

# Adaptive Generation-Based Evolution Control for Gaussian Process Surrogate Models

Jakub Repický<sup>1,2</sup>, Lukáš Bajer<sup>1,2</sup>, Zbyněk Pitra<sup>2,3,4</sup>, and Martin Holeňa<sup>2</sup>

<sup>1</sup> Faculty of Mathematics and Physics, Charles University in Prague  
Malostranské nám. 25, 118 00 Prague 1, Czech Republic

<sup>2</sup> Institute of Computer Science, Czech Academy of Sciences  
Pod Vodárenskou věží 2, 182 07 Prague 8, Czech Republic

<sup>3</sup> Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague  
Břehová 7, 115 19 Prague 1, Czech Republic

<sup>4</sup> National Institute of Mental Health  
Topolová 748, 250 67 Klecany, Czech Republic  
{repicky,bajer,pitra,martin}@cs.cas.cz

*Abstract:* The interest in accelerating black-box optimizers has resulted in several surrogate model-assisted version of the Covariance Matrix Adaptation Evolution Strategy, a state-of-the-art continuous black-box optimizer. The version called Surrogate CMA-ES uses Gaussian processes or random forests surrogate models with a generation-based evolution control. This paper presents an adaptive improvement for S-CMA-ES, in which the number of generations using the surrogate model before retraining is adjusted depending on the performance of the last instance of the surrogate. Three algorithms that differ in the measure of the surrogate model's performance are evaluated on the COCO/BBOB framework. The results show a minor improvement on S-CMA-ES with constant model lifelengths, especially when larger lifelengths are considered.

## 1 Introduction

The problem of optimization of real-valued functions without a known mathematical expression, arising in many engineering tasks, is referred to as continuous black-box optimization. Evolutionary strategies, a class of randomized population-based algorithms inspired by natural evolution, are a popular choice for continuous black-box optimization. Especially the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [6] is considered the state-of-the-art continuous black-box optimizer of the several past decades. Since values of a black-box function can only be obtained empirically and at considerable costs in practice, the number of function evaluations needed to obtain a desired function value is a key criterion for evaluating black-box optimizers.

The technique of surrogate modelling aims at saving function evaluations by building a surrogate model of the fitness and using that for a portion of function evaluations conducted in the course of the evolutionary search. Several surrogate model-assisted versions of the CMA-ES have been developed (see [11] for a recent comparison of some of the most notable algorithms). Surrogate CMA-ES (S-CMA-ES) [2] utilizes random forests- or Gaussian

processes-based surrogate models, which possess an inherent capability to quantify uncertainty of the prediction.

In order to control surrogate model's error, S-CMA-ES uses the surrogate model for a given number of generations  $g_m$  before a new instance of the model is trained on a population evaluated with the fitness, which is a strategy called generation-based evolution control [9]. In [2], two model lifelengths, in particular  $g_m \in \{1, 5\}$ , have been benchmarked on the COCO/BBOB framework. In many cases, the higher lifelength outperformed the lower one in earlier phases of the optimization, but the reverse order was observed towards later phases of the optimization. In this paper, we propose an adaptive version of the S-CMA-ES (A-S-CMA-ES), in which an adequate number of model-evaluated generations is estimated as a function of previous model's error. We restrict our attention to Gaussian processes, since they outperformed random forest-based surrogates [2]. We consider three different surrogate error measures and compare all of them on the COCO/BBOB framework.

The remainder of this paper is organized as follows. Section 2 outlines basic concepts of S-CMA-ES. The proposed adaptive improvement is described in Section 3. Experimental setup is given in Section 4. Experimental results are reported in Section 6. Section 7 concludes the paper.

## 2 Surrogate CMA-ES

The CMA-ES operates on a population of  $\lambda$  candidate solutions sampled from a multivariate normal distribution:

$$\mathbf{x}_k \sim \mathcal{N}(\mathbf{m}, \sigma^2 \mathbf{C}) \quad k = 1, \dots, \lambda, \quad (1)$$

where  $\mathcal{N}$  is the normal distribution function;  $\mathbf{m}$  and  $\mathbf{C}$  are the mean and the covariance matrix of the estimated search distribution, respectively; and the  $\sigma$  is the overall search step size. The candidate solutions are ranked according to their fitness values:

$$y_k = f(\mathbf{x}_k) \quad k = 1, \dots, \lambda. \quad (2)$$

Upon a (weighted) selection of  $\mu < \lambda$  highest ranked points, the mean and the covariance matrix of the multivariate normal distribution are adapted according to a procedure that takes as input, among other variables, a cumulation of the past search steps [5]. The S-CMA-ES modifies the CMA-ES by replacing its sampling (1) and fitness-evaluation (2) steps with a procedure depicted in Algorithm 1.

---

**Algorithm 1** Surrogate part of S-CMA-ES
 

---

**Input:**  $g$  (generation)

$g_m$  (number of model-evaluated generations)

$\sigma, \lambda, \mathbf{m}, \mathbf{C}$  (CMA-ES internal variables)

$r$  (maximal distance between  $\mathbf{m}$  and a training point)

$n_{\text{req}}$  (minimal number of points for training)

$n_{\text{max}}$  (maximal number of points for training)

$\mathcal{A}$  (archive),  $f_{\mathcal{M}}$  (model),  $f$  (fitness)

1:  $\mathbf{x}_k \sim \mathcal{N}(\mathbf{m}, \sigma^2 \mathbf{C}) \quad k = 1, \dots, \lambda$  {sampling}

2: **if**  $g$  is original-fitness-evaluated **then**

3:  $y_k \leftarrow f(\mathbf{x}_k) \quad k = 1, \dots, \lambda$  {fitness evaluation}

4:  $\mathcal{A} \leftarrow \mathcal{A} \cup \{(\mathbf{x}_k, y_k)\}_{k=1}^{\lambda}$

5:  $(\mathbf{X}_{\text{tr}}, \mathbf{y}_{\text{tr}}) \leftarrow$  choose  $n_{\text{tr}}$  training points within the Mahalanobis distance  $r$  from  $\mathcal{A}$ , assuring that  $n_{\text{req}} \leq n_{\text{tr}} \leq n_{\text{max}}$

6:  $f_{\mathcal{M}} \leftarrow \text{train\_model}(\mathbf{X}_{\text{tr}}, \mathbf{y}_{\text{tr}})$

7: mark  $(g+1)$  as model-evaluated

8: **else**

9:  $\hat{y}_k \leftarrow f_{\mathcal{M}}(\mathbf{x}_k)$  {model evaluation}

10: **if**  $g_m$  model generations have passed **then**

11: mark  $(g+1)$  as original-fitness-evaluated

12: **end if**

13: **end if**

**Output:**  $f_{\mathcal{M}}, \mathcal{A}, (y_k)_{k=1}^{\lambda}$

---

Depending on the generation number  $g$ , the procedure evaluates all candidate solutions either with the real fitness or with the model. In each case, the sampling of the estimated multivariate normal distribution is unchanged (step 1).

If the population is original-fitness-evaluated (step 3), the new evaluations are saved in an archive of known solutions (step 4). Afterwards, a new model is trained on a set of points within the Mahalanobis distance  $r$  from the current CMA-ES distribution  $\mathcal{N}(\mathbf{m}, \sigma \mathbf{C})$  (step 5).

In model-evaluated generations, the fitness values of the whole population of candidate solutions are estimated by the model (step 9).

## 2.1 Gaussian Processes

A Gaussian process (GP) is a collection of random variables  $(f(\mathbf{x}))_{\mathbf{x} \in \mathbb{R}^D}$ , such that any finite subcollection  $\mathbf{f} = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_N))$  has an  $N$ -dimensional normal distribution. A Gaussian process is defined by a mean function  $\mu(\mathbf{x})$  (often assumed to be zero) and a covariance function  $k(\mathbf{x}, \mathbf{x}; \theta)$ , where  $\theta$  is a vector of parameters of  $k$ , hence

hyperparameters of the Gaussian process. Given a set of training data  $X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ , the covariance matrix of a GP prior is  $\mathbf{K}_N + \sigma_n^2 \mathbf{I}_N$ , where  $\mathbf{K}_N$  is a  $N \times N$  matrix given by  $\{\mathbf{K}_N\}_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j; \theta)$  for all  $i, j = 1, \dots, N$ ;  $\sigma_n^2$  is the variance of an additive, i. i. d. noise and  $\mathbf{I}_N$  is a  $N \times N$  identity matrix. Given a new point  $\mathbf{x}_* \notin X$ , Gaussian process regression is derived by conditioning the joint normal distribution of  $(f(\mathbf{x}_1), \dots, f(\mathbf{x}_N), f(\mathbf{x}_*))$  on the prior, which yields a univariate Gaussian (see [12] for more details). The hyperparameters  $\theta$  of a GP regression model are estimated using the maximum likelihood estimation method.

## 3 Adaptive Evolution Control for Surrogate CMA-ES

The generation-based evolution strategy optimizes the fitness function and the surrogate model thereof in certain proportion. On problem areas that can be approximated well, a surrogate-assisted optimization might benefit from frequent utilization of the model, while on areas that are hard for the surrogate to approximate, frequent utilization of the model might degenerate the performance due to the model's inaccuracy.

The proposed evolution control adapts the lifelength of a new model depending on the previous model's error measured on the current generation.

Consider a generation  $g$  that is marked as original-fitness-evaluated, and a newly-trained surrogate model  $f_{\mathcal{M}}$ . If  $f_{\mathcal{M}}$  is the first surrogate trained so far, put  $g_m = 1$ . Otherwise, an error  $\varepsilon$  of a previous surrogate model  $f_{\mathcal{M}}^{\text{last}}$  is estimated on the newly evaluated population  $(\mathbf{x}_1^{(g+1)}, \dots, \mathbf{x}_{\lambda}^{(g+1)})$  (Algorithm 2). The error  $\varepsilon$  is then mapped into a number of consecutive generations  $g_m, g_m \in [0, g_m^{\text{max}}]$ , for which the surrogate  $f_{\mathcal{M}}$  will be used (Algorithm 3).

We consider three different model error measures. As the CMA-ES depends primarily on the ranking of candidate solutions, the first two proposed error measures, *Kendall correlation coefficient* and *Rank difference* are based on ranking. The third one, *Kullback-Leibler divergence* a. k. a. *information gain*, measures a difference between a multivariate normal distribution estimated from the fitness values  $\mathbf{y}$  and a multivariate normal distribution estimated for the predicted values  $\hat{\mathbf{y}}$ .

*Kendall rank correlation coefficient* Kendall rank correlation coefficient  $\tau$  measures similarity between two different orderings of the same set. Let  $\mathbf{y} = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_{\lambda}))$  and  $\hat{\mathbf{y}} = (f_{\mathcal{M}}^{\text{last}}(\mathbf{x}_1), \dots, f_{\mathcal{M}}^{\text{last}}(\mathbf{x}_{\lambda}))$  be the sequences of the fitness values and the predicted values of a population  $\mathbf{x}_1, \dots, \mathbf{x}_{\lambda}$ , respectively. A pair of indices  $(i, j)$ , such that  $i \neq j, i, j \in \{1, \dots, \lambda\}$ , is said to be concordant, if both  $y_i < y_j$  and  $\hat{y}_i < \hat{y}_j$  or if both  $y_i > y_j$  and  $\hat{y}_i > \hat{y}_j$ . A discordant pair  $(i, j), i \neq j, i, j \in \{1, \dots, \lambda\}$  is one fulfilling that both  $y_i < y_j$  and  $\hat{y}_i > \hat{y}_j$  or both  $y_i > y_j$  and  $\hat{y}_i < \hat{y}_j$ . Let  $n_c$  and  $d_c$  denote the number of concordant and discordant

**Algorithm 2** Model error estimation

---

**Input:** error\_type (one of {"Kendall", "Rank-Difference", "Kullback-Leibler"})  
 $g$  (CMA-ES generation number)  
 $\mathbf{x}_1^{(g+1)}, \dots, \mathbf{x}_\lambda^{(g+1)}$  (a newly sampled population)  
 $\mathbf{y}, \hat{\mathbf{y}}$  (fitness values and model predictions in generation  $g$ )  
 $\mathbf{c}_{\text{cma}} = (c_c, c_1, c_\mu, c_\sigma, d_\sigma)$  (CMA-ES constants)  
 $\mathbf{v}_{\text{cma}}^{(g)} = (\mathbf{m}^{(g)}, \mathbf{C}^{(g)}, \mathbf{p}_c^{(g)}, \mathbf{p}_\sigma^{(g)}, \sigma^{(g)})$  (CMA-ES variables at generation  $g$ )  
 $\varepsilon_{\text{max}}$  (maximal error so far)

- 1: **if** error\_type = "Kendall" **then**
- 2:    $\tau \leftarrow$  Kendall rank correlation coefficient between  $\mathbf{y}$  and  $\hat{\mathbf{y}}$
- 3:    $\varepsilon \leftarrow \frac{1}{2}(1 - \tau)$
- 4: **else if** error\_type = "Rank-Difference" **then**
- 5:    $\varepsilon \leftarrow \varepsilon_{\text{RD}}^\mu(\hat{\mathbf{y}}, \mathbf{y})$
- 6: **else if** error\_type = "Kullback-Leibler" **then**
- 7:    $(\mathbf{m}^{(g+1)}, \mathbf{C}^{(g+1)}, \sigma^{(g+1)}) \leftarrow$   
    cma\_update( $(\mathbf{x}_1^{(g+1)}, \dots, \mathbf{x}_\lambda^{(g+1)}), \mathbf{y}, \mathbf{c}_{\text{cma}}, \mathbf{v}_{\text{cma}}^{(g)}$ )
- 8:    $(\mathbf{m}_{\mathcal{M}}^{(g+1)}, \mathbf{C}_{\mathcal{M}}^{(g+1)}, \sigma_{\mathcal{M}}^{(g+1)}) \leftarrow$   
    cma\_update( $(\mathbf{x}_1^{(g+1)}, \dots, \mathbf{x}_\lambda^{(g+1)}), \hat{\mathbf{y}}, \mathbf{c}_{\text{cma}}, \mathbf{v}_{\text{cma}}^{(g)}$ )
- 9:    $\varepsilon \leftarrow D_{\text{KL}}(\mathcal{N}(\mathbf{m}_{\mathcal{M}}^{(g+1)}, \sigma_{\mathcal{M}}^{(g+1)}, \mathbf{C}_{\mathcal{M}}^{(g+1)}) \| \mathcal{N}(\mathbf{m}^{(g+1)}, \sigma^{(g+1)}, \mathbf{C}^{(g+1)}))$
- 10:   **if**  $\varepsilon > \varepsilon_{\text{max}}$  **then**
- 11:      $\varepsilon_{\text{max}} \leftarrow \varepsilon$
- 12:   **end if**
- 13:    $\varepsilon \leftarrow \frac{\varepsilon}{\varepsilon_{\text{max}}}$  {normalize in proportion to the historical maximum}
- 14: **end if**

**Output:**  $\varepsilon \in [0, 1]$

---

pairs of indices from  $\{1, \dots, \lambda\}$ , respectively. The Kendall correlation coefficient  $\tau$  between vectors  $\mathbf{y}$  and  $\hat{\mathbf{y}}$  is defined as:

$$\tau = \frac{2}{\lambda(\lambda - 1)}(n_c - n_d).$$

In the corresponding branch of Algorithm 2, the value  $\tau$  is decreasingly scaled into interval  $[0, 1]$ .

*Ranking difference error* The ranking difference error is a normalized version of a measure used in [10]. Given  $r_1(i)$  the rank of the  $i$ -th element of  $\hat{\mathbf{y}}$  and  $r_2(i)$  the rank of the  $i$ -th element of  $\mathbf{y}$ , the ranking difference error is the sum of element-wise differences between  $r_1$  and  $r_2$  taking into account only the  $\mu$  best-ranked points from  $\hat{\mathbf{y}}$ :

$$\varepsilon_{\text{RD}}^\mu(\hat{\mathbf{y}}, \mathbf{y}) = \frac{\sum_{i: r_1(i) \leq \mu} |r_2(i) - r_1(i)|}{\max_{\pi \in S_\lambda} \sum_{i: \pi(i) \leq \mu} |i - \pi(i)|},$$

where  $S_\lambda$  is the group of all permutations of set  $\{1, \dots, \lambda\}$ .

*Kullback-Leibler divergence* Kullback-Leibler divergence from a continuous random variable  $Q$  with probability

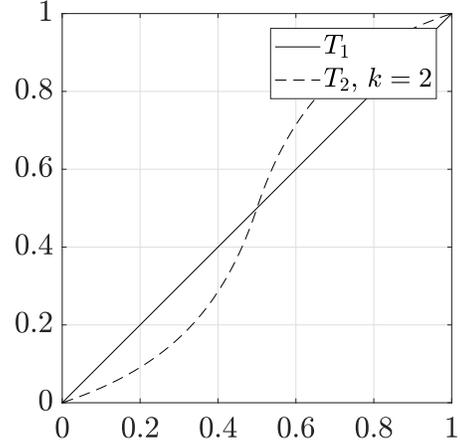


Figure 1: Model error transfer functions

density function  $q$  to a continuous random variable  $P$  with probability density function  $p$  is defined as:

$$D_{\text{KL}}(P \| Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx.$$

For two multivariate normal distributions  $\mathcal{N}_1(\mu_1, \Sigma_1)$  and  $\mathcal{N}_2(\mu_2, \Sigma_2)$  with the same dimension  $k$ , the Kullback-Leibler divergence from  $\mathcal{N}_2$  to  $\mathcal{N}_1$  is:

$$D_{\text{KL}}(\mathcal{N}_1 \| \mathcal{N}_2) = \frac{1}{2} \left( \text{tr}(\Sigma_2^{-1} \Sigma_1) + \ln \left( \frac{|\Sigma_2|}{|\Sigma_1|} \right) + (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) - k \right).$$

The algorithm of model error estimation (Algorithm 2) in generation  $g$  computes Kullback-Leibler divergence from a CMA-estimated multivariate normal distribution  $\mathcal{N}(\mathbf{m}^{(g+1)}, \mathbf{C}^{(g+1)})$  w.r.t. fitness values  $\mathbf{y}$  to a CMA-estimated multivariate normal distribution  $\mathcal{N}(\mathbf{m}_{\mathcal{M}}^{(g+1)}, \mathbf{C}_{\mathcal{M}}^{(g+1)})$  w.r.t. predicted values  $\hat{\mathbf{y}}$ . Procedure cma\_update in steps 7 and 8 refers to one iteration of the CMA-ES from the point when a new population has been sampled. The result is normalized by the historical maximum (step 13).

*Adjusting the number of model generations* The number of consecutive model generations  $g_m$  is updated by Algorithm 3. The history of surrogate model errors  $\varepsilon$  is exponentially smoothed with a rate  $r_u$  (step 1). The error is truncated at a threshold  $\varepsilon_T$  so that resulting  $g_m = g_m^{\text{max}}$  for all values  $\varepsilon \geq \varepsilon_T$  (step 3). We consider two different transfer functions  $T_1, T_2: [0, 1] \rightarrow [0, 1]$  (plotted in Figure 1) that scale the error into the admissible interval  $[0, g_m^{\text{max}}]$ :

$$T_1(x) = x \tag{3}$$

$$T_2(x; k) = \frac{(x - \frac{1}{2})(1 + \frac{1}{k})}{|2(x - \frac{1}{2})| + \frac{1}{k}} + \frac{1}{2}, k > 0. \tag{4}$$

**Algorithm 3** Updating the number of model generations

**Input:**  $\varepsilon$  (estimation of surrogate model error,  $\varepsilon \in [0, 1]$ )  
 $\varepsilon_T \in [0, 1]$  (a threshold at which the error is truncated to 1)  
 $\gamma: [0, 1] \rightarrow [0, 1]$  (transfer function)  
 $r_u$  (error update rate)  
 $\varepsilon_{\text{last}}$  (model error from the previous iteration)  
 $g_m^{\text{max}}$  (upper bound for admissible number of model generations)

- 1:  $\varepsilon \leftarrow (1 - r_u)\varepsilon_{\text{last}} + r_u\varepsilon$  {exponential smoothing}
- 2:  $\varepsilon_{\text{last}} \leftarrow \varepsilon$
- 3:  $\varepsilon \leftarrow \frac{1}{\varepsilon_T} \min\{\varepsilon, \varepsilon_T\}$  {truncation to 1}
- 4:  $g_m \leftarrow \text{round}(\gamma(1 - \varepsilon)g_m^{\text{max}})$  {scaling into the admissible interval}

**Output:**  $g_m$  – updated number of model-evaluated generations

Both functions are defined on  $[0, 1]$ , moreover,  $T_i(0) = 0$  and  $T_i(1) = 1$  for  $i = 1, 2$ . Transfer function  $T_2$  is a simple sigmoid function defined to be slightly less sensitive near the edges than in the middle. More control can thus be achieved in the regions of low and high error values. The parameter  $k$  determines the steepness of the sigmoid curve.

## 4 Experimental Setup

The proposed adaptive generation-based evolution control for the S-CMA-ES with three different surrogate model error measures is evaluated on the noiseless testbed of the COCO/BBOB (Comparing Continuous Optimizers / Black-Box Optimization Benchmarking) framework [7, 8] and compared with the S-CMA-ES and CMA-ES.

Each function is defined everywhere on  $\mathbb{R}^D$  and has its optimum in  $[-5, 5]^D$  for all dimensionalities  $D \geq 2$ . For every function and every dimensionality, 15 trials of the optimizer are run on independent instances, which differ in linear transformations of the  $x$ -space or shifts of the  $f$ -space. In our experiments, instances recommended for BBOB 2015 workshop, i. e.,  $\{1, \dots, 5, 41, \dots, 50\}$ , were used. Each trial is terminated when the  $f_{\text{opt}}$  is reached within a small tolerance  $\Delta f_i = 10^{-8}$  or when a given budget of function evaluations, 250D in our case, is used up. Experiments were run for dimensionalities 2, 3, 5, 10 and 20. The algorithms' settings are summarized in the following subsections.

### 4.1 CMA-ES

The CMA-ES results in BBOB format were downloaded from the BBOB 2010 workshop archive<sup>1</sup>. The CMA-ES used in those experiments was in version 3.40.beta and utilized a restart strategy (known as IPOP-CMA-ES), where the population size is increased by factor IncPopSize after

<sup>1</sup><http://coco.gforge.inria.fr/data-archive/bbob/2010/>

Table 1: Discretization of the A-S-CMA-ES parameters.

| Parameter       | Discretization       |
|-----------------|----------------------|
| $\gamma$        | $T_1$ (3), $T_2$ (4) |
| $\varepsilon_T$ | 0.5, 0.9             |
| $g_m$           | 5, 10, 20            |
| $r_u$           | 0.2, 0.5, 0.8        |

each restart [1]. The default parameter values employed in the CMA-ES are  $\lambda = 4 + \lceil 3\log D \rceil$ ,  $\mu = \lfloor \frac{\lambda}{2} \rfloor$ ,  $\sigma_{\text{start}} = \frac{8}{3}$ , IncPopSize = 2.

### 4.2 S-CMA-ES

The S-CMA-ES was tested with two numbers of model-evaluated generations,  $g_m = 1$  (further denoted as ‘‘GP-1’’) and  $g_m = 5$  (‘‘GP-5’’). All other S-CMA-ES settings were left as described in [2]. In particular, the Mahalanobis distance was  $r = 8$ , the starting values  $(\theta, l)$  of the Matérn covariance function  $k_{\text{Matérn}}^{v=5/2}$  were  $(0.5, 2)$  and the starting value of the GP noise parameter  $\sigma_n^2$  was 0.01. If not mentioned otherwise, the corresponding settings of adaptive versions of the S-CMA-ES are as just stated.

In order to find the most promising settings for each considered surrogate error measure, a full factorial experiment was conducted on one half of the noiseless testbed, namely on functions  $f_i$  for  $i \in \{2, 3, 6, 8, 12, 13, 15, 17, 18, 21, 23, 24\}$ . The discretization of continuous parameters  $(\gamma, \varepsilon_T, g_m^{\text{max}}, r_u)$  is reported in Table 1. All possible combinations of the parameters were ranked on the 12 selected functions according to the lowest achieved  $\Delta f^{\text{med}}$  (see Section 6) for different numbers of function evaluations #FEs/ $D = 25, 50, 125, 250$ . The best settings were chosen according to the highest sum of 1-st rank counts. Ties were resolved according to the lowest sum of ranks. All of the best settings included maximum model-evaluated generations  $g_m^{\text{max}} = 5$ . The remaining of the winning values are summarized in the following paragraphs.

*Kendall correlation coefficient (ADA-Kendall)* Transfer function  $\gamma = T_2$ , error threshold  $\varepsilon_T = 0.5$  and update rate  $r_u = 0.2$ .

*Ranking difference error (ADA-RD)* The same, except transfer function was  $\gamma = T_1$ .

*Kullback-Leibler divergence (ADA-KL)* Transfer function  $\gamma = T_2$ , error threshold  $\varepsilon_T = 0.9$  and update rate  $r_u = 0.5$ .

## 5 CPU Timing

In order to assess computational costs other than the number of function evaluations, we calculate CPU timing per

Table 2: The time in seconds per function evaluation for the Adaptive S-CMA-ES.

| Algorithm   | 2D   | 3D   | 5D   | 10D  | 20D  |
|-------------|------|------|------|------|------|
| ADA-KL      | 0.38 | 0.26 | 0.34 | 0.69 | 3.36 |
| ADA-Kendall | 0.47 | 0.45 | 0.61 | 1.29 | 6.27 |
| ADA-RD      | 0.57 | 0.60 | 0.71 | 1.63 | 7.90 |

function evaluation for each algorithm and each dimensionality. Each experiment was divided into jobs by dimensionalities, functions and instances. All jobs were run in a single thread on the Czech national grid MetaCentrum. The average time per function evaluation for each algorithm and each tested dimensionality is summarized in Table 2.

## 6 Results

We test the difference in algorithms’ convergence for significance on the whole noiseless testbed with the non-parametric Friedman test [3]. The algorithms are ranked on each BBOB function with respect to medians of log-scaled minimal distance  $\Delta f$  from the function optimum, denoted as  $\Delta f^{\text{med}}$ , at a fixed budget of function evaluations.

To account for different optimization scenarios, the test is conducted separately for all considered dimensionalities of the input space and two function evaluation budgets, a higher and a lower one. Let  $\#FE_t$  be the smallest number of function evaluations at which at least one algorithm reached the target, i. e., satisfied  $\Delta f^{\text{med}} \leq \Delta f_t$ , or  $\#FE_t = 250D$  if the target has not been reached. We set the higher budget for the tests to  $\#FE_t$  and the lower budget to  $\frac{\#FE_t}{3}$ .

Mean ranks from the Friedman test are given in Table 3. The critical value for the Friedman test is 2.29.

The mean ranks differ significantly for all tested scenarios except for both tested numbers of function evaluations in 2D and the higher tested number of function evaluations in 3D. Starting from 5D upwards, the lowest mean rank is achieved either by ADA-Kendall or ADA-RD at both tested  $\#FEs$ .

In order to show pairwise differences, we perform a pairwise  $N \times N$  comparison of the algorithms’ average ranks by the post-hoc Friedman test with the Bergmann-Hommel correction of the family-wise error [4] in cases when the null hypothesis of equal algorithms’ performance was rejected. To better illustrate algorithms differences, we also count the number of benchmark functions at which one algorithm achieved a higher rank than the other. The pairwise score and the statistical significance of the pairwise mean rank differences are reported in Table 4. In the post-hoc test, ADA-Kendall significantly outperforms both the CMA-ES and GP-5 in 10D and 20D. It also significantly outperforms GP-1 in 10D at the higher tested  $\#FEs$ .

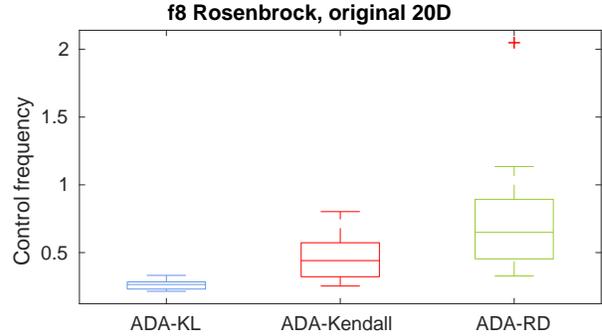


Figure 2: Average control frequency (the ratio of the number of total original-fitness-evaluated generations to the number of total model-evaluated generations) in A-S-CMA-ES measured in 15 trials of each algorithm on  $f_8$  in 20D.

For illustration, the average control frequency given by the ratio of the number of total original-fitness-evaluated generations to the number of total model-evaluated generations within one trial, for data from 15 trials on  $f_8$  (Rosenbrock’s function) in 20D is given in Figure 2. The algorithm ADA-KL led to generally lower control frequencies than its competitors, which might explain its slightly inferior performance. Similar results were observed for the remaining functions and dimensionalities.

The cases when ADA-Kendall and ADA-RD are able to switch between more exploitation-oriented and more data-gathering-oriented behaviour can be studied on the results from COCO’s postprocessing. GP-5 outperforms both GP-1 and the CMA-ES on the lower and middle parts of the empirical distribution functions (ECDFs) basically for all dimensionalities (Figure 3). On the other hand, GP-1 outperforms GP-5 especially in later phases of the search (Figure 3).

The ability of ADA-Kendall and ADA-RD to switch to a less-exploitation mode when appropriate is eminent on the ECDFs plots in 20D, especially on the moderate and the all-function groups (top right and bottom right on Figure 3), with exception of the well structured multimodal group (middle right), when they fail in the middle part and the weakly structured multimodal group (bottom left), when they fail towards the end of the search.

## 7 Conclusion

In this paper, we implemented several modifications of the Surrogate CMA-ES (S-CMA-ES), an algorithm using generation-based evolution control in connection with GPs. We considered three measures of surrogate model error according to which an adequate number of upcoming model-evaluated generations could be estimated online. Three resulting algorithms were compared on the COCO/BBOB framework with the S-CMA-ES parametrized by two different numbers of consecutive model-evaluated

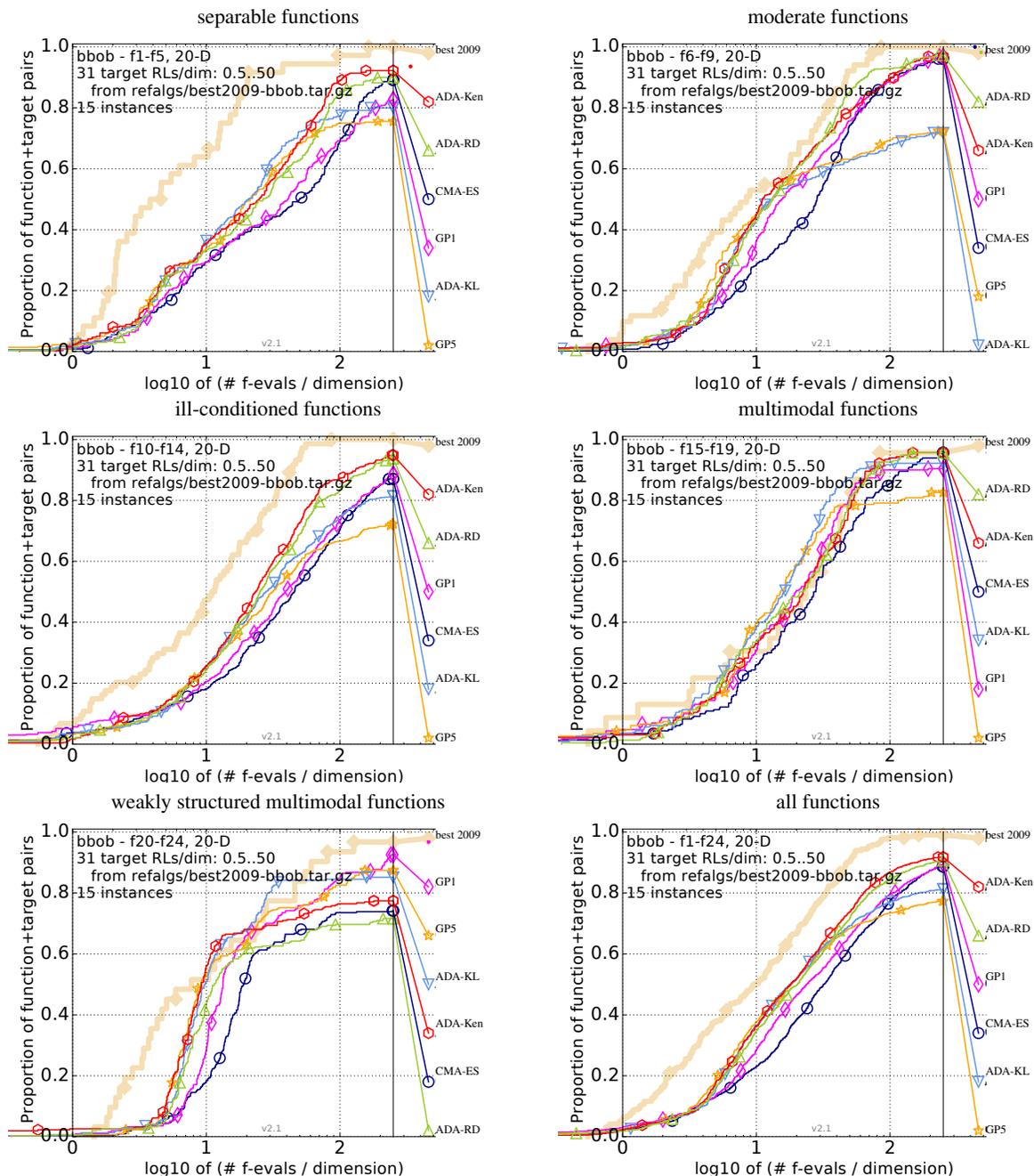


Figure 3: Bootstrapped empirical cumulative distribution of the number of objective function evaluations divided by dimension (FEvals/DIM) for all functions and subgroups in 20-D. The targets are chosen from  $10^{[-8..2]}$  such that the best algorithm from BBOB 2009 just not reached them within a given budget of  $k \times \text{DIM}$ , with 31 different values of  $k$  chosen equidistant in logscale within the interval  $\{0.5, \dots, 50\}$ . The “best 2009” line corresponds to the best aRT observed during BBOB 2009 for each selected target.

Table 3: Mean ranks of the CMA-ES, the S-CMA-ES and all A-S-CMA-ES versions over the BBOB and the Iman-Davenport variant of the Friedman test for the 10 considered combinations of dimensionalities and evaluation budgets. The lowest value is highlighted in bold. Statistically significant results at the significance level  $\alpha = 0.05$  are marked by an asterisk.

| Dim     | 2D            |             | 3D            |             | 5D            |             | 10D           |             | 20D           |             |
|---------|---------------|-------------|---------------|-------------|---------------|-------------|---------------|-------------|---------------|-------------|
|         | $\frac{1}{3}$ | 1           |
| CMA-ES  | 4.04          | 4.25        | 4.25          | 3.96        | 4.38          | 3.58        | 4.67          | 3.83        | 4.58          | 4.42        |
| GP-1    | 3.38          | 3.94        | 4.21          | 3.62        | 3.92          | 4.02        | 3.69          | 3.92        | 3.54          | 3.27        |
| GP-5    | 3.54          | 3.12        | <b>2.83</b>   | 4.08        | 3.81          | 4.35        | 4.25          | 4.42        | 4.23          | 4.52        |
| ADA-KL  | 3.23          | <b>2.85</b> | 3.29          | 3.44        | 3.69          | 3.73        | 3.60          | 4.04        | 3.15          | 3.65        |
| ADA-Ken | 3.98          | 3.46        | 3.25          | <b>2.90</b> | 2.90          | 2.96        | <b>2.27</b>   | <b>2.40</b> | <b>2.33</b>   | <b>2.23</b> |
| ADA-RD  | <b>2.83</b>   | 3.38        | 3.17          | 3.00        | <b>2.31</b>   | <b>2.35</b> | 2.52          | 2.40        | 3.17          | 2.92        |
| $F_F$   | 1.48          | 1.89        | 2.52*         | 1.67        | 4.47*         | 4.13*       | 7.82*         | 6.50*       | 5.35*         | 6.62*       |

generations. Since the work on the adaptive extension is still in progress, the presented results summarize the performance of all compared algorithms on the whole BBOB framework or its function groups. We found two error measures, the Kendall rank correlation and the rank difference error, that significantly outperformed the S-CMA-ES used with a higher number of model-evaluated generations, especially in higher dimensionalities of the input space. However, both of these algorithms provided only a minor improvement of the S-CMA-ES used with a lower number of model-evaluated generations and in some tested scenarios fell behind both tested settings of the S-CMA-ES. An area for further research is the adjustment of other surrogate model parameters beside the control frequency, such as the number of the training points or the radius of the area from which they are selected.

## 8 Acknowledgments

The research reported in this paper has been supported by the Czech Science Foundation (GAČR) grant 17-01251.

Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated.

## References

- [1] A. Auger and N. Hansen. A restart CMA evolution strategy with increasing population size. In *2005 IEEE Congress on Evolutionary Computation*. IEEE, 2005.
- [2] L. Bajer, Z. Pitra, and M. Holeňa. Benchmarking Gaussian processes and random forests surrogate models on the BBOB noiseless testbed. In *Proceedings of the Companion Publication of the 2015 on Genetic and Evolutionary Computation Conference - GECCO Companion '15*. Association for Computing Machinery (ACM), 2015.
- [3] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, December 2006.
- [4] S. García and F. Herrera. An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *J. Mach. Learn. Res.*, 9:2677–2694, 2008.
- [5] N. Hansen. The CMA evolution strategy: A tutorial. *CoRR*, abs/1604.00772, 2016.
- [6] N. Hansen, A. Auger, R. Ros, S. Finck, and P. Pošík. Comparing results of 31 algorithms from the Black-box Optimization Benchmarking BBOB-2009. In *Proceedings of the 12th annual conference comp on Genetic and evolutionary computation - GECCO '10*, pages 1689–1696, New York, NY, USA, 2010. ACM.
- [7] N. Hansen, S. Finck, R. Ros, and A. Auger. Real-parameter Black-Box Optimization Benchmarking 2009: Noiseless functions definitions. Technical report, INRIA, 2009, updated 2010.
- [8] N. Hansen, S. Finck, R. Ros, and A. Auger. Real-parameter Black-Box Optimization Benchmarking 2012: Experimental setup. Technical report, INRIA, 2012.
- [9] Y. Jin and B. Sendhoff. Fitness approximation in evolutionary computation—A survey. In *Proceedings of the Genetic and Evolutionary Computation Conference, GECCO '02*, pages 1105–1112, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.
- [10] S. Kern, N. Hansen, and P. Koumoutsakos. *Local Meta-models for Optimization Using Evolution Strategies*, pages 939–948. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.
- [11] Z. Pitra, L. Bajer, J. Repický, and M. Holeňa. Overview of surrogate-model versions of covariance matrix adaptation evolution strategy. In *Proceedings of the Genetic and Evolutionary Computation Conference 2017, Berlin, Germany, July 15–19, 2017 (GECCO '17)*. ACM, July 2017. To appear.
- [12] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning series. MIT Press, 2006.

Table 4: A pairwise comparison of the algorithms in 2D, 3D, 5D, 10D and 20D over the BBOB for 2 different evaluation budgets. The comparison is based on medians over runs on 15 instances for each of all the 24 functions. The number of wins of  $i$ -th algorithm against  $j$ -th algorithm over all benchmark functions is given in  $i$ -th row and  $j$ -th column. The asterisk marks the row algorithm achieving a significantly lower value of the objective function than the column algorithm according to the Friedman post-hoc test with the Bergmann-Hommel correction at family-wise significance level  $\alpha = 0.05$ .

| <b>2D</b>             | CMA-ES |     | GP-1 |     | GP-5 |     | ADA-KL |     | ADA-Ken |    | ADA-RD |    |
|-----------------------|--------|-----|------|-----|------|-----|--------|-----|---------|----|--------|----|
| #FEs/#FE <sub>t</sub> | 1/3    | 1   | 1/3  | 1   | 1/3  | 1   | 1/3    | 1   | 1/3     | 1  | 1/3    | 1  |
| CMA-ES                | —      | —   | 8    | 8   | 11   | 8   | 10     | 7   | 11      | 10 | 7      | 9  |
| GP-1                  | 16     | 16  | —    | —   | 12   | 9   | 13     | 8   | 13      | 9  | 9      | 6  |
| GP-5                  | 13     | 16  | 12   | 15  | —    | —   | 11     | 10  | 14      | 13 | 9      | 14 |
| ADA-KL                | 14     | 17  | 11   | 15  | 13   | 13  | —      | —   | 16      | 14 | 12     | 15 |
| ADA-Ken               | 13     | 14  | 11   | 15  | 10   | 10  | 8      | 9   | —       | —  | 7      | 11 |
| ADA-RD                | 17     | 15  | 15   | 17  | 15   | 9   | 12     | 9   | 17      | 12 | —      | —  |
| <b>3D</b>             | CMA-ES |     | GP-1 |     | GP-5 |     | ADA-KL |     | ADA-Ken |    | ADA-RD |    |
| #FEs/#FE <sub>t</sub> | 1/3    | 1   | 1/3  | 1   | 1/3  | 1   | 1/3    | 1   | 1/3     | 1  | 1/3    | 1  |
| CMA-ES                | —      | —   | 11   | 9   | 7    | 13  | 8      | 10  | 9       | 9  | 7      | 8  |
| GP-1                  | 13     | 15  | —    | —   | 7    | 14  | 7      | 11  | 6       | 8  | 10     | 9  |
| GP-5                  | 17     | 11  | 17   | 10  | —    | —   | 15     | 9   | 13      | 6  | 14     | 9  |
| ADA-KL                | 16     | 14  | 17   | 13  | 9    | 15  | —      | —   | 13      | 11 | 10     | 9  |
| ADA-Ken               | 15     | 15  | 18   | 16  | 11   | 17  | 11     | 13  | —       | —  | 11     | 12 |
| ADA-RD                | 17     | 16  | 14   | 15  | 10   | 14  | 14     | 15  | 13      | 10 | —      | —  |
| <b>5D</b>             | CMA-ES |     | GP-1 |     | GP-5 |     | ADA-KL |     | ADA-Ken |    | ADA-RD |    |
| #FEs/#FE <sub>t</sub> | 1/3    | 1   | 1/3  | 1   | 1/3  | 1   | 1/3    | 1   | 1/3     | 1  | 1/3    | 1  |
| CMA-ES                | —      | —   | 8    | 12  | 11   | 14  | 11     | 14  | 7       | 10 | 2      | 8  |
| GP-1                  | 16     | 12  | —    | —   | 11   | 12  | 11     | 12  | 9       | 7  | 3      | 4  |
| GP-5                  | 13     | 10  | 13   | 12  | —    | —   | 10     | 6   | 9       | 5  | 8      | 7  |
| ADA-KL                | 13     | 10  | 13   | 11  | 14   | 18  | —      | —   | 7       | 10 | 8      | 5  |
| ADA-Ken               | 17     | 14  | 15   | 17  | 15   | 19  | 17     | 14  | —       | —  | 10     | 9  |
| ADA-RD                | 22*    | 16  | 21*  | 20* | 16*  | 17* | 16     | 19  | 14      | 14 | —      | —  |
| <b>10D</b>            | CMA-ES |     | GP-1 |     | GP-5 |     | ADA-KL |     | ADA-Ken |    | ADA-RD |    |
| #FEs/#FE <sub>t</sub> | 1/3    | 1   | 1/3  | 1   | 1/3  | 1   | 1/3    | 1   | 1/3     | 1  | 1/3    | 1  |
| CMA-ES                | —      | —   | 7    | 12  | 13   | 14  | 10     | 14  | 1       | 8  | 1      | 4  |
| GP-1                  | 17     | 12  | —    | —   | 15   | 15  | 14     | 14  | 4       | 5  | 5      | 4  |
| GP-5                  | 11     | 10  | 9    | 9   | —    | —   | 8      | 11  | 6       | 4  | 8      | 5  |
| ADA-KL                | 14     | 10  | 10   | 10  | 16   | 13  | —      | —   | 8       | 5  | 9      | 8  |
| ADA-Ken               | 23*    | 16* | 20   | 19* | 18*  | 20* | 16     | 19* | —       | —  | 13     | 12 |
| ADA-RD                | 23*    | 20* | 19   | 20* | 16*  | 19* | 15     | 15* | 11      | 12 | —      | —  |
| <b>20D</b>            | CMA-ES |     | GP-1 |     | GP-5 |     | ADA-KL |     | ADA-Ken |    | ADA-RD |    |
| #FEs/#FE <sub>t</sub> | 1/3    | 1   | 1/3  | 1   | 1/3  | 1   | 1/3    | 1   | 1/3     | 1  | 1/3    | 1  |
| CMA-ES                | —      | —   | 7    | 5   | 11   | 12  | 9      | 13  | 4       | 3  | 3      | 5  |
| GP-1                  | 17     | 19  | —    | —   | 14   | 17  | 12     | 16  | 7       | 6  | 9      | 8  |
| GP-5                  | 13     | 12  | 10   | 7   | —    | —   | 4      | 5   | 6       | 5  | 10     | 6  |
| ADA-KL                | 15     | 11  | 12   | 8   | 20   | 19  | —      | —   | 9       | 8  | 13     | 10 |
| ADA-Ken               | 20*    | 21* | 17   | 18  | 18*  | 19* | 15     | 16  | —       | —  | 17     | 17 |
| ADA-RD                | 21     | 19* | 15   | 16  | 14   | 18* | 11     | 14  | 7       | 7  | —      | —  |