PSA-CMA-ES: CMA-ES with population size adaptation

Conference Paper · July 2018

DOI: 10.1145/3205455.3205467

2 authors:

Kouhei Nishida
Shinshu University
7 PUBLICATIONS **49** CITATIONS
SEE PROFILE

Youhei Akimoto
University of Tsukuba
139 PUBLICATIONS **1,504** CITATIONS
SEE PROFILE

PSA-CMA-ES: CMA-ES with Population Size Adaptation

Kouhei Nishida
Interdisciplinary Graduate School of Science and Technology, Shinshu Univercity
17st208e@shinshu-u.ac.jp

Youhei Akimoto
Faculty of Engineering, Information and Systems
University of Tsukuba
akimoto@cs.tsukuba.ac.jp

ABSTRACT

The population size, i.e., the number of candidate solutions generated at each iteration, is the most critical strategy parameter in the covariance matrix adaptation evolution strategy, CMA-ES, which is one of the state-of-the-art search algorithms for black-box continuous optimization. The population size is required to be larger than its default value when the objective function is well-structured multimodal and/or noisy, while we want to keep it as small as possible for optimization speed. However, the strategy parameter tuning based on trial and error is, in general, prohibitively expensive in black-box optimization scenario. This paper proposes a novel strategy to adapt the population size for CMA-ES. The population size is adapted based on the estimated accuracy of the update of the normal distribution parameters. The CMA-ES with the proposed population size adaptation mechanism, PSA-CMA-ES, is tested both on noiseless and noisy benchmark functions, and compared with existing strategies. The results revealed that the PSA-CMA-ES works well on well-structured multimodal and/or noisy functions, but causes inefficient increase of the population size on unimodal functions. Furthermore, it is shown that the PSA-CMA-ES can tackle noise and multimodality at the same time.

CCS CONCEPTS

• **Mathematics of computing → Continuous optimization**; **Bio-inspired optimization**;

KEYWORDS

CMA-ES, population size adaptation, multimodal functions, noisy functions

ACM Reference Format:
Kouhei Nishida and Youhei Akimoto. 2018. PSA-CMA-ES: CMA-ES with Population Size Adaptation. In *GECCO '18: Genetic and Evolutionary Computation Conference, July 15–19, 2018, Kyoto, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3205455.3205467

1 INTRODUCTION

The covariance matrix adaptation evolution strategy (CMA-ES) [7, 10, 11] is recognized as a state-of-the-art stochastic algorithm for black-box continuous optimization problems. The CMA-ES samples candidate solutions from a multivariate normal distribution, evaluates them on the objective of the problem, then updates the parameters of the multivariate normal distribution. The CMA-ES exhibits several invariance properties including the invariance to order preserving transformation of the objective function and the invariance to linear transformation of coordinate system of the search space. They are essential for black-box optimization where good problem feature such as well-conditioning and variable independence can not be assumed.

One very important feature of the CMA-ES that attracts practitioners is that it is a quasi-parameter-free algorithm. That is, all the strategy parameters, e.g., the learning rate, have the default values that only depend on the dimension of the search space. In the black-box optimization the parameter tuning is often the bottleneck of the optimization process. Since prior knowledge is limited and the parameter tuning requires extremely expensive trial-and-error.

The only but most important strategy parameter that is sometimes required to tune is the population size, i.e., the number of samples at each iteration. A population size larger than its default value, which is logarithmic in the dimension, improves the quality of the solution obtained by the CMA-ES when we apply it to a noisy function or a multimodal function with globally well structure [5, 6, 9]. Such a function is called a well-structured multimodal function, also known as a big valley function. The other multimodal functions are called weakly-structured multimodal function. In particular, if the objective function is noisy and its noise-to-signal ratio (NSR) increases as it gets closer to the optimum, e.g., the additive steady noise case, a fixed population size will not be sufficient and it needs to increase correspondingly to the NSR. On the other hand, a larger population size generally results in spending more function evaluations until the algorithm converges. Therefore the population size is desired to be as small as possible. However, the tuning of the population size is a difficult task and its reasonable value depends on the stage of the optimization process: a large population size might be necessary at the beginning on a well-structured multimodal function, but a smaller population size is sufficient after the sampling distribution becomes so small that the candidate solutions are all sampled in a area where the objective function can be approximated by a unimodal function. Moreover, if the objective function is a weakly-structured multimodal function, a larger population size may cause that the distribution converges into a dominant but non-optimal point.

A simple, yet promising approach is to restart the algorithm with different configurations [4, 5]. For example, the IPOP restart strategy [4] doubles the population size every restart. The BIPOP restart strategy [5] has two regimes: one for the IPOP regime and the other for the local search regime where a relatively small population size and a relatively small initial distribution variance are employed. The

GECCO '18, July 15–19, 2018, Kyoto, Japan

ACM ISBN 978-1-4503-5618-3/18/07…$15.00
https://doi.org/10.1145/3205455.3205467

IPOP regime is helpful for solving a well-structured multimodal function, while the local search regime is helpful for solving a weakly-structured multimodal function where a large population size is barely useful.

Another approach is to adapt the population size [12, 14, 15]. Online adaptation of the population size has potential advantages over restart strategies: the population size is increased only when it is necessary (typically on noiseless multimodal functions), and it is gradually increasing if necessary (typically on noisy functions). Moreover, the efficiency of restart strategies heavily relies on reasonable stopping mechanisms, which is sometimes difficult to design, especially on noisy functions. Recently, several adaptation mechanisms are proposed. The pcCMSA-ES [12] is proposed for noisy optimization and it adapts the population size based on the estimated noise strength on the objective. The CMAES-APOP [14] is proposed for noiseless multimodal optimization and is based on the function value decrease over iterations. These mechanisms focus on the function values to adapt the population size, where as the population size adaptation for the information geometric optimization [15] adapts the population size based on the accuracy of the parameter update and is applied both for multimodal and/or noisy functions. However, the population size adaptation mechanism [15] is only applied with a simplified variant of the CMA-ES, called the pure rank-μ update CMA-ES, since the idea is based on the natural gradient perspective of the CMA-ES.

In this paper, we propose a population size adaptation mechanism for the most commonly used CMA-ES that employs weighted recombination, cumulative step-size adaptation, rank-one and rank-μ covariance matrix update. We extend the idea of the previous study [15] so that we can deal with algorithmic components that are not derived from the natural gradient perspective. The main idea is to regard the distribution parameter update as a stochastic approximation of a deterministic update using finite candidate solutions, and to control the accuracy of the parameter update by adapting the population size. To treat the change of the optimal step-size due to the change of the population size that is derived theoretically based on the quality gain analysis [1], we introduce a step-size correction mechanism, without which the step-size and the population size tend to fluctuate. The resulting algorithm is called the PSA-CMA-ES, the CMA-ES with population size adaptation.

The rest of this paper is organized as follows. Section 2 is devoted to explain the baseline CMA-ES algorithm. In Section 3 we extend the population size adaptation introduced in the previous work [15] so that we can incorporate all the algorithmic components in the standard CMA-ES. Additionally, the step-size correction mechanism is introduced to make the adaptation stable. In Section 4 we investigate the behavior of the PSA-CMA-ES on noiseless and noisy test functions. The performance is also compared with other existing population size adaptation mechanisms. We conclude with summary and direction of future work in Section 5.

2 CMA-ES

The CMA-ES maintains a multivariate normal distribution from which candidate solutions are generated. The CMA-ES parameterizes the multivariate normal distribution $\mathcal{N}(\boldsymbol{m}, \sigma^2 \boldsymbol{C})$ by three components, the mean vector $\boldsymbol{m}$, the step-size σ and the covariance matrix $\boldsymbol{C}$. The mean vector $\boldsymbol{m}$ is the center of the distribution. The step-size σ represents the spread of the distribution and the covariance matrix $\boldsymbol{C}$ mainly determines the shape of the distribution.

At $t+1$-st iteration ($t = 0, 1, \dots$), the candidate solutions $\boldsymbol{x}$ are generated as

$$\boldsymbol{x}_i^{(t+1)} \sim \boldsymbol{m}^{(t)} + \sigma^{(t)} \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}^{(t)}) \quad \text{for } i = 1, \dots, \lambda, \tag{1}$$

where λ is the population size. After the evaluation of these candidate solutions on the objective, they are sorted based on their function values. Let $\boldsymbol{m}_{i:\lambda}$ be the i-th best candidate solution among the λ candidate solutions.

The mean vector $\boldsymbol{m}$ is updated as

$$\boldsymbol{m}^{(t+1)} = \boldsymbol{m}^{(t)} + c_m \sum_{i=1}^{\lambda} w_i (\boldsymbol{x}_{i:\lambda}^{(t+1)} - \boldsymbol{m}^{(t)}), \tag{2}$$

where w_i is the weight corresponding to the i-th best candidate solution $\boldsymbol{x}_{i:\lambda}$ and c_m is the learning rate of the mean vector update.

The CMA-ES employs two evolution paths that accumulate successive steps of the mean vector update for the cumulative step-size adaptation (CSA) and the rank-one update of the covariance matrix.

The evolution path $\boldsymbol{p}_\sigma$ for CSA is updated as

$$\boldsymbol{p}_\sigma^{(t+1)} = (1 - c_\sigma)\boldsymbol{p}_\sigma^{(t)} + \sqrt{c_\sigma(2 - c_\sigma)\mu_{\text{eff}}} \left(\boldsymbol{C}^{(t)}\right)^{-\frac{1}{2}} \frac{\Delta \boldsymbol{m}^{(t)}}{\sigma^{(t)}}, \tag{3}$$

where $\Delta \boldsymbol{m}^{(t)} = \boldsymbol{m}^{(t+1)} - \boldsymbol{m}^{(t)}$ is the step of the mean vector update, $\boldsymbol{C}^{-1/2}$ is the inverse of the matrix square root computed as $\boldsymbol{C}^{-1/2} = \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{B}^T$ with the eigendecomposition $\boldsymbol{C} = \boldsymbol{B}\boldsymbol{D}^2\boldsymbol{B}^T$ of the covariance matrix, μ_{eff} is so called effective variance selection mass defined as $\mu_{\text{eff}} = 1/\sum_{i=1}^{\lambda} w_i^2$ and c_σ is the cumulation factor. The step $\Delta \boldsymbol{m}$ is normalized by the factor $\sigma^{-1}\boldsymbol{C}^{-1/2}$ on the right-hand side (RHS) of (3) so that $\boldsymbol{p}_\sigma^{(t+1)}$ is comparable with $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

The evolution path $\boldsymbol{p}_c$ for the rank-one covariance matrix update is updated as

$$\boldsymbol{p}_c^{(t+1)} = (1 - c_c)\boldsymbol{p}_c^{(t)} + h_\sigma \sqrt{c_c(2 - c_c)\mu_{\text{eff}}} \frac{\Delta \boldsymbol{m}^{(t)}}{\sigma^{(t)}}, \tag{4}$$

where c_c is the cumulation factor and $h_\sigma^{(t+1)}$ is the heaviside function defined, with the dimension n of the objective function, as follows:

$$h_\sigma^{(t+1)} = \begin{cases} 1 & \text{if } \|\boldsymbol{p}_\sigma^{(t+1)}\| < \left(1.4 + \frac{2}{n+1}\right) \chi_n \sqrt{\gamma_\sigma^{(t+1)}} \\ 0 & \text{otherwise} \end{cases}, \tag{5}$$

where χ_n is the expected norm $\mathbb{E}[\|\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})\|]$ of the n-variate standard normal distribution and we use the approximated value $\sqrt{n}(1 - 1/(4n) + 1/(21n^2))$ instead. Usually, the former case is selected. The division by the step-size σ on the RHS of (4) makes $\boldsymbol{p}_c^{(t+1)}$ comparable with $\mathcal{N}(\boldsymbol{0}, \boldsymbol{C}^{(t)})$.

To derive cleanly the population size adaptation mechanism that we present in Sec.3, we introduce the normalization factor for each evolution path. They are updated as follows:

$$\gamma_\sigma^{(t+1)} = (1 - c_\sigma)^2 \gamma_\sigma^{(t)} + c_\sigma(2 - c_\sigma) \tag{6}$$

$$\gamma_c^{(t+1)} = (1 - c_c)^2 \gamma_c^{(t)} + h_\sigma^{(t+1)} c_c(2 - c_c). \tag{7}$$

Note that, they converge to 1 as t increases. The effect of the introduction of these normalization factors is barely recognizable in practice, and they are only for clean derivation.

The step-size σ is updated as

$$\sigma^{(t+1)} = \sigma^{(t)} \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|\boldsymbol{p}_\sigma^{(t+1)}\|}{\chi_n} - \sqrt{\gamma_\sigma^{(t+1)}}\right)\right), \tag{8}$$

where d_σ is the damping factor of the step-size update.

The covariance matrix is updated as

$$\begin{aligned} \boldsymbol{C}^{(t+1)} = \boldsymbol{C}^{(t)} &+ c_1\left(\boldsymbol{p}_c^{(t+1)}\left(\boldsymbol{p}_c^{(t+1)}\right)^T - \gamma_c^{(t+1)}\boldsymbol{C}^{(t)}\right) \\ &+ c_\mu \sum_{i=1}^{\lambda} w_i \left(\left(\boldsymbol{x}_{i:\lambda}^{(t+1)} - \boldsymbol{m}^{(t)}\right)\left(\boldsymbol{x}_{i:\lambda}^{(t+1)} - \boldsymbol{m}^{(t)}\right)^T - \boldsymbol{C}^{(t)}\right). \end{aligned} \tag{9}$$

The second term on the RHS of (9) is the rank-one update and the third term is called the rank-μ update. The learning rate c_1 and c_μ in (9) corresponds to the rank-one and rank-μ update, respectively.

3 POPULATION SIZE ADAPTATION

Our population size adaptation mechanism for the CMA-ES is based on the previous study [15], where the CMA-ES variant that derives from the information geometric optimization (IGO) framework [16] is considered. It is called the pure rank-μ update CMA-ES. In the IGO framework, the parameter update is understood as a stochastic natural gradient ascent, where the population size is translated as the number of Monte-Carlo samples for the gradient estimation. A key observation is that there is less tendency of the estimated natural gradients over iterations on multimodal functions and noisy functions than on noiseless unimodal functions such as the sphere function. Based on this observation, the population size is adapted so that the accuracy, more precisely signal-to-noise level, of the natural gradient estimate is kept at a fixed level. To estimate the estimation accuracy of the natural gradient under the current population size in the current situation, an evolution path in the distribution parameter space is introduced.

We extend the idea of the previous work and apply it to the (standard) CMA-ES. Similarly to the pure rank-μ update CMA-ES, we observe that the parameter update in the CMA-ES becomes more and more deterministic as the population size increases. Therefore, we treat the accuracy of the natural gradient estimate in the previous work as the accuracy of the distribution parameter update that may contain components that are not derived from the natural gradient perspective. The evolution path is introduced with a novel normalization factor. It quantifies the tendency of the parameter updates in the distribution parameter space. The population size is adapted based on the length of the evolution path, where the Fisher metric is taken into account. An additional component is introduced to incorporate the step-size adaptation. That is, when the population size is changed, the step-size is scaled since the optimal step-size depends on the population size. The step-size correction factor is derived from the quality gain analysis [1]. This correction make the algorithm more stable.

3.1 Quantification of Update Accuracy

To quantify accuracy of the parameter update, we introduce the evolution path in the parameter space $\boldsymbol{\Theta}$ of the probability distribution. Let $\boldsymbol{\theta} = (\boldsymbol{m}, \mathrm{vech}(\Sigma))$ represent the parameter vector of the multivariate normal distribution, where $\Sigma = \sigma^2 \boldsymbol{C}$ and $\mathrm{vech}(\boldsymbol{A})$ denotes the vector consisting of the upper triangular elements of the symmetric matrix $\boldsymbol{A}$. We also define the parameter movement vector $\Delta\boldsymbol{\theta}^{(t+1)} = (\Delta\boldsymbol{m}^{(t+1)}, \mathrm{vech}(\Delta\Sigma^{(t+1)}))$ from t-th iteration to $t+1$-st iteration. Its components are computed as

$$\Delta\boldsymbol{m}^{(t+1)} = \boldsymbol{m}^{(t+1)} - \boldsymbol{m}^{(t)}, \tag{10}$$

$$\Delta\Sigma^{(t+1)} = (\sigma^{(t+1)})^2\boldsymbol{C}^{(t+1)} - (\sigma^{(t)})^2\boldsymbol{C}^{(t)}. \tag{11}$$

An evolution path is introduced to accumulate successive parameter movements $\Delta\boldsymbol{\theta}$. A measure of accuracy of the parameter update should not depend on parameterization of the probability distribution, however the parameter movement $\Delta\boldsymbol{\theta}$ does. Moreover, the norm of the parameter movement vector depends on the dimension and the strategy parameters such as the population size and the learning rates. To deal with these issues and realize stable measurement of tendency of the parameter update, we define a novel evolution path as follows:

$$\boldsymbol{p}_\theta^{(t+1)} = (1-\beta)\boldsymbol{p}_\theta^{(t)} + \sqrt{\beta(2-\beta)}\frac{\boldsymbol{\mathcal{I}}_{\theta^{(t)}}^{\frac{1}{2}}\Delta\boldsymbol{\theta}^{(t+1)}}{\mathbb{E}[\|\boldsymbol{\mathcal{I}}_{\theta^{(t)}}^{\frac{1}{2}}\Delta\boldsymbol{\theta}^{(t+1)}\|^2]^{\frac{1}{2}}}, \tag{12}$$

where β denotes the cumulation factor of the evolution path and $\mathbb{E}[\cdot]$ indicates the expectation under the random selection. The left multiplication of the square root $\boldsymbol{\mathcal{I}}_{\theta^{(t)}}^{1/2}$ of the Fisher information matrix in the second term of the RHS of (12) provides the invariance property against parameterization of the probability distribution. Additionally, the parameter movement is normalized by $\mathbb{E}[\|\boldsymbol{\mathcal{I}}_{\theta^{(t)}}^{\frac{1}{2}}\Delta\boldsymbol{\theta}^{(t+1)}\|^2]^{\frac{1}{2}}$ to avoid scaling of the parameter movement due to the change of the population size. We use the approximated value below

$$\begin{aligned} \mathbb{E}[\|\boldsymbol{\mathcal{I}}_{\theta^{(t)}}^{\frac{1}{2}}\Delta\theta^{(t+1)}\|^2] \approx{}& \frac{nc_m^2}{\mu_\mathrm{w}} + \frac{2n(n-\chi_n^2)}{\chi_n^2}\gamma_\sigma^{(t+1)}\left(\frac{c_\sigma}{d_\sigma}\right)^2 \\ &+ \frac{1}{2}\left[1 + 8\gamma_\sigma^{(t+1)}\frac{n-\chi_n^2}{\chi_n^2}\left(\frac{c_\sigma}{d_\sigma}\right)^2\right]\left[\frac{(n^2+n)c_\mu^2}{\mu_\mathrm{w}}\right. \\ &+ (n^2+n)c_c(2-c_c)c_1c_\mu\mu_\mathrm{w}\textstyle\sum_{i=1}^{\lambda}\mathrm{w}_i^3 \\ &\left.+ c_1^2((\gamma_c^{(t+1)})^2n^2 + (1-2\gamma_c^{(t+1)} + 2(\gamma_c^{(t+1)})^2)n)\right]. \end{aligned} \tag{13}$$

See the supplementary material for the derivation.

Analogously to the normalization factors for $\boldsymbol{p}_\sigma$ and $\boldsymbol{p}_c$, the normalization factor for $\boldsymbol{p}_\theta$ is updated as

$$\gamma_\theta^{(t+1)} = (1-\beta)^2\gamma_\theta^{(t)} + \beta(2-\beta), \tag{14}$$

where $\gamma_\theta^{(0)} = 0$ since we initialize $\boldsymbol{p}_\theta^{(0)} = \boldsymbol{0}$ in this paper. Note that, the normalization factor γ_θ converges to 1 as t increases.

Note that, this evolution path is defined in $\mathbb{R}^{n(n+3)/2}$ in contrast to the existing evolution paths $\boldsymbol{p}_c, \boldsymbol{p}_\sigma$ defined in $\mathbb{R}^n$. That is, the latter focus on the movement of the mean vector, but the former considers changes of all the distribution parameters. When the objective function is linear, the mean vector keeps to move toward the almost same direction and the step-size keeps increasing, therefore the evolution path gets longer. When the objective function is the

sphere function, the mean vector is updated like a random walk when it is relatively close to the optimal point. The step-size and the covariance matrix however gets smaller continuously. In this case, the evolution path gets longer due to the high directional change of Σ. The latter case is captured thanks to the evolution path in the distribution parameter space, rather than in the search space.

We utilize the squared norm $\|\boldsymbol{p}_\theta^{(t)}\|^2$, i.e. length, of the proposed evolution path as the measure of accuracy of the parameter update. The expectation $\mathbb{E}[\|\boldsymbol{p}_\theta^{(t)}\|^2]$ under the random selection converges to 1 as t increases. Therefore, when the length $\|\boldsymbol{p}_\theta^{(t)}\|^2$ is close to 1, we can assume that the parameter update has less accuracy. On the other side, we consider that the parameter update has relatively high accuracy when the length $\|\boldsymbol{p}_\theta^{(t)}\|^2$ is greater than 1.

3.2 Population Size Update

We adapt the population size to keep the parameter update sufficiently accurate. That is, we adapt the population size so that the length $\|\boldsymbol{p}_\theta\|^2$ of evolution path is close to a target value α. The update rule is as follows:

$$\lambda^{(t+1)} = \lambda^{(t)} \exp\left(\beta\left(\gamma_\theta^{(t+1)} - \frac{\|\boldsymbol{p}_\theta^{(t+1)}\|^2}{\alpha}\right)\right). \tag{15}$$

We restrict the population size as

$$\lambda^{(t+1)} = \min(\max(\lambda^{(t+1)}, \lambda_{\min}), \lambda_{\max}) \tag{16}$$

after (15) if necessary. The lower limit should be $\lambda_{\min} \geqslant 4$ due to the basic property of the CMA-ES. The default value $\lambda_{\text{def}} = 4 + \lfloor 3\ln(n) \rfloor$ in the CMA-ES may be a good candidate, and we set $\lambda_{\min} = \lambda_{\text{def}}$ in this paper. Note that $\lambda^{(t+1)}$ is real-valued. We use the rounded population size $\lambda_{\text{r}}^{(t)} = \text{round}(\lambda^{(t)})$ in the CMA-ES, but keep $\lambda^{(t)}$ real-valued in the adaptation mechanism.

We can introduce the damping factor as is introduced in the step-size adaptation (8). The damping can reduce variance of the population size change, however, it makes the adaptation slower at the same time. It is often better to increase immediately if the population size is too small. Therefore we omit the damping in this paper.

3.3 Step-Size Correction

It has been derived from the quality gain analysis [2, 3] that the optimal standard deviation of the sampling distribution is proportional to μ_{w} on a convex quadratic function in the limit of n to infinity. It implies that the step-size is increased when the population size is increased, and vice versa. A more explicit scaling factor of the optimal standard deviation has been provided in [1].

A practical step-size adaptation in the CMA-ES usually well follows the optimal value [8, 13]. It indicates that the step-size will be increased when the population size is increased. Then, an artificial tendency of the step-size change is captured in the evolution path, resulting in decreasing the population size. This unpleasant behavior was observed in our preliminary study and led to unstable adaptation of the population size and the step-size.

Algorithm 1: PSA-CMA-ES

input : $\boldsymbol{m}^{(0)} \in \mathbb{R}^n, \sigma^{(0)} \in \mathbb{R}_+$
set : $c_m = 1$, $\alpha = 1.4$, $\beta = 0.4$, $\lambda_{\min} = \lambda_{\text{def}}$, $\lambda_{\max} = \infty$
initialize : $\boldsymbol{C}^{(0)} = \boldsymbol{I}, \boldsymbol{p}_c^{(0)} = \boldsymbol{0}, \boldsymbol{p}_\sigma^{(0)} = \boldsymbol{0}, \boldsymbol{p}_\theta^{(0)} = \boldsymbol{0}, \gamma_c^{(0)} = 0, \gamma_\sigma^{(0)} = 0, \gamma_\theta^{(0)} = 0, \lambda^{(0)} = \lambda_r^{(0)} = \lambda_{\text{def}}, t = 0$

while *not terminate* **do**

- `// (re-)compute parameters depending on` λ
- $\mu \leftarrow \lfloor \lambda_r^{(t)}/2 \rfloor$
- $w_i \leftarrow \frac{\log(\mu+0.5)-\log i}{\sum_{i=1}^{\mu}(\log(\mu+0.5)-\log i)} \quad (i = 1, \ldots, \mu)$
- $w_i \leftarrow 0 \quad (i = \mu+1, \ldots, \lambda_r^{(t)})$
- $\mu_{\text{eff}} \leftarrow 1/\sum_{i=1}^{\lambda_r^{(t)}} w_i^2$
- $c_\sigma \leftarrow (\mu_{\text{eff}} + 2)/(n + \mu_{\text{eff}} + 5)$
- $d_\sigma \leftarrow 1 + 2\max(0, \sqrt{(\mu_{\text{eff}} - 1)/(n+1)} - 1) + c_\sigma$
- $c_c \leftarrow (4 + \mu_{\text{eff}}/n)/(n + 4 + 2\mu_{\text{eff}}/n)$
- $c_1 \leftarrow 2/((n+1.3)^2 + \mu_{\text{eff}})$
- `// perform a CMA-ES iteration`
- (1),...,(9) in Sec. 2
- `// update evolution path and its factor`
- $\boldsymbol{p}_\theta^{(t+1)} \leftarrow (1-\beta)\boldsymbol{p}_\theta^{(t)} + \sqrt{\beta(2-\beta)}\frac{\boldsymbol{I}_{\theta^{(t)}}^{\frac{1}{2}} \Delta\boldsymbol{\theta}^{(t+1)}}{\mathbb{E}[\|\boldsymbol{I}_{\theta^{(t)}}^{\frac{1}{2}} \Delta\boldsymbol{\theta}^{(t+1)}\|^2]^{\frac{1}{2}}}$
- $\gamma_\theta^{(t+1)} \leftarrow (1-\beta)^2\gamma_\theta^{(t)} + \beta(2-\beta)$
- `// update population size`
- $\lambda^{(t+1)} \leftarrow \lambda^{(t)} \exp\left(\beta\left(\gamma_\theta^{(t+1)} - \frac{\|\boldsymbol{p}_\theta^{(t+1)}\|^2}{\alpha}\right)\right)$
- $\lambda^{(t+1)} \leftarrow \min(\max(\lambda^{(t+1)}, \lambda_{\min}), \lambda_{\max})$
- $\lambda_r^{(t+1)} \leftarrow \text{round}(\lambda^{(t+1)})$
- `// step-size correction`
- $\sigma^{(t+1)} \leftarrow \sigma^{(t+1)}\frac{\sigma^*(\lambda_{\text{r}}^{(t+1)})}{\sigma^*(\lambda_{\text{r}}^{(t)})}$
- $t \leftarrow t + 1$

end

To make the population size adaptation stable, we update the step-size after updating the population size according to

$$\sigma^{(t+1)} \leftarrow \sigma^{(t+1)}\frac{\sigma^*(\lambda_{\text{r}}^{(t+1)})}{\sigma^*(\lambda_{\text{r}}^{(t)})} \tag{17}$$

where $\sigma^*(\lambda_{\text{r}})$ is the scaling factor of the optimal standard deviation derived in [1] given a rounded population size λ_{r}, whose approximated value below is used in our implementation,

$$\sigma^*(\lambda_{\text{r}}) = \frac{c \cdot n \cdot \mu_{\text{w}}}{n - 1 + c^2 \cdot \mu_{\text{w}}}\ , \tag{18}$$

where $c = -\sum_{i=1}^{\lambda_{\text{r}}} w_i \mathbb{E}[\mathcal{N}_{i:\lambda_{\text{r}}}]$ is the weighted average of the expected value of the normal order statistics from λ_{r} population and is usually in $O(1)$.

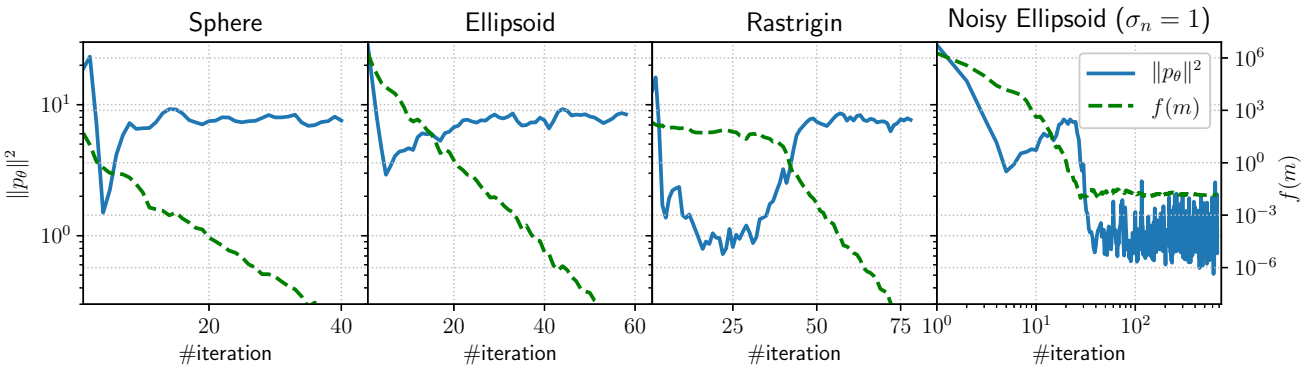


Figure 1: Typical behavior of the proposed evolution path in the CMA-ES with a fixed population size ($\lambda = 400$) on 10d functions.

Table 1: Function Definition

Function	Init
$f_{\text{Sphere}}(x) = \sum_{i=1}^{n} x_i^2$	$[1, 5]^n$
$f_{\text{Ellipsoid}}(x) = \sum_{i=1}^{n} 10^{\frac{6(i-1)}{n-1}} x_i^2$	$[1, 5]^n$
$f_{\text{Rastrigin}}(x) = \sum_{i=1}^{n} \left(x_i^2 + 10(1 - \cos 2\pi x_i)\right)$	$[1, 5]^n$
$f_{\text{Schaffer}}(x) = \sum_{i=1}^{n-1} [x_i^2 + x_{i+1}^2]^{1/4} \cdot [\sin^2(50(x_i^2 + x_{i+1}^2)^{0.1}) + 1]$	$[10, 100]^n$

3.4 PSA-CMA-ES

The resulting algorithm is called the PSA-CMA-ES, the CMA-ES with population size adaptation. At each iteration, we perform a single CMA-ES iteration explained in Sec. 2 with population size $\lambda_{\text{r}}^{(t)} = \text{round}(\lambda^{(t)})$. The strategy parameters depending directly or indirectly on the population size are recomputed beforehand. After the main iteration, the evolution path is updated (Sec. 3.1) and the population size is adapted (Sec. 3.2), followed by the step-size correction (Sec. 3.3). Its pseudo-code is provided in Algorithm 1.

4 EXPERIMENT

In this section, we show how the evolution path behaves when the CMA-ES with a fixed population size is applied to some functions. Then, we demonstrate how the population size is adapted in the proposed algorithm on noiseless and noisy functions and compare it to the CMA-ES with a fixed population size. We finally compare the proposed algorithm with existing population size adaptive algorithms, namely the CMAES-APOP and the pcCMSA-ES.

4.1 Experiment Setting

The test functions are defined in the Table 1. For all the functions, the global optimal solution is located at $\boldsymbol{x}^* = \mathbf{0}$ and its function value $f(\boldsymbol{x}^*)$ is zero. In noisy scenario, we consider the additive Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma_{\text{n}}^2)$ with the standard deviation σ_{n}.

The initial mean vector $\boldsymbol{m}^{(0)}$ is sampled uniformly within the initialization interval shown in Table 1 and the initial step-size $\sigma^{(0)}$ is set to the half of the initialization interval length.

The strategy parameters for the PSA-CMA-ES is: $\lambda_{\text{min}} = \lambda_{\text{def}} = 4 + \lfloor 3 \ln(n) \rfloor$ (default λ for the CMA-ES), $\lambda_{\text{max}} = \infty$, $\alpha = 1.4$, and $\beta = 0.4$. The default values for α and β were determined based on our preliminary parameter survey. The other strategy parameters appeared in the CMA-ES are set to their default values. See Algorithm 1 (Line 2-10) for the default values.

4.2 Behavior of the Evolution Path

We demonstrate a typical behavior of the proposed evolution path in the CMA-ES with a fixed population size on 10d functions in Figure 1. The horizontal axis indicates the number of iterations. The population size is set to 400 , which provides a relatively good performance on the 10d Rastrigin function [9].

On the unimodal functions, where the default population size is sufficient, the length of the evolution path was kept greater than 1. On the other hand, when solving the Rastrigin function, which has many local minima and a large population size is preferable on, we observed a phase transition of the length of the evolution path. In the early phase, the length $\|\boldsymbol{p}_\theta\|^2$ was nearly 1, which means the parameter update is nearly at random. In the latter phase where the convergence slope became as steep as it was on the Sphere function, the length $\|\boldsymbol{p}_\theta\|^2$ increased up to a value similar to that observed on the unimodal functions.

When the objective function is the noisy Ellipsoid function, the evolution path dynamics also had two phases. In the early phase when the selection is not corrupted by a noise, the length $\|\boldsymbol{p}_\theta\|^2$ was sufficiently greater than 1. It behaved similarly as it did on the noiseless Ellipsoid function. In the latter phase where the interquartile range (IQR) of the function values of candidate solutions became close to the IQR of the noise (approximately 1.3489), the length $\|\boldsymbol{p}_\theta\|^2$ decreased to nearly 1.

4.3 Behavior of the Population Size

Figure 2 shows a typical behavior of the PSA-CMA-ES on noiseless unimodal/multimodal and noisy unimodal/multimodal functions.

On the unimodal functions (Sphere and Ellipsoid), the population size is kept in relatively small values. The stable value was a bit greater than the default value, especially at the beginning on the Ellipsoid function. We consider that it is not a big issue as the

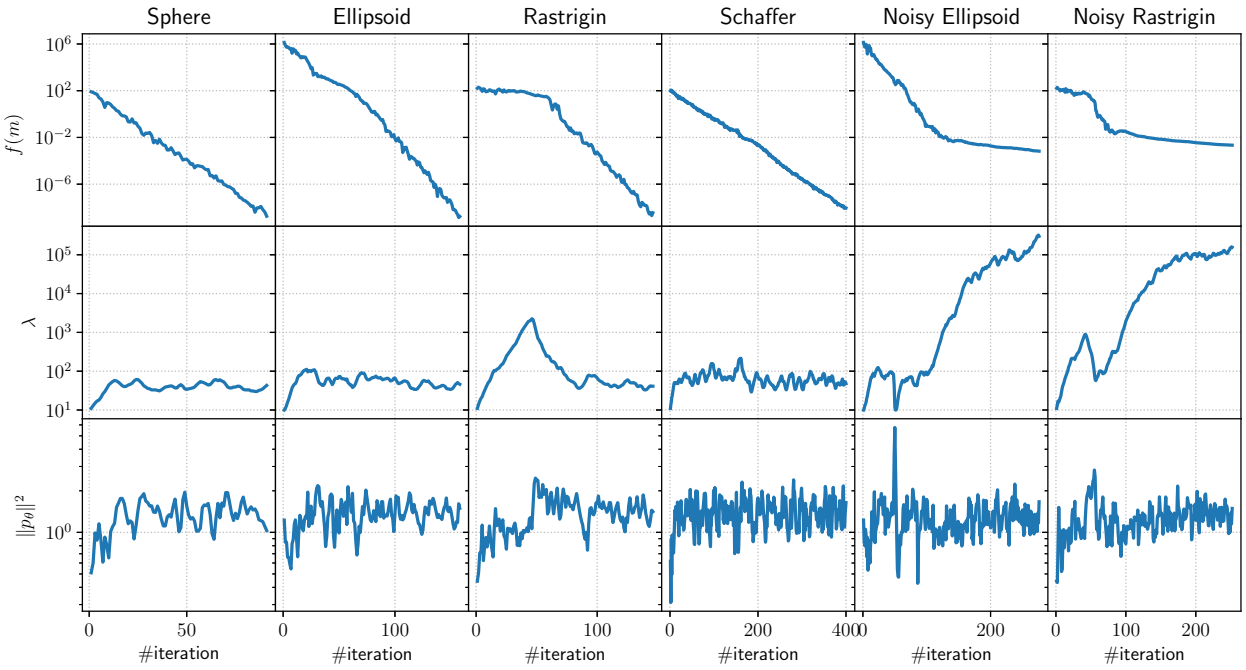


Figure 2: Typical runs of the PSA-CMA-ES on 10d functions. The standard deviation σ_n of noise is set to 1 in noisy scenario.

covariance matrix adaptation can benefit from a large population. We can decrease the adapted population size by tuning the parameters, i.e., by decreasing α and increasing β, but it often degrades the performance on multimodal functions.

On the multimodal functions (Rastrigin and Schaffer), the population size increased at the beginning and decreased after the probability mass of the sampling distribution was considered to be concentrated at a basin of a local minimum. The population size decreased to a similar value as it was kept on the Sphere function.

When the objective function has a noise, the population size behaved as if it is on the noiseless function at an early phase where the selection was not corrupted by a noise. When the noise started to affect the selection, the population size increased corresponding to the NSR. Noteworthy, the PSA-CMA-ES can tackle a function with both noise and multimodality.

4.4 Comparison

Figure 3 and Figure 4 show the performance of the PSA-CMA-ES, CMA-ES with a fixed population size ($\lambda = \lambda_{\text{def}}, \lambda^2_{\text{def}}, \lambda^3_{\text{def}}$), the CMA-ES-APOP [14] and the pcCMSA-ES [12]. We employ the SP1 [4], the average number of function evaluations until reaching the target function value among successful trials divided by the success rate, as the performance measurement on the noiseless testbed. To measure and compare the performance of different algorithms, we also employ the empirical cumulative density function used in the COmparing Continuous Optimizers (COCO) framework[1]. We define N_{target} target values. We record the number of function evaluations spent until the noiseless function value $f(\boldsymbol{m})$ hits a smaller value than each target value for the first time. The data is collected by running N_{trial} independent trials. In total, we have $N_{\text{target}} \cdot N_{\text{trial}}$ targets for each setting. Figure 4 shows the proportion of the target values reached within each number of function evaluations. The target values are set to $10^{6-9(i-1)/(N_{\text{target}}-1)}$ for $i = 1, \ldots, N_{\text{target}}$, and the number of trials is $N_{\text{trial}} = 20$, resulting in 500 targets for each setting. Note that the figures are essentially related with the convergence graph in log-scale in the range of $[10^{-3}, 10^{6}]$ if they are flipped vertically, and they agree if $N_{\text{trial}} = 1$.

[1] http://coco.gforge.inria.fr/

4.4.1 Noiseless Testbed. We compare the PSA-CMA-ES, pcCMSA-ES, and the CMA-ES with different population sizes in Figure 3. The vertical axis indicates the SP1 of each algorithm divided by dimension and the horizontal axis indicates the dimension ($n = 2, 3, 5, 10, 20, 40$) of the objective function.

The CMA-ES with the default population size outperformed the other algorithms in the unimodal functions. On the other hand, it could not find the global minimum in most cases when solving multimodal functions. The larger population size improved the performance of the CMA-ES, however, the best population size depends on the function to be solved. In contrast, the PSA-CMA-ES showed relatively good performance on all of the functions without tuning the population size in advance.

The PSA-CMA-ES outperformed the CMAES-APOP on all the functions excluding the Rastrigin function and the 20d Schaffer function. However, the superiority in the unimodal functions may not be important because the CMA-ES-APOP have been proposed with a restart strategy, where the CMA-ES with the default population size is applied for the first search. A comparison of the

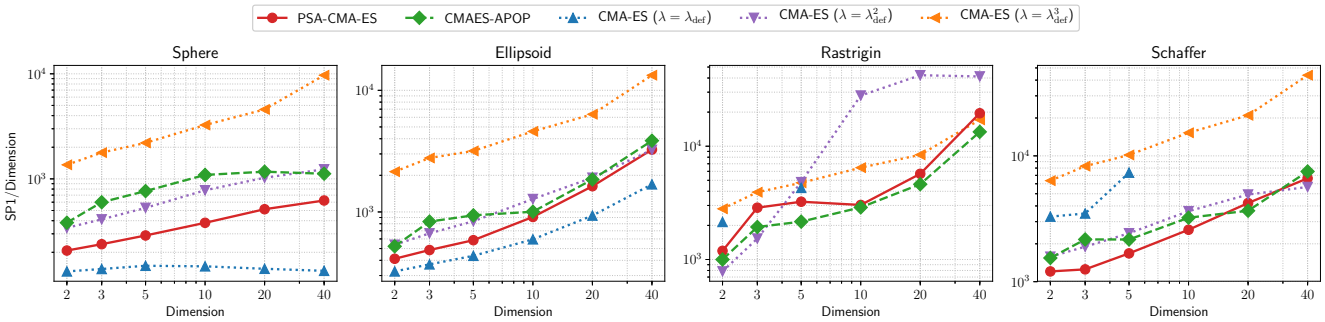
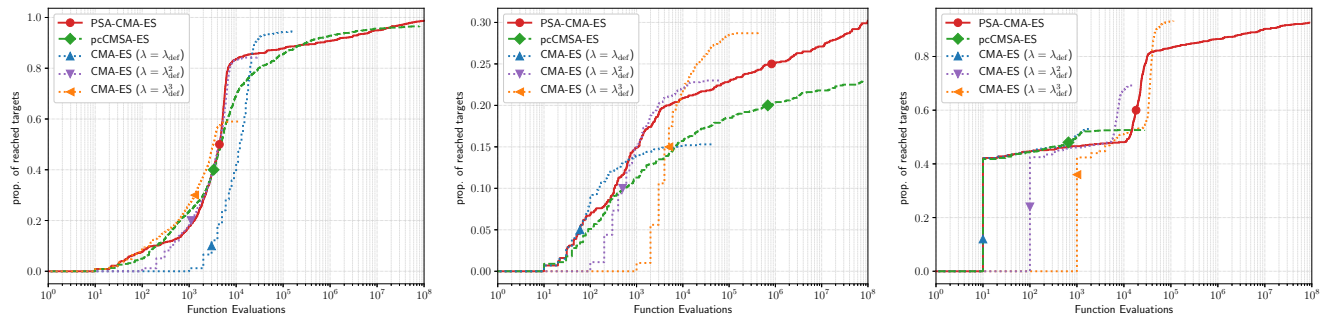


Figure 3: SP1 divided by dimension versus dimension. (noiseless testbed)



(a) Noisy Ellipsoid ($\sigma_{\mathrm{n}} = 1$) **(b) Noisy Ellipsoid ($\sigma_{\mathrm{n}} = 10^6$)** **(c) Noisy Rastrigin ($\sigma_{\mathrm{n}} = 1$)**

Figure 4: Performance measure on 10d noisy functions

PSA-CMA-ES and the CMAES-APOP using the same restart strategy on a wider testbed is required in the future work.

The performance of the PSA-CMA-ES got a little worse as the dimension n increased. It might be because the cumulation factor β of the evolution path $\boldsymbol{p}_\theta$ is set to a constant value independently of the dimension, unlikely to the cumulation factors for the other evolution paths in the CMA-ES. A further investigation of the hyperparameter of the PSA-CMA-ES is necessary.

4.4.2 Noisy Testbed. Figure 4 shows the empirical cumulative distribution graph for the PSA-CMA-ES, pcCMSA-ES, and the CMA-ES with different population sizes.

With a fixed population size, the CMA-ES stopped improving the solution quality if the objective function is corrupted by a noise. The saturation point depends heavily on the population size, that is, a larger population size results in a better but sub-optimal solution. Therefore, a fixed population size can not be sufficient to approach the global optimum under an additive Gaussian noise.

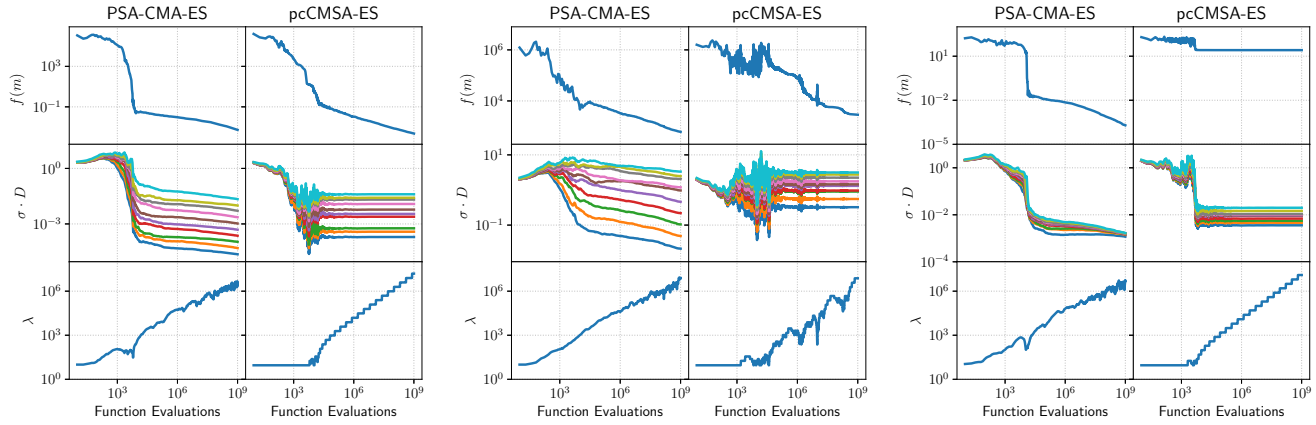
The PSA-CMA-ES kept improving the solution quality during the optimization of noisy functions, including the noisy Rastrigin function that has both noise and multimodality. The pcCMSA-ES also kept improving the solution quality on the noisy Ellipsoid function, but the proportion of reached targets tended to level out. Moreover, the pcCMSA-ES reached only the number of targets similar to that of the CMA-ES with the default population size when it applied to the noisy Rastrigin function.

Figure 5 shows that the covariance matrix approximated the inverse Hessian of the expected objective function, f, in the PSA-CMA-ES, whereas the pcCMSA-ES stops adapting the covariance matrix once it detects a noise, hence the covariance matrix did not approximate the inverse Hessian and the speed of approaching the global optimum was slower than the PSA-CMA-ES. Moreover, since the pcCMSA-ES is designed for noisy, but essentially unimodal functions, it could not increase the population size when it is necessary to find a basin of attraction of the global optimum on the noisy Rastrigin function. It started increasing the population size once a noise was detected, but at that time the search distribution was already concentrated at a basin of attraction of a local minimum and ended up exploitation of a local minimum.

5 CONCLUSION

We have presented the novel population size adaptation mechanism for the CMA-ES, called PSA-CMA-ES. The population size is adapted based on the accuracy of the update of the distribution parameters. The accuracy is quantified by the length of the evolution path in the distribution parameter space. The evolution path is almost invariant against parameterization of the distribution. The mechanism includes the step-size correction to stabilize the population size dynamics.

We have evaluated the PSA-CMA-ES in the noiseless and noisy testbeds. On the noiseless unimodal functions, the population size



(a) Noisy Ellipsoid ($\sigma_n = 1$) **(b) Noisy Ellipsoid ($\sigma_n = 10^6$)** **(c) Noisy Rastrigin ($\sigma_n = 1$)**

Figure 5: Typical behavior of the PSA-CMA-ES and the pcCMSA-ES on 10d noisy functions. $\sigma \cdot D$ indicates the square root of the eigenvalues of the covariance matrix $\Sigma = \sigma^2 \cdot C$. It is desirable for efficient search that the ratio of its maximum value and its minimum value is close to 10^3 (Ellipsoid) or 1 (Rastrigin).

was kept in the relatively small values. On the noiseless multimodal functions, the population size increased at the beginning and decreased after the probability mass of the sampling distribution is concentrated at a single basin of a local minimum. When the objective function has a noise, the population size is adapted corresponding to the noise-to-signal strength. The PSA-CMA-ES could hence tackle functions with noise and multimodality. The comparison with the CMA-ES with fixed population sizes, the CMA-ES-APOP, and the pcCMSA-ES revealed a relatively good performance of the PSA-CMA-ES in all the test scenarios in this paper without tuning any hyper-parameter problem-by-problem. One of the strong point of the PSA-CMA-ES compared with the other population size adaptation mechanisms is that it is effective on a function with both noise and multimodality.

On the other hand, we found that the PSA-CMA-ES loses a little efficiency as the dimension increases. To tackle this issue, we have two directions of the future work. One is to set the hyper-parameter of the proposed mechanism depending on the dimension to relax the dependency of the performance on the dimension. The other is to utilize a restart strategy where the CMA-ES with the default population size is applied for the first run. A restart strategy is expected to be effective on weakly-structured multimodal functions as well. By incorporating a restart regime using a relatively small initial step size as in the BIPOP strategy, the PSA-CMA-ES is expected to be useful on weakly-structured multimodal functions. Designing a restart strategy specialized for the PSA-CMA-ES is an important future work. In the future work, we will also conduct a thorough experiments using a standard benchmarking such as BBOB.

REFERENCES

[1] Youhei Akimoto, Anne Auger, and Nikolaus Hansen. 2017. Quality Gain Analysis of the Weighted Recombination Evolution Strategy on General Convex Quadratic Functions. (2017). arXiv:1608.04813v5

[2] Youhei Akimoto, Anne Auger, and Nikolaus Hansen. 2017. Quality Gain Analysis of the Weighted Recombination Evolution Strategy on General Convex Quadratic Functions. In *Foundations of Genetic Algorithms, FOGA XIV, Copenhagen, Denmark, January 12-15, 2017*. ACM, 111–126.

[3] Dirk V. Arnold. 2005. Optimal weighted recombination. In *Foundations of Genetic Algorithms*. Springer, 215–237.

[4] Anne Auger and Nikolaus Hansen. 2005. A Restart CMA Evolution Strategy With Increasing Population Size. In *2005 IEEE Congress on Evolutionary Computation*. Ieee, 1769–1776.

[5] Nikolaus Hansen. 2009. Benchmarking a BI-population CMA-ES on the BBOB-2009 function testbed. In *Workshop Proceedings of the GECCO Genetic and Evolutionary Computation Conference*. ACM Press, New York, NY, USA, 2389–2395.

[6] Nikolaus Hansen. 2009. Benchmarking a BI-population CMA-ES on the BBOB-2009 noisy testbed. In *GECCO '09: Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference: Late Breaking Papers*. ACM Request Permissions.

[7] Nikolaus Hansen. 2016. The CMA Evolution Strategy: A Tutorial. *ArXiv e-prints* (April 2016). arXiv:cs.LG/1604.00772

[8] Nikolaus Hansen, Asma Atamna, and Anne Auger. 2014. How to Assess Step-Size Adaptation Mechanisms in Randomised Search. In *Parallel Problem Solving from Nature–PPSN XIII*. Springer, 60–69.

[9] Nikolaus Hansen and Stefan Kern. 2004. Evaluating the CMA Evolution Strategy on Multimodal Test Functions. In *Parallel Problem Solving from Nature - PPSN VIII*. Springer, 282–291.

[10] Nikolaus Hansen, Sibylle D. Muller, and Petros Koumoutsakos. 2003. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evolutionary Computation* 11, 1 (2003), 1–18.

[11] Nikolaus Hansen and Andreas Ostermeier. 2001. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation* 9, 2 (2001), 159–195.

[12] Michael Hellwig and Hans-Georg Beyer. 2016. Evolution Under Strong Noise: A Self-Adaptive Evolution Strategy Can Reach the Lower Performance Bound-The pcCMSA-ES. In *International Conference on Parallel Problem Solving from Nature*. Springer, 26–36.

[13] Oswin Krause, Tobias Glasmachers, and Christian Igel. 2017. Qualitative and Quantitative Assessment of Step Size Adaptation Rules. In *Proceedings of the 14th ACM/SIGEVO Conference on Foundations of Genetic Algorithms (FOGA '17)*. ACM, New York, NY, USA, 139–148.

[14] Duc Manh Nguyen and Nikolaus Hansen. 2017. Benchmarking CMAES-APOP on the BBOB Noiseless Testbed. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion (GECCO '17)*. ACM, New York, NY, USA, 1756–1763.

[15] Kouhei Nishida and Youhei Akimoto. 2016. Population Size Adaptation for the CMA-ES Based On the Estimation Accuracy of the Natural Gradient. In *Genetic and Evolutionary Computation Conference, GECCO 2016, Denver, Colorado, USA, July 20-24, 2016*. ACM, 237–244.

[16] Yann Ollivier, Ludovic Arnold, Anne Auger, and Nikolaus Hansen. 2017. Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles. *Journal of Machine Learning Research* 18, 18 (2017), 1–65.