. Due to the high rate of change in μ, CSA is constantly adapting to the changing population size. In the worst case, one observes divergence due to σ^* being outside the positive progress range $\sigma^* \in (0, \sigma^*_{\varphi_0})$, see Fig. 1a. Note that CSA (4a) shows consistent results as N is increased. On the other hand, CSA (4c) converges with $r_\sigma = 1$ due to the higher damping at small $N = 10$ (although with large oscillations of R), but diverges due to lower damping at large $N = 1000$. In Fig. 4b, the waiting time $\Delta_g = \lceil \sqrt{N} \rceil$ provides more time for the CSA to adapt accordingly. Again, (25b) shows the most stable results. As expected, $\Delta_g > 0$ stabilizes all CSA variants and for all r_σ-choices. The obtained results show that (25b) is the preferred rescaling of σ on the sphere in terms of stability. In Sec. V, it will be shown that



(a) No waiting time $\Delta_g = 0$ (left: $N = 10$, right: $N = 1000$).

(b) Waiting time $\Delta_g = \lceil \sqrt{N} \rceil$ (left: $N = 10$, right: $N = 1000$).

Fig. 4: Stability of CSA on the sphere using predefined $\mu^{(g)}$-schedule. The dynamics of R and μ are displayed for (25a), (25b), and (25c), see legend. The tested CSAs are (4a), (4b), and (4c), from top to bottom, respectively for (a) and (b).

certain PCS are more sensitive to changes in σ than others. The influence of r_σ with active PCS will also be studied.

V. Adaptive Population Control Strategies

For the subsequent investigations, three state-of-the-art PCS are implemented based on [8] (APOP), [6] (pcCSA), and [7] (PSA). These algorithms were designed and tested for adaptive population control on ES for a broader range of functions (noisy and/or multimodal, details below). As for the analysis, the idea is to study simplified versions of the PCS. Hence, only the performance measuring routines are implemented. Note that the population control subroutines of Alg. 1 are given in Alg. 2, including the pseudocode for APOP, pcCSA, and PSA. A simple routine for the μ-update is realized according to Alg. 1 (Lines 31-37). We omit the introduction of additional damping parameters for the μ-change (as done in [7], [8], analogous to D in (2)) since it adds more complexity. Furthermore, (25b) is used for the σ-rescaling of all methods.

APOP was introduced in [8] for the CMA-ES and tested on a larger set of noiseless functions. The basic idea is to count the number of (median-)fitness deteriorations within the last L generations (in [8] they use a fixed length $L = 5$). Defining the median over selected f-values as $f_{\text{med}}^{(g)} := \text{median}(f_{m;\lambda}^{(g)})$, $m = 1, ..., \mu$, the difference is given by

$$\Delta f^{(g)} = f_{\text{med}}^{(g)} - f_{\text{med}}^{(g-1)}. \tag{26}$$

Then, the ratio P_f is evaluated by counting the occurrence of $\Delta f^{(g)} > 0$. There are $L - 1$ differences for L values of $f_{\text{med}}^{(g)}$. Using g_0 (Line 20) and the indicator function $\mathbb{1}$, one has

$$P_f = \frac{1}{L-1} \sum_{i=g_0+1}^{g} \mathbb{1}[\Delta f^{(i)} > 0]. \tag{27}$$

A threshold $\mathcal{T}_f = 1/5$ is chosen in [8] based on empirical studies and will be adopted. $P_f > \mathcal{T}_f$ triggers a population increase due to insufficient performance. For $P_f = \mathcal{T}_f$ no change occurs, and for $P_f < \mathcal{T}_f$ the population is decreased. As for all the PCS considered here, the respective threshold value $\mathcal{T}$ has significant influence on the performance of the PCS. Furthermore, exchanging the CSA will lead to notable performance differences on simple test functions.

The pcCMSA was introduced in [6] and further analyzed in [17]. It was originally introduced as a covariance matrix self-adaptive ES for population control on noisy functions and later tested on multimodal functions in [7]. In this paper, only the hypothesis test for convergence is implemented using a CSA, calling it pcCSA. In [6], it is argued that stagnation or divergence behavior coincides with a non-negative trend within the observed fitness value dynamics of the ES (for minimization). For a trend analysis, a regression model of the parental recombinant fitness sequence $f(\mathbf{y}^{(g)})$ of length L is used and a hypothesis test on the slope is performed. The (fluctuating) fitness dynamics f is modeled assuming a linear model with slope a, intercept b, and random normal fluctuations ϵ_i as $f^{(g)} = ag + b + \epsilon_i$. Denoting the standard error of the estimated slope $\hat{a}$ as $s_{\hat{a}}$, the (standardized) test statistic is a t-distributed variate with $L-2$ degrees of freedom

$$T_{L-2} \sim (\hat{a} - a)/s_{\hat{a}}. \tag{28}$$

The hypothesis test is defined as H_0: $a \geq 0$, indicating no significant trend and insufficient performance. The alternative H_1: $a < 0$ indicates sufficient performance. We will evaluate the P-value of the test ($P_{t,L-2}$ denoting the distribution function of T with $L - 2$ degrees of freedom)

$$P_H := P_{t,L-2}(\hat{a}/s_{\hat{a}}), \tag{29}$$

rejecting H_0 at a significance level $P_H < 0.05$ ($\mathcal{T}_H = 0.05$). In this case, μ is decreased, or otherwise increased.

The core ideas of the PSA-CMA-ES were introduced in [18] and further extended in [7]. It is based on the idea of measuring the change of search space parameters, namely covariance matrix $\mathbf{C}$, mutation strength σ, and the mean search vector $\mathbf{y}$. By accumulating the information using two merged cumulation paths, the resulting length is used as an indicator of ES performance. The key element is to distinguish random selection (insufficient performance) from non-random selection (sufficient performance). The measured path length is used to change the population size, given a certain threshold.

Algorithm 2 Population Control Strategies (PCS)

1: Measure performance $\mathcal{P}$ in Alg. 1 by evaluation of:

get_apop()

2: $\mathbf{d} \leftarrow \text{diff}(f_{\text{med}}^{(g_0:g)})$ $\triangleright$ $g_0 = g - L + 1$ (Line 1)
3: $P_f \leftarrow \text{sum}(\mathbf{d} > 0)/(L-1)$
4: $\mathcal{P} \leftarrow \text{perf}(P_f < \mathcal{T}_f\text{: } 1; \quad P_f > \mathcal{T}_f\text{: } {-1}; \quad P_f = \mathcal{T}_f\text{: } 0)$

get_pccsa()

5: $\mathbf{g} \leftarrow [g_0 : g], \quad \mathbf{f} \leftarrow [f_{\text{rec}}^{(g_0:g)}]$ $\triangleright$ $g_0 = g - L + 1$ (Line 1)
6: $\bar{g} \leftarrow \text{mean}(\mathbf{g}), \quad \bar{f} \leftarrow \text{mean}(\mathbf{f})$
7: $\hat{a} \leftarrow \frac{\sum_{i=g_0}^{g}(i-\bar{g})(f_i-\bar{f})}{\sum_{i=g_0}^{g}(i-\bar{g})^2}$
8: $s_{\hat{a}} \leftarrow \sqrt{\frac{\sum_{i=g_0}^{g}(f_i-\bar{f})^2}{(L-2)\sum_{i=g_0}^{g}(i-\bar{g})^2}}$
9: $t \leftarrow \hat{a}/s_{\hat{a}}$
10: $P_H \leftarrow \text{tcdf}(t, L-2)$ $\triangleright$ distribution function of T_{L-2}
11: $\mathcal{P} \leftarrow \text{perf}(P_H < \mathcal{T}_H\text{: } 1; \quad P_H > \mathcal{T}_H\text{: } {-1}; \quad P_H = \mathcal{T}_H\text{: } 0)$

get_psa()

12: $E_F \leftarrow N/\mu^{(g)}$
13: $\widetilde{\Delta}_m \leftarrow \langle \mathbf{z} \rangle^{(g+1)}$
14: $\widetilde{\Delta}_c \leftarrow \frac{1}{\sqrt{2}}((\sigma^{(g+1)}/\sigma^{(g)})^2 - 1)\mathbf{1}$
15: $\mathbf{p}_m^{(g+1)} \leftarrow (1-\beta)\mathbf{p}_m^{(g)} + \sqrt{\beta(2-\beta)/E_F}\,\widetilde{\Delta}_m$
16: $\mathbf{p}_c^{(g+1)} \leftarrow (1-\beta)\mathbf{p}_c^{(g)} + \sqrt{\beta(2-\beta)/E_F}\,\widetilde{\Delta}_c$
17: $||\mathbf{p}_\theta^{(g+1)}||^2 \leftarrow ||\mathbf{p}_m^{(g+1)}||^2 + ||\mathbf{p}_c^{(g+1)}||^2$
18: $\mathcal{P} \leftarrow \text{perf}(\|\mathbf{p}_\theta\|^2 < \mathcal{T}_\theta\text{: } {-1}; \quad \|\mathbf{p}_\theta\|^2 > \mathcal{T}_\theta\text{: } 1; \quad \|\mathbf{p}_\theta\|^2 = \mathcal{T}_\theta\text{: } 0)$

In contrast to pcCSA and APOP, the performance is measured solely in search space and not in f-space [2]. The main idea is to define two cumulation paths for the change of the mean $\Delta\mathbf{m}^{(g+1)} = \mathbf{y}^{(g+1)} - \mathbf{y}^{(g)}$, and the change of $\mathbf{C}$ and σ as $\Delta\mathbf{\Sigma}^{(g+1)} = (\sigma^{(g+1)})^2\mathbf{C}^{(g+1)} - (\sigma^{(g)})^2\mathbf{C}^{(g)}$. Then, $\Delta\mathbf{m}^{(g+1)}$ and $\Delta\mathbf{\Sigma}^{(g+1)}$ are transformed using their respective Fisher information matrices to achieve invariance w.r.t. the given normal search space distribution. The detailed derivation is presented in the supplementary material A-C. Using a CSA-ES, i.e., $\mathbf{C} = \mathbf{I}$ (identity matrix), the update equations simplify significantly. The corresponding cumulation paths from (A.11) and (A.13) yield

$$\mathbf{p}_m^{(g+1)} = (1-\beta)\mathbf{p}_m^{(g)} + \sqrt{\beta(2-\beta)\mu/N}\langle\mathbf{z}\rangle^{(g+1)}. \tag{30}$$

$$\mathbf{p}_c^{(g+1)} = (1-\beta)\mathbf{p}_c^{(g)} + \sqrt{\frac{\beta(2-\beta)\mu}{(2N)}}\left[\frac{(\sigma^{(g+1)})^2}{(\sigma^{(g)})^2} - 1\right]\mathbf{1}, \tag{31}$$

with $\mathbf{1} = [1, \ldots, 1]$ of length N. The PSA measures the squared norm of (30) and (31). Aggregating both vectors into a single update vector $\mathbf{p}_\theta^{(g+1)} = (\mathbf{p}_m^{(g+1)}, \mathbf{p}_c^{(g+1)})$, one evaluates

$$\|\mathbf{p}_\theta^{(g+1)}\|^2 = \|\mathbf{p}_m^{(g+1)}\|^2 + \|\mathbf{p}_c^{(g+1)}\|^2 \tag{32}$$

The results of (30) and (31) are notable. While (30) mirrors the cumulation of CSA in (1) with different cumulation constant β and normalization, the path (31) aggregates relative σ-changes. The PSA operates by measuring $\|\mathbf{p}_\theta\|^2 = \|\mathbf{p}_m\|^2 + \|\mathbf{p}_c\|^2$ and comparing the result to the given threshold $\mathcal{T}_\theta = 1.4$. Under random selection, one observes $\|\mathbf{p}_m\|^2 \approx 1$ and $\|\mathbf{p}_c\|^2 \gtrapprox 0$.

[2]Note that [7] and [18] elaborate the similarities between their update equations and the natural gradient interpretation of the CMA-ES. This is omitted at this point since we will investigate a simplification of the PSA-CMA-ES by only including the CSA without covariance matrix adaptation.

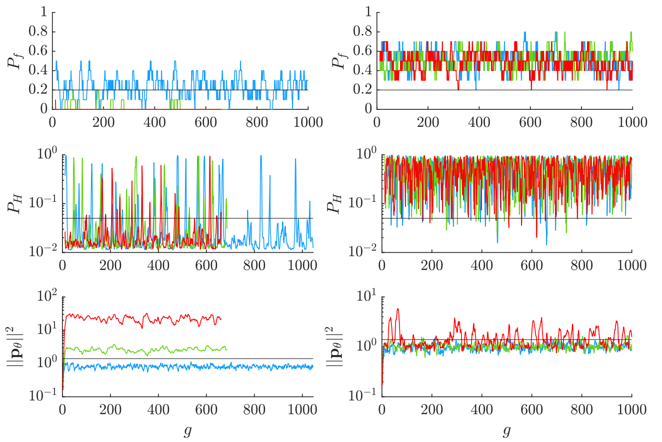


Fig. 5: Sphere (left column) and random function (right column) with CSA (4a) for $N = 100$ using APOP (top, $L = 10$), pcCSA (center, $L = 10$), and PSA (bottom $\beta = 1/10$). The performance is measured with deactivated population control at $\mu = 10, 100, 1000$ (blue, green, and red signals, respectively). The threshold $\mathcal{T}$ is shown in solid black.

PSA associates $\|\mathbf{p}_\theta\|^2 < \mathcal{T}_\theta$ with insufficient performance and increases μ accordingly. Similarly, μ is decreased if $\|\mathbf{p}_\theta\|^2 > \mathcal{T}_\theta$ is measured. The steady-state expected value under random selection yields $\mathrm{E}[\|\mathbf{p}_m^{(g+1)}\|^2] = \mathrm{E}[\|\mathbf{p}_m^{(g)}\|^2] = \|\mathbf{p}_m\|^2$, $\mathrm{E}[\mathbf{p}_m^{(g)}\langle\mathbf{z}\rangle^{(g+1)}] = 0$ (no preferred search direction), and $\mathrm{E}[\|\langle\mathbf{z}\rangle\|^2] = N/\mu$ from [12, (5.2)], such that one gets $\|\mathbf{p}_m\|^2 = 1$. The analogous calculation holds for the steady-state $\mathrm{E}[\|\mathbf{s}\|^2] = N$ of the CSA under random selection. From the updates (2) and (3) one infers that σ does not change in expectation. Hence, the contribution of (31) vanishes, giving $\|\mathbf{p}_c\|^2 = 0$. If the selection is not random, both $\|\mathbf{p}_m\|^2$ and $\|\mathbf{p}_c\|^2$ yield relevant contributions, controlling μ to maintain $\|\mathbf{p}_\theta\|^2 \approx \mathcal{T}_\theta$. The sphere steady-state $\|\mathbf{p}_m\|^2$ and $\|\mathbf{p}_c\|^2$ of the PSA are derived in the supplementary material A-D. In (A.21) it is shown that $\|\mathbf{p}_m\|^2 < 1$ for $\gamma \to 1$, while (A.30) predicts linear growth $\|\mathbf{p}_c\|^2 \propto \mu$ with μ. The results demonstrate that PSA will increase μ on the sphere to reach $\|\mathbf{p}_\theta\|^2 \approx \mathcal{T}_\theta = 1.4$, which is not necessarily desired. While this behavior is explained for CSA, it also appears when considering CMA. This indicates that, at least partially, the results and explanations might also hold for CMA.

Before continuing the analysis, the performance measures of the PCS are evaluated with *deactivated* population control. In Fig. 5, the sphere (33a) and random function (33b) are evaluated at $N = 100$ for three constant μ-values. The respective quantities P_f, P_H, and $\|\mathbf{p}_\theta\|^2$ are measured. We are interested in the signal levels in relation to their thresholds. P_f from APOP decreases for increasing μ on f_{sph}. Larger populations yield less deterioration of the median fitness due to a more robust search. Note that at $\mu = 10$, $P_f > \mathcal{T}_f = 0.2$ most of the time. With active population control, this would trigger an undesired μ-increase on the sphere, which decreases the progress efficiency in terms of function evaluations [14, Sec. 6.1.3.4]. On f_{ran}, the signals fluctuate around $P_f \approx 0.5$ (above $\mathcal{T}_f$), successfully indicating bad performance due to random selection. For the pcCSA, P_H mostly stays below the significance level $\mathcal{T}_H = 0.05$ on the sphere, which indicates good performance. On f_{ran}, P_H lies mostly above the threshold, correctly indicating bad performance. Sporadic f-fluctuations may (falsely) indicate good performance. For the PSA, $\|\mathbf{p}_\theta\|^2$ shows different levels depending on μ. At $\mu = 10$, $\|\mathbf{p}_\theta\|^2$ is smaller than the threshold $\mathcal{T}_\theta = 1.4$. At $\mu = 100, 1000$, it lies above it. Similar to the APOP, the PSA will increase μ on the sphere function to reach $\|\mathbf{p}_\theta\|^2 \approx 1.4$. This effect is due to the aggregation of relative σ-changes via (31) which increase for larger μ due to faster convergence (large $\|\mathbf{p}_c\|^2$-contributions compared to $\|\mathbf{p}_m\|^2$, see supplementary material A-D). On the random function, the three signals lie in the vicinity of $\|\mathbf{p}_\theta\|^2 \approx 1$ mostly due to $\|\mathbf{p}_m\|^2 \approx 1$ and vanishing $\|\mathbf{p}_c\|^2 \gtrapprox 0$ (fluctuations can be observed). In general, the PSA detects random selection well.

	L	β	α_μ	Δ_g	r_σ
P1	$\lceil N^{1/2} \rceil$	$1/N^{1/2}$	1.05	0	$(\mu^{(g+1)}/\mu^{(g)})^{1/2}$
P2	10	$1/10$	2	10	$(\mu^{(g+1)}/\mu^{(g)})^{1/2}$

TABLE I: Two parameter sets P1 and P2 for comparing the adaptive PCS.

A. Method Comparison

For the comparison of the introduced PCS, two sets of parameters will be chosen according to Tab. I by the following argumentation. The goal is to align the time scales of the PCS by which the performance of the ES is evaluated. Both pcCSA and APOP aggregate f-values over a length of L generations. Using the derived time scale (17), we set $L = \lceil N^{1/2} \rceil$ for P1. Similarly, the backward time horizon of the cumulation path scales as $1/\beta$, see [10]. To align the PSA with the array length L, we choose $\beta = 1/N^{1/2}$. Note that the CSA (4a) also operates at $c_\sigma = 1/N^{1/2}$. For P1 one chooses a small $\alpha_\mu = 1.05$, no waiting Δ_g, and r_σ (25b) according to Sec. IV to reduce the stress on the σ-adaptation ($\alpha_\mu = 1.05 \ll 2$). An alternative to P1 is the parameter set P2, including waiting times and performing larger μ-changes. This should stabilize the CSA additionally, see Fig. 4b. Since $G \propto \sqrt{N}$ scales weakly with N, one may choose a constant value $L = 10$ (and $\beta = 1/10$). As an example, $N = 10$ would yield $L = \lceil \sqrt{N} \rceil = 4$ and for $N = 1000$ one has $L = 32$. Both results are not too far away from $L = 10$. A constant L and β schedule was also applied in [8] and [7], however, large N were not investigated. Both sets require a comparable number of generations to reach $\mu_{\max} = 1024$ starting from $\mu_{\min} = 4$, with P1 yielding $g = 89$ and P2 $g = 78$ (if persistent μ-increase is triggered).

The test functions to be investigated are chosen as follows. The goal is to have a simple benchmark set to test basic, but essential properties of the PCS together with the underlying CSA. The properties are performance on a unimodal function, behavior under high noise, and performance on a highly multimodal function, all at varying dimensionality. To this end,

	P1	μ_{25}	μ_{med}	μ_{75}	P2	μ_{25}	μ_{med}	μ_{75}
APOP	S10	14	20	37	S10	8	16	32
	S100	23	95	318	S100	16	16	32
	S1000	29	130	523	S1000	16	16	32
	N10	1024	1024	1024	N10	1024	1024	1024
	N100	1024	1024	1024	N100	1024	1024	1024
	N1000	1024	1024	1024	N1000	1024	1024	1024
pcCSA	S10	129.5	1024	1024	S10	4	4	8
	S100	4	4	7	S100	4	4	8
	S1000	4	4	4	S1000	4	4	8
	N10	1024	1024	1024	N10	1024	1024	1024
	N100	1024	1024	1024	N100	1024	1024	1024
	N1000	1024	1024	1024	N1000	1024	1024	1024
PSA	S10	13	17	20	S10	8	8	16
	S100	33	41	56	S100	32	32	64
	S1000	97	126	163	S1000	256	256	512
	N10	29	37	56	N10	16	32	64
	N100	343	594	927	N100	256	512	1024
	N1000	1024	1024	1024	N1000	1024	1024	1024

TABLE II: CSA (4a) with APOP, pcCSA, PSA (top to bottom, respectively) with P1 and P2 (left and right column).

	P1	μ_{25}	μ_{med}	μ_{75}	P2	μ_{25}	μ_{med}	μ_{75}
APOP	S10	29	40	53	S10	32	32	64
	S100	38	137	429	S100	128	128	256
	S1000	37	177	761	S1000	512	1024	1024
	N10	1024	1024	1024	N10	1024	1024	1024
	N100	1024	1024	1024	N100	1024	1024	1024
	N1000	1024	1024	1024	N1000	1024	1024	1024
pcCSA	S10	383	1024	1024	S10	4	8	16
	S100	593	800	975	S100	64	256	512
	S1000	11	18	31	S1000	256	512	1024
	N10	1024	1024	1024	N10	1024	1024	1024
	N100	1024	1024	1024	N100	1024	1024	1024
	N1000	1024	1024	1024	N1000	1024	1024	1024
PSA	S10	29	33	37	S10	8	16	16
	S100	429	474	550	S100	256	512	512
	S1000	1024	1024	1024	S1000	1024	1024	1024
	N10	133	201	287	N10	64	128	256
	N100	1024	1024	1024	N100	1024	1024	1024
	N1000	1024	1024	1024	N1000	1024	1024	1024

TABLE III: CSA (4b) with APOP, pcCSA, PSA (top to bottom, respectively) with P1 and P2 (left and right column).

we select

$$f_{\text{sph}}(\mathbf{y}) := \textstyle\sum_{i=1}^{N} y_i^2 \tag{33a}$$

$$f_{\text{ran}}(\mathbf{y}) := \mathcal{N}(0,1) \tag{33b}$$

$$f_{\text{ras}}(\mathbf{y}) := \textstyle\sum_{i=1}^{N} [y_i^2 + A(1-\cos(\alpha y_i))]. \tag{33c}$$

On unimodal functions, the population size should be kept small to reduce the number of function evaluations. On the random function, it should reach its maximum value and be kept large. The reason is that large populations reduce the expected residual distance for a sphere under high noise [1]. On the Rastrigin function, it should be dynamically changed since local attraction is only relevant within a certain range [2]. For both large and small $R^2 = \sum_i y_i^2$, f_{ras} behaves like a quadratic function.

A series of experiments on f_{sph} (denoted by "S") and f_{ran} ("N" for noise) is shown in Tab. II, Tab. III, and Tab. IV. The number denotes the dimensionality (e.g. S10 is f_{sph} at $N = 10$). μ_{25}, μ_{med}, and μ_{75} measure the 25-th, 50-th, and 75-th percentile of the occurred $\mu^{(g)}$. They are an indicator of the overall μ-level and distribution width. Exemplary dynamics from the tables are shown in Figs. 6, 7, and 8. In general, one observes high variation of the performance on f_{sph} and f_{ran}, depending on the CSA. The best configuration for APOP and pcCSA is shown in Tab. II with CSA (4a) for P2 (see top right and center right). They maintain low μ-levels on the f_{sph} and high, stable μ-levels on f_{ran} (examples in Figs. 6 and 7). The PSA shows mixed results (bottom right). The main issue is insufficient performance (low μ) for N10. This is critical, e.g., when optimizing multimodal functions at small N where high

	P1	μ_{25}	μ_{med}	μ_{75}	P2	μ_{25}	μ_{med}	μ_{75}
APOP	S10	438	689	975	S10	1024	1024	1024
	S100	97	399	1024	S100	128	512	1024
	S1000	32	147	594	S1000	16	32	32
	N10	1024	1024	1024	N10	1024	1024	1024
	N100	1024	1024	1024	N100	1024	1024	1024
	N1000	1024	1024	1024	N1000	1024	1024	1024
pcCSA	S10	1024	1024	1024	S10	512	1024	1024
	S100	4	6	9	S100	4	4	8
	S1000	4	4	4	S1000	4	4	8
	N10	1024	1024	1024	N10	1024	1024	1024
	N100	1024	1024	1024	N100	1024	1024	1024
	N1000	1024	1024	1024	N1000	1024	1024	1024
PSA	S10	928	1024	1024	S10	16	16	32
	S100	40.5	59	103	S100	32	64	512
	S1000	88	109	135	S1000	128	256	256
	N10	740	974	1024	N10	128	512	1024
	N100	761	1024	1024	N100	512	1024	1024
	N1000	1023	1024	1024	N1000	512	1024	1024

TABLE IV: CSA (4c) with APOP, pcCSA, PSA (top to bottom, respectively) with P1 and P2 (left and right column).

μ-levels are needed for global convergence. The overall best configuration for the PSA uses CSA (4c) for which it was designed in [7], see Tab. IV (bottom right). However, it shows somewhat elevated μ-levels on f_{sph} (also reported by [7]).

Now, the examples of Figs. 6, 7, and 8 are discussed in more detail. On the left side, the dynamics are displayed (see legend). On the right, the corresponding performance measures are shown in black with P_f (APOP), P_H (pcCSA), and $\|\mathbf{p}_\theta\|^2$ (PSA), with the threshold shown in dashed magenta. The respective top and center plots show examples where the PCS works relatively well. The bottom plot shows an example where undesired behavior was observed. Note that no f-values are shown for f_{ran} (random noise).

The APOP in Fig. 6 on f_{sph} controls μ, such that P_f is close to the threshold 0.2. On f_{ran}, $P_f \approx 0.5$ lies above the threshold due to selection on a random function. Hence, μ is kept at the maximum value. In the bottom plot, zero waiting $\Delta_g = 0$ (via P1) introduces unnecessary oscillations of μ. They are related to constant changes in the f-distribution, such that (26) falsely detects fitness deterioration. Waiting $\Delta_g > 0$ is necessary to remove these issues (used in P2). An example of the changing f-distribution is shown in the supplementary material A-E.

The pcCSA in Fig. 7 shows very low μ-levels on f_{sph}, which is desired. On f_{ran}, the hypothesis test detects stagnation reliably and maintains a large μ. In the bottom plot, the small data set $L = \lceil N \rceil = 4$ yields large standard errors in (29), falsely indicating bad performance and increasing μ (example shown in the supplementary material A-F). The pcCSA requires a larger sample size L to overcome this issue (see P2 with $L = 10$).

The PSA in Fig. 8 measures $\|\mathbf{p}_\theta\|^2$ (black) via $\|\mathbf{p}_m\|^2$ (blue) and $\|\mathbf{p}_c\|^2$ (red). On f_{sph}, $\|\mathbf{p}_m\|^2$ and $\|\mathbf{p}_c\|^2$ show similar contributions. On f_{ran}, $\|\mathbf{p}_m\|^2 \approx 1$ and $\|\mathbf{p}_c\|^2 \gtrapprox 0$, indicating random selection (see discussion below (31)). However, spurious fluctuations of σ may generate notable $\|\mathbf{p}_c\|^2$-contributions, yielding $\|\mathbf{p}_\theta\|^2$ above the threshold and reducing μ. The bottom example shows issues of the PSA to increase μ on f_{ran} due to the mentioned fluctuations. This occurs for CSA (4a), but is less pronounced for CSA (4c) since the latter adapts σ significantly slower due to $D \propto \sqrt{\mu/N}$ from (22). The PSA requires slow adaptation for small N to correctly detect random selection. Note that this effect is not related to the applied $r_\sigma = \sqrt{\mu^{(g+1)}/\mu^{(g)}}$, but to the underlying CSA

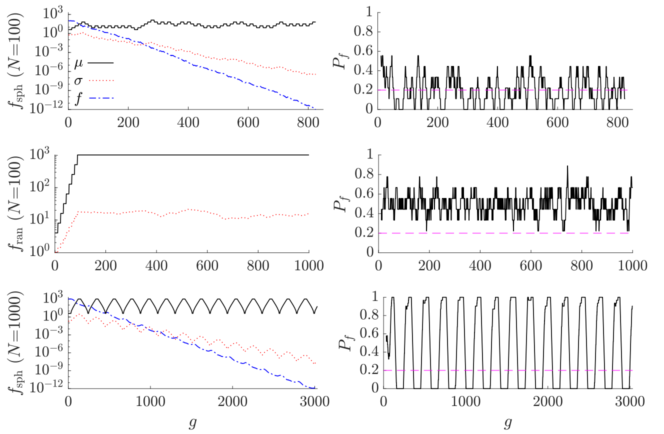


Fig. 6: APOP-dynamics with CSA (4a) for S100 (P2) at the top, N100 (P2) in the center, and undesired μ-oscillations for S1000 (P1) at the bottom.

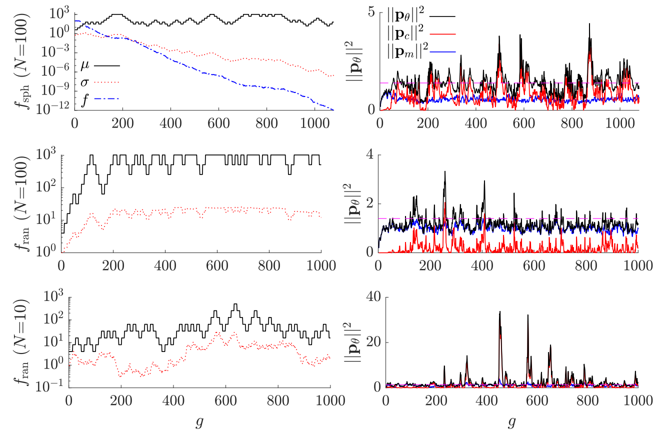


Fig. 8: PSA-dynamics with CSA (4c) for S100 (P2) at the top, N100 (P2) in the center, and insufficient μ-level for N10 with CSA (4a) at the bottom.

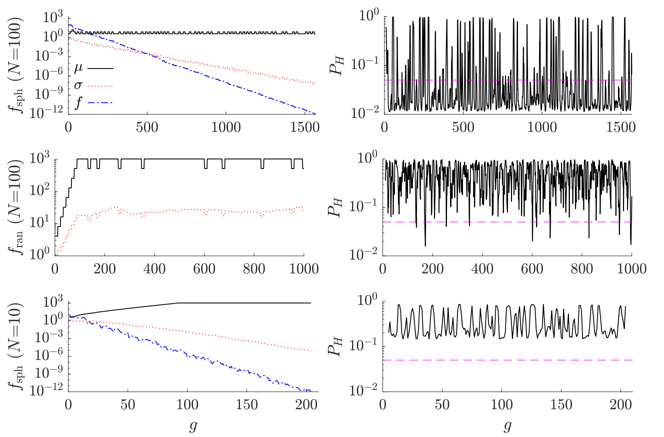


Fig. 7: pcCSA-dynamics with CSA (4a) for S100 (P2) at the top, N100 (P2) in the center, and undesired μ-increase for S10 (P1) at the bottom.

(see supplementary material A-G).

One major observation can be made from Tab. III and Tab. IV. Recall that both CSA (4b) and (4c) show increasingly slow adaptation within respective limits, see $\gamma \to 1$ in Fig. 2c. Very slow adaptation can significantly deteriorate the PCS performance on $f_{\rm sph}$. In Tab. III, larger N show large $\mu_{\rm med}$-levels and in Tab. IV, large $\mu_{\rm med}$-levels occur for small N. In general, suboptimal performance with slow adaptation on $f_{\rm sph}$ can be explained in terms of measured f-values using Fig. 1a. Since the adaptation is very slow, the ES operates at very large σ^* close to $\sigma^*_{\varphi_0}$. It progresses very slowly and selects worse fitness values more often. The deterioration is reflected within the pcCSA by the hypothesis test mostly indicating stagnation. Similarly, the APOP counts significantly more worse fitness changes via (27), leading to unnecessary μ-increase on $f_{\rm sph}$. The PSA shows mixed results. In Tab. III, it works well for S10 (small $\mu_{\rm med}$), but yields $\mu_{\rm med} = \mu_{\rm max}$ for S1000. It requires a different CSA, i.e., adaptation speed as a function of μ and N, to achieve better performance (see Tab. IV).

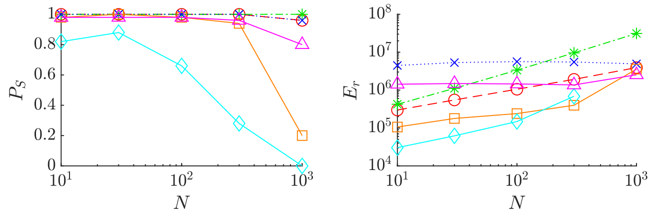
Comparing the experiments of parameter sets P1 and P2 in Tab. II, Tab. III, and Tab. IV shows mixed results. There are cases in which one option is either better or worse than the other and the results depend on both CSA and PCS. However, P2 with $\Delta_g > 0$ is preferred for APOP due to stability (see discussion of Fig. 6). Furthermore, P2 yields improved results for pcCSA (hypothesis test for small N) and in general good results for PSA. Hence, the parameter set P2 will be used as the default in Sec. V-B for further investigations.

B. Comparison on the Rastrigin Function

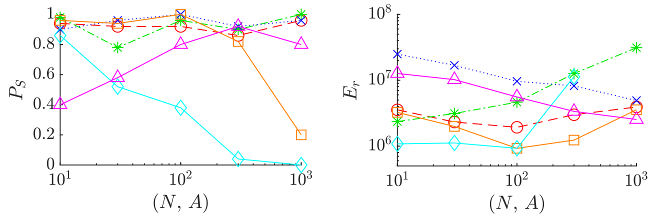
Now that basic properties of the PCS and CSA variants have been studied on $f_{\rm sph}$ and $f_{\rm ran}$, a simplified benchmark is performed on the Rastrigin function $f_{\rm ras}$. The reason for choosing Rastrigin is that sufficiently large populations are needed to find the global attractor [2], [4]. However, local attraction has only a limited range, such that adaptive PCS should keep the population size small far away from the global attractor (global quadratic structure) and within a local (or global) attractor. The goal is to investigate the performance of the best configurations from Sec. V-A. Furthermore, the PCS will be compared against the CSA variants with constant $\mu = \mu_{\rm max}$.

The tested parameter sets of Rastrigin are chosen as follows. The first experiment is performed for varying N with constant multimodality parameters $A = 3$ and $\alpha = 2\pi$. Hence, $f_{\rm ras}$ becomes more difficult to optimize with increasing N. For the second experiment, one varies N together with A. Large A are chosen for small N and vice-versa. The values are chosen such that CSA (4a) yields roughly a constant success rate $P_S \approx 0.9$ at $\mu = \mu_{\rm max}$. The test provides "similar difficulty" for an ES operating at constant population size and will serve as a reference.

For the initialization, one chooses $\mathbf{y}^{(0)} = 2\lceil \alpha A/2 \rceil \mathbf{1}$ (outside the local attraction region), $\sigma^{(0)} = \sigma^*_{\varphi_0} \| \mathbf{y}^{(0)} \| / N$ via (6) ensuring a large initial step-size, and $\mathbf{s}^{(0)} = \mathbf{1}$ for the CSA. The runs terminate for $f < f_{\rm stop} = 10^{-3}$ (global convergence) or $\sigma < 10^{-3}$ with $f \geq f_{\rm stop}$ (local convergence). For each parameter set, 50 trials are evaluated and the success

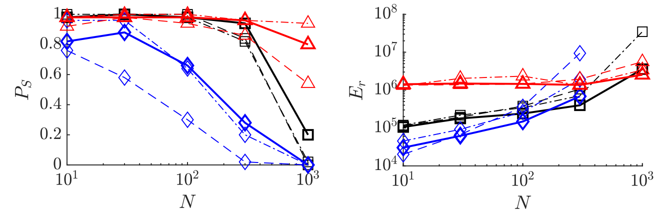


(a) Variation of N for $A = 3$ and $\alpha = 2\pi$.

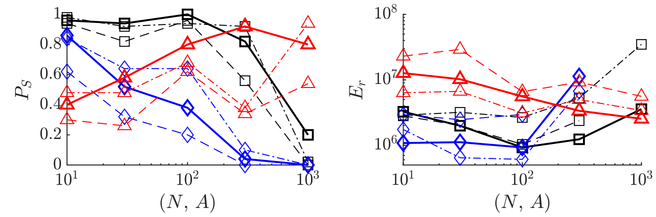


(b) $(N, A) = (10, 65), (30, 33), (100, 12), (300, 7), (1000, 3)$.

Fig. 9: Success rate P_S and expected runtime E_r on f_{ras} ($\alpha = 2\pi$). The CSA with $\mu = \mu_{\max}$ are shown as $\circ$ (4a), $*$ (4b), and $\times$ (4c). The PCS are shown as $\square$ (35a), $\diamond$ (35b), and $\triangle$ (35c).



(a) Variation of N for $A = 3$ and $\alpha = 2\pi$.



(b) $(N, A) = (10, 65), (30, 33), (100, 12), (300, 7), (1000, 3)$.

Fig. 10: Setup as shown in Fig. 9. The PCS are shown as $\square$ (35a), $\diamond$ (35b), and $\triangle$ (35c). Given P2 in Tab. I, rescaling r_σ is varied using (25b) (solid, bold, same data as in Fig. 9), (25c) (dash-dotted), and (25a) (dashed).

rate P_S is measured. We measure the expected runtime E_r (in function evaluations) to reach global convergence by including the number of function evaluations of successful (F_s) and unsuccessful (F_u) runs. For $P_S > 0$, it is estimated as [19]

$$E_r = (F_s + F_u)/P_S. \tag{34}$$

For the PCS, we choose the following parameters (see Sec. V-A)

$$\begin{aligned} &\text{APOP:} && \text{CSA (4a) and P2 (Tab. I)} && \text{(35a)}\\ &\text{pcCSA:} && \text{CSA (4a) and P2 (Tab. I)} && \text{(35b)}\\ &\text{PSA:} && \text{CSA (4c) and P2 (Tab. I).} && \text{(35c)} \end{aligned}$$

The population size limits are $\mu^{(0)} = \mu_{\min} = 4$ and $\mu_{\max} = 1024$. There are two reasons for limiting μ. One reason are limited computational resources, especially when optimizing Rastrigin at large $N = 1000$. The second reason is that the PCS are compared to runs with constant $\mu = \mu_{\max}$. Certain examples will show that $\mu_{\max}$ is not reached by the PCS, even though it should be attained to achieve better performance.

The results are shown in Fig. 9 in terms of P_S and E_r. The following observations can be made. At constant $\mu = \mu_{\max}$, CSA (4a) yields very good overall performance. While P_S is comparable to CSA (4b) and (4c), its E_r levels are the lowest (the only exception is $N = 10$ in Fig. 9b). Furthermore, E_r remains approximately constant in Fig. 9b, while CSA (4b) and (4c) show notable upward and downward trends for E_r, respectively, due to their different adaptation characteristics. With active μ-control, APOP and pcCSA stay mostly below $\mu = \mu_{\max}$ of CSA (4a), which is the CSA they utilize. This is important as it illustrates a more effective search with active μ-control. The highest efficiency (lowest E_r) for small N is achieved by pcCSA, which is related to very low μ-levels in the sphere limits (see Tab. II). However, the pcCSA shows a significant drop in P_S (and increase of E_r) for large N. The APOP shows very good overall results, showing relatively high P_S-levels and satisfactory E_r for the tested configurations. However, its performance also decreases to some extent for large N. The PSA shows lower E_r-values than its CSA (4c) at $\mu = \mu_{\max}$, which is good. In Fig. 9b, its E_r-slope closely follows CSA (4c), which illustrates that the CSA properties are recovered with active population control. In contrast to APOP and pcCSA, the PSA performs best at large $N = 1000$. At small $N = 10$, it tends to pick up more fluctuations of σ that falsely indicate good performance when measuring $\|\mathbf{p}_\theta\|^2$ (similar to the bottom plot in Fig. 8). At large N, the fluctuations of the contributing terms $\|\mathbf{p}_m\|^2$ and $\|\mathbf{p}_c\|^2$ are reduced and the detection of insufficient performance becomes more robust. Exemplary dynamics of the PCS on f_{ras} are given in the supplementary material A-H.

Figure 10 shows the same experiment as in Fig. 9, now with σ-rescaling r_σ being varied for each of the three PCS. For APOP, one observes improved P_S (left) and E_r (right) for (25b) most of the time. For pcCSA, mixed results are obtained. Either (25b) or (25c) yields higher P_S and lower E_r, respectively. Similar (mixed) results are observed for PSA. Linear scaling improves E_r for small N in Fig. 10b, but not at large N. Recall that a linear rescaling was originally used in [7], see also discussion of (20). In nearly all cases, no rescaling (25a) yields worse results for all PCS. The stability tests with predefined μ-schedule in Sec. IV have shown clear advantages of (25b), which is not observed in Fig. 10. Using active PCS introduces feedback that controls μ based on the measured $\mathcal{P}$. The realized changes in μ (over multiple generations) are usually moderate compared to a fixed schedule. This leads to smaller σ-changes due to rescaling. Furthermore, the waiting time $\Delta_g > 0$ helps to stabilize the σ-dynamics for any r_σ, see also Fig. 4b. Experiments on f_{sph} and f_{ran} with varying r_σ (not shown) yield similar inconclusive results. The only exception is PSA which on f_{sph} appears more susceptible to larger σ-changes introduced by (25c).

VI. Conclusions and Outlook

In this paper, adaptive (online) population control strategies (PCS) were investigated for a multi-recombinative $(\mu/\mu_I, \lambda)$-ES with cumulative step-size adaptation (CSA) and isotropic mutations. To this end, the adaptation properties of standard CSA variants were discussed on the sphere as a function of population size μ and dimensionality N. Furthermore, scaling laws for the generation number ($\sqrt{N}$-law) and rescaling of the mutation strength ($\sqrt{\mu}$-law) were derived on the sphere for large population sizes with $\mu \gg \sqrt{N}$.

The obtained scaling laws were implemented into three state-of-the-art PCS, namely APOP [8], pcCSA [6], and PSA [7]. The experimental (and theoretical) analysis of the PCS has shown significant dependence of their respective performance measures on the CSA variants. On the selected test bed (sphere, random, and Rastrigin functions), APOP and pcCSA using CSA (4a) (cumulation constant $= 1/\sqrt{N}$, damping $= \sqrt{N}$) perform well due to comparably fast σ-adaptation. This improves overall progress and also helps the PCS to discern good from insufficient ES-performance. PSA works more efficiently using CSA (4c) due to its underlying cumulation paths. Especially on the Rastrigin function, the PCS benefit from short data aggregation periods ($\sqrt{N}$-law) together with equally short waiting times for the decorrelation of μ-dependent data. For mutation strength rescaling r_σ, stability tests on the sphere have shown clear advantages of a $\sqrt{\mu}$-law. However, with active PCS the effects of r_σ are less pronounced. Strengths and weaknesses of the individual PCS have been discussed. In conclusion, there is no clear winner among the PCS. Best overall performance was achieved by APOP, providing good results across all parameter variations. On the other hand, pcCSA was the most efficient PCS at low N, while PSA achieved the best results at large N.

Of course, the performance of individual PCS can be improved by selective parameter tuning. However, as the aim of the present work was to investigate basic properties of CSA and PCS, this is left for future research. Further research aims at testing favorable PCS configurations on a larger benchmark set of noisy and multimodal functions. The goal is to have good performance over a large dimensionality range, while maintaining efficiency on unimodal functions. Another potential line of research is the improvement (or hybridization) of different population control methods, which may evaluate the ES performance in both search and fitness space. Future research should also investigate the interaction of PCS with covariance matrix adaptation (CMA). These investigations should, hopefully, yield improved algorithms that perform well on a variety of test functions and also scale well with the dimensionality.

References

[1] D. Arnold and H.-G. Beyer, "Performance Analysis of Evolution Strategies with Multi-Recombination in High-Dimensional $\mathbb{R}^N$-Search Spaces Disturbed by Noise," *Theoretical Computer Science*, vol. 289, pp. 629–647, 2002.

[2] L. Schönenberger and H.-G. Beyer, "On a Population Sizing Model for Evolution Strategies Optimizing the Highly Multimodal Rastrigin Function," in *Proceedings of the Genetic and Evolutionary Computation Conference*, ser. GECCO '23, New York, NY, USA, 2023, p. 848–855. [Online]. Available: https://doi.org/10.1145/3583131.3590451

[3] A. Omeradzic and H.-G. Beyer, "Convergence Properties of the $(\mu/\mu_I, \lambda)$-ES on the Rastrigin Function," in *Proceedings of the 17th ACM/SIGEVO Conference on Foundations of Genetic Algorithms*, ser. FOGA '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 117–128. [Online]. Available: https://doi.org/10.1145/3594805.3607126

[4] ——, "Self-Adaptation of Multi-Recombinant Evolution Strategies on the Highly Multimodal Rastrigin Function," *IEEE Transactions on Evolutionary Computation*, 2024. [Online]. Available: https://doi.org/10.1109/TEVC.2024.3400857

[5] H.-G. Beyer and B. Sendhoff, "Evolution Strategies for Robust Optimization," in *Proceedings of the WCCI'06 Conference*. Piscataway, NJ: IEEE Press, 2006, pp. 4489–4496.

[6] M. Hellwig and H.-G. Beyer, "Evolution under Strong Noise: A Self-Adaptive Evolution Strategy Can Reach the Lower Performance Bound - the pcCMSA-ES," in *Parallel Problem Solving from Nature XXIV*, E. Hart et al., Ed. Berlin: Springer, 2016, pp. 26–36, https://doi.org/10.1007/978-3-319-45823-6_3.

[7] K. Nishida and Y. Akimoto, "PSA-CMA-ES: CMA-ES with Population Size Adaptation," in *Proceedings of the Genetic and Evolutionary Computation Conference*, ser. GECCO '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 865–872. [Online]. Available: https://doi.org/10.1145/3205455.3205467

[8] D. M. Nguyen and N. Hansen, "Benchmarking CMAES-APOP on the BBOB Noiseless Testbed," in *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, ser. GECCO '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 1756–1763. [Online]. Available: https://doi.org/10.1145/3067695.3084207

[9] N. Hansen, A. Auger, R. Ros, O. Mersmann, T. Tušar, and D. Brockhoff, "COCO: A platform for comparing continuous optimizers in a black-box setting," *Optimization Methods and Software*, vol. 36, pp. 114–144, 2021.

[10] N. Hansen, "The CMA Evolution Strategy: A Tutorial," 2023. [Online]. Available: https://www.arxiv.org/abs/1604.00772

[11] N. Hansen and A. Ostermeier, "Completely Derandomized Self-Adaptation in Evolution Strategies," *Evolutionary Computation*, vol. 9, no. 2, pp. 159–195, 2001, https://doi.org/10.1162/106365601750190398.

[12] D. Arnold, *Noisy Optimization with Evolution Strategies*. Dordrecht: Kluwer Academic Publishers, 2002.

[13] N. Hansen, "Verallgemeinerte individuelle Schrittweitenregelung in der Evolutionsstrategie," Doctoral thesis, Technical University of Berlin, Berlin, 1998.

[14] H.-G. Beyer, *The Theory of Evolution Strategies*, ser. Natural Computing Series. Heidelberg: Springer, 2001, DOI: 10.1007/978-3-662-04378-3.

[15] A. Omeradzic and H.-G. Beyer, "Mutation Strength Adaptation of the $(\mu/\mu_I, \lambda)$-ES for Large Population Sizes on the Sphere Function," 2024. [Online]. Available: https://arxiv.org/abs/2408.09761

[16] Y. Akimoto, J. Sakuma, I. Ono, and S. Kobayashi, "Functionally specialized CMA-ES: a modification of CMA-ES based on the specialization of the functions of covariance matrix adaptation and step size adaptation," in *Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation*, ser. GECCO '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 479–486. [Online]. Available: https://doi.org/10.1145/1389095.1389188

[17] H.-G. Beyer and M. Hellwig, "Analysis of the pcCMSA-ES on the noisy ellipsoid model," in *GECCO'17: Proceedings of the Genetic and Evolutionary Computation Conference*. New York: ACM, 2017, pp. 689–696, DOI: 10.1145/3071178.3079195.

[18] K. Nishida and Y. Akimoto, "Population size adaptation for the CMA-ES based on the estimation accuracy of the natural gradient," in *GECCO'16: Proceedings of the Genetic and Evolutionary Computation Conference*. New York: ACM, 2016, pp. 237–244.

[19] A. Auger and N. Hansen, "Performance Evaluation of an Advanced Local Search Evolutionary Algorithm," in *Congress on Evolutionary Computation, CEC'05*, vol. 2. IEEE, 2005, pp. 1777–1784.