Design of a Surrogate Model Assisted $(\mu/\mu, \lambda)$-ES

JINGYUN YANG, Faculty of Computer Science, Dalhousie University

Surrogate models have been widely used to assist evolutionary algorithms (EAs) to avoid unnecessary objective function evaluations. But those surrogate assisted EAs are usually complicated and the behaviour of the algorithm is not well understood. A recent analysis of a surrogate model assisted (1+1)=ES has helped understand the behaviour of the algorithm and resulted in a step size adaptation mechanism. The goal of this thesis is to conduct a similar analysis for $(/mu/\mu, \lambda)$-ES that potentially more fully exploits the surrogate model in a sense a population of candidate solutions are evaluated by the surrogate in each iteration. It is unclear whether any additional performance advantage can be derived from this.

Additional Key Words and Phrases: $(\mu/\mu, \lambda)$-ES, Surrogate Model, Evolutionary algorithms(EAs), Gaussian Process

1 INTRODUCTION

Evolution strategies (ESs) have been widely utilized to solve optimization problems where the true objective function evaluation is computationally-intensive. Various attempts have been made to reduce the cost by extracting the information obtained from points evaluated in previous iterations, such information yields insights into better selection and recombination that help generate potential promising offspring. One way is to use a surrogate model, an approximation model trained based on the candidate solutions evaluated by the true objective function in previous iterations. The surrogate model acts as a substitution of the true objective function that gives an inaccurate estimate of the objective function value at much lower cost compared with using the exact objective function. The surrogate modeling can be helpful if the computational saving in using the true objective function outshines the potential poor step size resulted from the inaccurate surrogate estimation of the candidate solution.

Recent paper in surrogate assisted EAs by Kayhani and Arnold analyze surrogate assisted (1+1)-ES

surrogate model assisted

There are a range of surrogate models

One way is to use the cumulative step size adaptation (CSA) [19] that builds an evolution path based on the history step size (mutation) of ESs, the population in the next iteration is generated based on the mutation adapted by the evolution path.

The history information could be used to construct a surrogate model, referred either as a local approximation or a global approximation to the true objective function [16]. There are a range of surrogate models and a survey of the development can be found by Jin [14] and Loshchilov [18]. Those algorithms are usually heuristic by nature and the behaviour of each step is likely not well interpreted. Recent work in surrogate assisted EAs tend to use sophisticated algorithm where surrogates are combined or the model is updated online according to some heuristic. Comparison is often made by comparing the performance using the algorithm with and without model assistance where the behaviour of the surrogate is not well simulated. In this context, an approach that could simulate the surrogate would be helpful in understanding the surrogate behaviour, leading to potential modification for surrogate update or parameter-setting. A surrogate that models the objective function with desired precise gains benefit especially for algorithms that requires a large population size for good performance.The computational saving largely lies in the saved evaluations outshine the potential poor step resulted from relative inaccurate estimation of candidate solutions.

This thesis intend to analyze and understand the surrogate-assisted $(\mu/\mu, \lambda)$-ES on simple test functions following the analysis of surrogate model-assisted (1+1)-ES [17] and exploit the potential benefit of using an extensive sampling with surrogate model assistance. The paper is organized as follows: In Section 2 we give a brief review of related background, in Section 3 we propose a local surrogate model-assisted and study its behaviour on sphere functions. Based on the analysis, in Section 4, we first apply cumulative step size adaptation (CSA) to the algorithm. Using result obtained from CSA, we then propose a new step size adaptation mechanism for this algorithm the performance on several test functions are recorded. The experimental result is followed by a discussion and future work in Section 5.

2 BACKGROUND

2.1 Evolution Strategies

Evolution strategies (ESs), an category of Evolutionary Algorithms (EAs), is a nature-inspired direct search method that address optimization problems by using stochastic variation and selection. In each iteration, new offspring are generated from the parental population by mutation, followed by a selection based on the fitness of the offspring. Offspring selected as refereed to as the parental population for the next iteration.

ES are commonly used in black-box optimization where the N-dimensional search space $\mathbb{R}^N$ is continuous, whereas the solution space $\mathbb{R}$ is 1-dimensional. We consider minimization of an objective function $f : \mathbb{R}^N \to \mathbb{R}$ maps the search space to the solution space i.e. maps a point (individual) in the search space to a value (its fitness) in the solution space. Despite the objective function being continuous, there is no assumption on the objective function, such optimization problem are referred to as black box optimization.

2.1.1 $(\mu/\rho^{+}_{,}\lambda) - ES$.

A general algorithm for ES can be defined as follows. Assume a parental population size μ, number of parent for recombination (each offspring generated) ρ and the number of offspring generated in each iteration λ, where μ, ρ, λ are positive integers with $\rho \leq \mu$. $^{+}_{,}$, plus- or comma-selection refers to how the parental population is updated. If a plus-selection is applied, only the best μ individuals are chosen considering both the parental population and the offspring generated in this iteration (i.e. totally $\mu + \lambda$ individuals are considered for selection). Whereas a comma-selection only chooses the offspring to update the parental population, no individual from previous parental population can be chosen (i.e. only λ individuals are considered for selection).

2.1.2 Step size adaptation.

- The 1/5th Success Rule
- Cumulative Step-Size Adaptation
- Covariance Matrix Adaptation

2.1.3 Analyzing ES on sphere function.

2.2 Surrogate Model

1. Polynomials/Response Surface
2. Gaussian processes/Kriging
3.

3 RELATED WORK

3.1 Surrogate Model

3.1.1 Surrogate model uses EAs to optimize.

3.1.2 Surrogate model adapted online. Using an approximate model to reduce computational cost can be traced back to 1960s [7]. Some successful surrogated models include but are not limited to Polynomial Regression (PR, response surface methodology) [11], Gaussian Process (GP, Kriging models) [13], Artificial neural networks [23]. There are two types of surrogate models, global surrogate model and local surrogate model, . ES using global surrogate model based on Kring was examined by Ratle [20]. Another ES using global surrogate model based on Artificial neural networks was constructed by Jin [15] which gives an imperial criterion on using the true objective function or the surrogate model to evaluate the offspring. Ulmer et al [24] and Buche et al [5] also applied GP as surrogate models in ES. But the performance of global surrogate models degrade as the dimension of the data increases, known as *curse of dimensionality*. Since the performance of ES is straightly affected by the surrogate model accuracy, online surrogates has been introduced by using a surrogate-adaptation mechanism that updated the model according to some heuristic. Loshchilov et al [12] uses . Online local surrogate models [25] can be constructed using methods like radial basis function (RBF) [8] to replace the global surrogate model, where the surrogate model is updated online, giving a more accurate estimation compared with the global surrogate model.

There are various surrogate-assisted EAs integrating global and local surrogate models or using a combination of heuristics. These methods tend to be sophisticated for good performance, while few literatures have *systematically investigated*??? the surrogated-assisted $(\mu/\mu, \lambda)$-ES. One exception is what Chen and Zou [6] proposed but yet incomplete in terms of two aspects. Firstly, it uses a linear surrogate that cannot give a precise estimate when coordinate transform is applied, the precondition to solve a generalized optimization problem [17]. Secondly, it does not include a step size adaptation mechanism. Besides that, Ulmer et al [9] proposed a Model Assisted Steady-State Evolution Strategy (MASS-ES), which is a $(\mu + \lambda)$-ES that is a (1+1)-ES when we set $\mu = \lambda = 1$. But the behavior of step size adaptation is unclear given the proposed conditions.

wonder should focus more on $surrogate assisted(1 + 1) - ES or surrogate assisted mml - ES, possibliy most CMA - ES$

There is a wealth of literatures for solving black box optimization using (1+1)-ES on unimodal test problems given the convergence property of convex functions. Kayhani and Arnold [17] proposed a surrogated-assisted (1+1)-ES that investigates the acceleration and single step behaviour of the algorithm using GP based local surrogate. In this algorithm, the local surrogate acts as a filter and is updated every time when a true objective function is made. Since (1+1)-ES generate a single offspring per iteration and is not as robust as $(\mu/\mu, \lambda)$ especially in the presence of surrogate (bias due to choice of points), we argue that it is natural to ask to what degree the choice of population can benefit the ES in terms of robustness and acceleration.

3.2 Step size adaptation

The step size of $(\mu/\mu, \lambda)$-ES is commonly adapated using cumulative step size adaptation (CSA) proposed by Ostermeier et al [19]. In each iteration, $(\mu/\mu, \lambda)$-ES generate λ candidate solutions $y_i \in \mathbb{R}^N, i = 1, ..., \lambda$ from a parental population $x_i \in \mathbb{R}^N i = 1, ..., \mu$ and the centroid of the parent population is $x = 1/\mu \sum_{i=1}^{\mu} x_i$, where $\mu < \lambda$. The parental population is replaced by the best μ candidate solutions gennerated by $y_i = x + \sigma z$ where $\sigma \in \mathbb{R}$ is a scalar referred to as the step size and $z \in \mathbb{R}^N$ as the mutation. For a strategy with ideally adapted step size, each step should be uncorrelated. If the

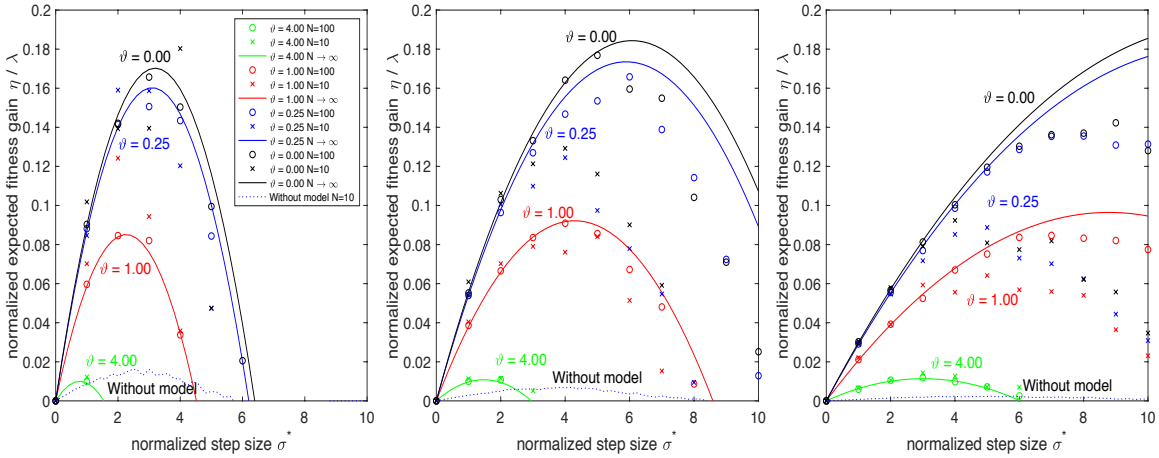

Fig. 1. The figures from left to right shows the expected single step behaviour of the surrogate model assisted $(\mu/\mu, \lambda)$-ES with unbiased Gaussian distributed surrogate error with $\lambda = 10, 20, 40$ respectively where $\mu = \lceil \lambda/4 \rceil$. The solid lines are the results obtained analytically when $n \to \infty$, while the dotted line below illustrates the corresponding performance ($n = 10$) of the $(\mu/\mu, \lambda)$-ES without model assistance. The dots represents the experimental result for $n = 10$ (crosses) and $n = 100$ (circles).

connective are negatively correlated, the step size should be decreased. In contrast, if the connective steps are positively correlated, meaning the steps are pointing to the same direction. Then a number of small steps can be replaced by fewer large steps and therefore, the step size should increase.

To decide the correlation, information from previous steps and mutations are cumulated. By comparing the step size with its expected length under random selection, the step size is adapted according to its expected length. Step size increases if the length is less than expected and decrease otherwise.

Define the search path as

$$p_{k+1} \leftarrow (1-c)p_k + \sqrt{\mu c(2-c)}z, \tag{1}$$

where $0 < c \geq 1$ is the proportion of history information retained and passed to the evolution path in the next iteration, $\sqrt{\mu c(2-c)}$ is a normalization constant that updates the evolution path from the mutation of this iteration and z, the mutation obtained by averaging the best μ candidate solutions generated.

The step size is adapated

$$\sigma \leftarrow \sigma \exp\left(\frac{c}{d}\left(\frac{\|p\|}{E\|N(O,I)\|}\right)\right), \tag{2}$$

where $E\|N(0,I)\|$ is the expected length of the search path p that can be approximated as $E\|N(0,I)\| \approx \sqrt{n}(1 - 1/4n + 1/21n^2)$. In Section 4, the details of parameter setting will be discussed.

wonder if something is missing feels not sufficient

4 ANALYSIS

To understand the potential implications of using surrogate models in EAs with varying population size, in this section, we use a simple model that applis a surrogate on the population. Specifically, we propose an EA that, in each iteration, a population of λ new candidate solutions are generated and then evaluated by the surrogate instead of true objective function calls and a selection based on the inaccurate surrogate estimate is done followed by a true objective function evaluation for the centroid of the selected referred to as the parent for next iteration. We assume that the inaccurate estimate of the surrogate model is a Gaussian random variable with mean equals the true objective function value of the candidate solution with some variance that describes the accuracy of the surrogate model. So, we can apply the technique of analyzing ESs's behaviours in the presence of Gaussian noise [3]. The analysis could be extend to biased surrogate models where the distribution mean is different from the exact objective function value[17]. not very sure what to follow

Consider the minimization of the quadratic sphere $f : \mathbb{R}^N \to \mathbb{R}$ with $f(x) = x^T x$ where the surrogate model assisted $(\mu/\mu.\lambda)$-ES is applied. This section will use the surrogate model described above to replace the true objective function calls of candidate solutions in each iteration, inaccurate but at vanishing cost. We first consider a simple iteration of the strategy. In each iteration, a population size of λ new candidate solutions $y_i \in \mathbb{R}^N, i = 1, ..., \lambda$ are generated from μ parents $x_i \in \mathbb{R}^N, i = 1, ..., \mu$, where $\lambda > \mu$. The parental population with size μ are replaced by the best μ candidate solutions $y_{i;\lambda}, i = 1, 2, ..., \mu$ evaluated by the surrogate model with fitness estimate $f_\epsilon(y_{i;\lambda}) \leq f_\epsilon(y_{j,\lambda}), 1 \leq i < j \leq \lambda$ at vanishing cost. For each of the λ candidate solutions $y_i = x + \sigma z$ïijN̈ where the parent $x = \sum_{i=1}^{n} x_i/\mu$ also the centroid of the parental population is obtained through intermediate recombination, $z \in \mathbb{R}^N$ is a standard normally distributed random vector, $\sigma > 0$ is the step size of the strategy, the adaptation is discussed in Section 4. The strategy uses the surrogate model to obtain a fitness estimate of the candidate solution $f_\epsilon(y_i), 1 \leq i \geq \lambda$ and by assumption the estimate has mean $f(y_i)$ with some standard deviation $\sigma_\epsilon > 0$ (*surroagte model error* also referred to as *fitness noise* [2]). Better surrogate model results in smaller model error σ_ϵ. For the λ new candidate solutions, $f_\epsilon(y_i) < f_\epsilon(y_j), 1 \leq i < j \leq \lambda$ indicates the estimated objective function value of y_i is superior to y_j and therefore the best μ candidate solutions are selected, replacing the old parental population of size μ (used for offspring generation in next iteration), at the same time, the other inferior candidate solutions are discarded. Therefore, only one objective function call is made per iteration in evaluating the fitness of the parent (centroid of parental population). The surrogate essentially does a pre-selection for $(\mu/\mu, \lambda)$-ES over candidate solutions, avoiding the necessary objective function calls determined by the surroagte model.

Decomposition of z, first proposed by Rechenberg [21] can be used to study the expected step size of the strategy. We can decompose the vector z as a vector sum $z = z_1 + z_2$, where z_1 is in the direction of the negative gradient of the objective function $\nabla f(x)$, while z_2 orthogonal to z_1. We have z_1 standard normally distributed, while $\|z_2\|^2$ χ-distributed with $N - 1$ degree of freedom and $\|z_2\|^2/N \overset{N\to\infty}{=} 0$ (see reference theorem [*dirk's slides*]). Denote $\delta = N(f(x) - f(y))/(2R^2)$, where $R = \|x\|$ is the distance to the optimal, we further introduce normalized step size $\sigma^* = N\sigma/R$ and $z_{\text{step}} = \sum_{i=1}^{\mu} z_{i;\lambda}$ (the averaged z taken by the best *mu* candidate solutions). The normalized fitness advantage of y over x follows

$$\begin{aligned}
\delta &= \frac{N}{2R^2}(x^T x - (x + \sigma z_{\text{step}})^T (x + \sigma z_{\text{step}})) \\
&= \frac{N}{2R^2}(-2\sigma x^T z_{\text{step}} - \sigma^2 \|z_{\text{step}}\|^2) \\
&\overset{N\to\infty}{=} \sigma^* z_{\text{step},1} - \frac{\sigma^{*2}}{2},
\end{aligned} \tag{3}$$

where $z_{\text{step},1}$, the component of z_{step} pointing to the negative graident of $f(x)$, is normally distributed and $\overset{N\to\infty}{=}$ denotes the convergence of the distribution $\|z_{\text{step}}\|^N/N = 1$. We further introduces $\sigma_\epsilon^* = N\sigma_\epsilon/(2R^2)$, the normalized surrogate model error (also referred to as the normalized fitness noise in Noise Sphere from Arnold and Beyer [2]). The estimate of true objective function value of y_i is $f_\epsilon(y_i) = f(y_i) + \sigma_\epsilon z_\epsilon, z_\epsilon \in \mathbb{R}$ is standard normally distributed.

Replacing $f(y)$ with $f_\epsilon(y)$, the actual normalized fitness advantage of y using the surrogate model is

$$\delta_\epsilon = \delta + \sigma_\epsilon^* z_\epsilon \tag{4}$$

The expected value of the normalized change in objective function value

$$\begin{aligned}\Delta &= -\frac{N}{2} E\left[\log f(y) - \log f(x)\right] \\ &= -\frac{N}{2} E\left[\log \frac{f(x^{t+1})}{f(x^t)}\right],\end{aligned} \tag{5}$$

where y^t is the centroid of parental population in timestamp t, the equation is normalized in terms of dimensionality.

Since the fitness of λ offspring generated are evaluated by the surrogate model with vanishing cost. The objective function evaluation per iteration is 1 instead of λ (for $(\mu/\mu, \lambda)$-ES), therefore the normalized progress rate when dimensionality $N \to \infty$, by substituting λ with 1 in equation (7) from [4] is

$$\eta = \frac{1}{1} E[\Delta] = \frac{\sigma^* c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2}} - \frac{(\sigma^*)^2}{2\mu}, \tag{6}$$

where $\vartheta = \sigma_\epsilon^*/\sigma^*$ is the noise-to-signal ratio, defined to measure the quality of surrogate model relative to the algorithm's step size, $c_{\mu/\mu,\lambda}$ is the $(\mu/\mu, \lambda)$-progess coefficient derived by Arnold and Beyer [1] that follows

$$c_{\mu/\mu,\lambda} = \frac{\lambda-\mu}{2\pi}\binom{\lambda}{\mu}\int_{-\infty}^{\infty} e^{-x^2}\left[\Phi(x)\right]^{\lambda-\mu-1}\left[1-\Phi(x)\right]^{\mu-1}\mathrm{d}x, \tag{7}$$

where Φ^{-1} is the inverse function of Φ, the normal cumulative distribution function. The integral can be solve numerically.

To obtain the opt. expected fitness gain η_{opt} and its corresponding opt. normalized step size σ^*_{opt}, we take derivative of equation (6) over σ^* and get the following

$$\sigma^*_{opt} = \frac{\mu c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2}} \tag{8}$$

$$\eta_{opt} = \frac{\sigma^*_{opt} c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2}} - \frac{(\sigma^*_{opt})^2}{2\mu} \tag{9}$$

The expected fitness gain is normalized in terms of the population size λ for easy comparison. The normalized fitness gain against the normalized step size for $(\mu/\mu, \lambda)$-ES with population size $\lambda = 10, 20, 40$ corresponding $\mu = 3, 5, 10$ are plotted in 1 from left to right respectively. The line shows the result obtained from Eqs. (6) (7). The dots represent the experimental result for unbiased Gaussian surrogate error for $n \in \{10, 100\}$ obtained by averaging 100 runs. The result obtained for $n \to \infty$ are considered to be cases with a large normalized step size with very small noise to signal ratio.

It can be inferred from Fig. 1, for a fixed population size, the expected fitness gain decreases with an increasing noise-to-signal-ratio. When $\vartheta \to \infty$, the surrogate model becomes useless and the strategy becomes a random search. For moderate noise-to-signal ratio ϑ, the surrogate model assisted algorithm can achieve much larger value for expected fitness gain at a larger normalized step size. When $\vartheta = 1$, the maximal expected fitness gain achievable for $(3/3, 10)$-ES,$(5/5, 20)$-ES and $(10/10, 40)$-ES are 0.8507, 1.841, 3.808 with $\sigma^* = 2.254, 4.251, 8.738$ respectively. Compared with the

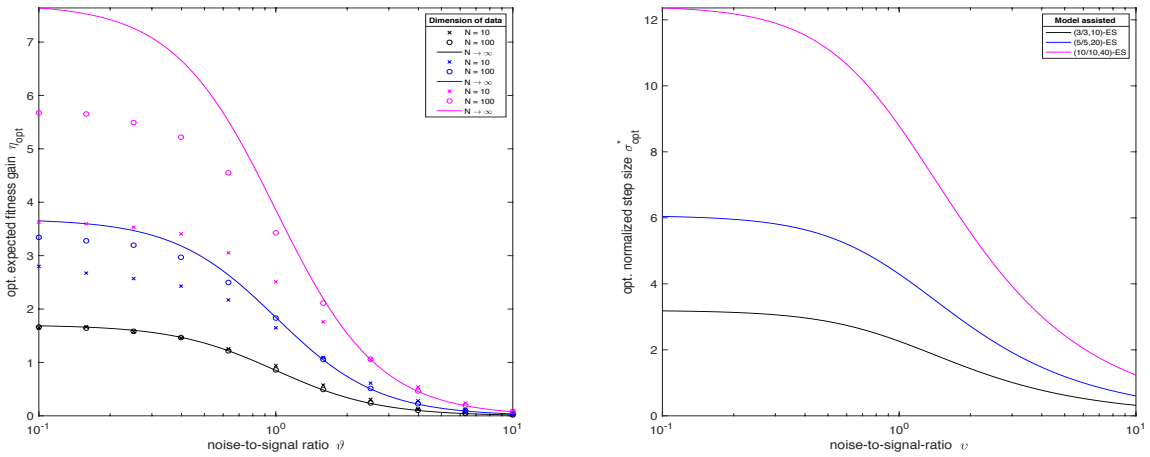

Fig. 2. Opt. expected fitness gain and corresponding opt. normalized step size of the surrogate model assisted $(\mu/\mu, \lambda)$-ES plotted against the noise-to-signal ratio. The line and dots with colour black, blue green represent (3/3, 10)-ES, (3/3, 10)-ES, (3/3, 10)-ES The solid line represents the results obtained analytically when $n \to \infty$.

result of the surrogate assisted (1+1)-ES [17] where maximal fitness gain is 0.548 achieved at $\sigma^* = 1.905$, $(\mu/\mu, \lambda)$-ES does benefit from using a larger population from the analysis. For $\vartheta = 0$ (the surrogate models the objective function exactly), from equation (8) we can obtain the maximal expected fitness gain is achieved at $\sigma^*_{opt} = \mu c_{\mu/\mu,\lambda}$ with value $\eta_{opt} = \mu(c_{\mu/\mu,\lambda})^2/2$. Even if this indicates the potential benefit the strategy may gain with a growing population, it is important to note the analytical results derived when $n \to \infty$ is an approximation for the finite-dimensional case. Fig. 2 shows the relation of optimal expected fitness gain and the corresponding optimal normalized step size over noise-to-signal ratio derived analytically in the limit of $n \to \infty$ for three different population sizes. The optimal expected fitness gain is also measured experimentally for $n \in \{10, 100\}$.

The speed-up is the ratio of median number of objective function evaluations used for surrogate assisted (1+1)-ES divide by that of surrogate assisted $(\mu/\mu, \lambda)$-ES. For a finite-dimension, the speed-up achieved with surrogate model assistance for small noise-to-signal ratio with $n = 10$ appears to be around one and two for a population size equals 10, two and three for $\lambda = 20$, four and five for $\lambda = 40$ respectively. The speed-up of $n = 100$ for each population size fall into almost the same range as is the case for $n = 10$. Wonder if the speed up should be calculated as (1+1)-ES/mml-ES both with model assistance and could it be reported in the table later as in each cell numOfObjFunCalls(speed-up)

There is a significant speed-up following the analysis and it seems the expected fitness gain of surrogate assisted $(\mu/\mu, \lambda)$-ES will increase as the population size λ grows.

5 STEP SIZE ADAPTATION

5.1 Cumulative step size adaptation

Even though the analysis in Section 3 suggests a potential better performance for the surrogate-assisted $(\mu/\mu, \lambda)$-ES. There is no guarantee the step size of the strategy can be properly adapted and further the analysis is very inaccurate in terms of finite dimension. In this section we experiment the surrogate model assisted $(\mu/\mu, \lambda)$-ES using the cumulative

Algorithm 1 A Surrogate Assisted $(\mu/\mu, \lambda)$-ES

$c \leftarrow \frac{\mu+2}{n+\mu+5}$
$d \leftarrow 1 + 2\max(0, \sqrt{\frac{\mu-1}{n+1}} - 1)$
$p \leftarrow 0$
while not terminate() **do**
  for $i = 1, 2, ..., \lambda$ **do**
    Generate standard normally distributed $z_i \in \mathbb{R}^N$
    $y_i \leftarrow x + \sigma z_i$
    Evaluate y_i using the surrogate model, yieding $f_\epsilon(y_i)$
  end for
  $z = \frac{1}{\mu} \sum_{i=1}^{\mu} z_{i;\lambda}$
  $y = x + \sigma x$
  Evaluate y using true objective function, yieding $f(y)$
  Update surrogate modle
  $s \leftarrow (1-c)s + \sqrt{c(2-c)\mu z}$
  $\sigma \leftarrow \sigma \times \exp\left(\frac{c}{d} \frac{\|X\|}{E\|N(0,I)\|} - 1\right)$
end while

Table 1. Median test results using CSA.

	Median number of objective function calls (speed-up)			
Test functions	$(1+1)$-ES	$(3/3, 10)$-ES	$(5/5, 20)$-ES	$(10/10, 40)$-ES
`linear sphere`	505	754(0.67)	689(0.73)	755(0.67)
`quadratic sphere`	214	310(0.69)	245(0.87)	228(0.93)
`cubic sphere`	202	274(0.74)	250(0.81)	254(0.80)
`Schwefel' s function`	1496	$+\infty$(/)	$+\infty$(/)	$+\infty$(/)
`quartic function`	1244	1006(1.2)	750(1.7)	662(1.9)

step size adaptation described in Section 2.2 and exploit the potential insight it may offer. The strategy is evaluated by using a Gaussian Process based surrogate model replacing the simple model that simulates the surrogate behaviour in Section 3. Several test functions are used for testing the strategy.

Here, we use the well established parameters from Hansen's CMA tutorial [10] that follows

$$\begin{cases} c = (\mu+2)/(N+\mu+5) \\ d = 1 + 2\max\left(0, \sqrt{(\mu-1)/(N+1)-1}\right) + c. \end{cases} \tag{10}$$

One signle iteration fo the surrogate model assisted $(\mu/\mu, \lambda)$-ES using CSA is shown in Alg. 1.

Five ten-dimensional test problems are used to test if the step size of the strategy has been appropriately adapted, namely sphere functions $f(x) = (x^T x)^{\alpha/2}$ for $\alpha = \{1, 2, 3\}$ referred to as linear, quadratic and cubic spheres, $f(x) = \sum_{i=1}^{n}(\sum_{j=1}^{i} x_j)^2$ (i.e. a convex quadratic function with condition number of the Hessian approximately equal to 175.1) referred to as Schwefel's Problem 1.2 [22]) and quartic function [17] defined as $f(x) = \sum_{i=1}^{n-1} \left[\beta(x_{i+1} - x_i^2)^2 + (1-x_i)^2\right]$ where $\beta = 1$. The quartic function becomes the Rosenbrock function when the condition number of the Hessian at the optimizer exceeds 3,500, making it very hard to find the global optima without adapting the shape of mutation distribution. We use the quartic function in the context with $\beta = 1$ and condition number of the Hessian at the optimizer

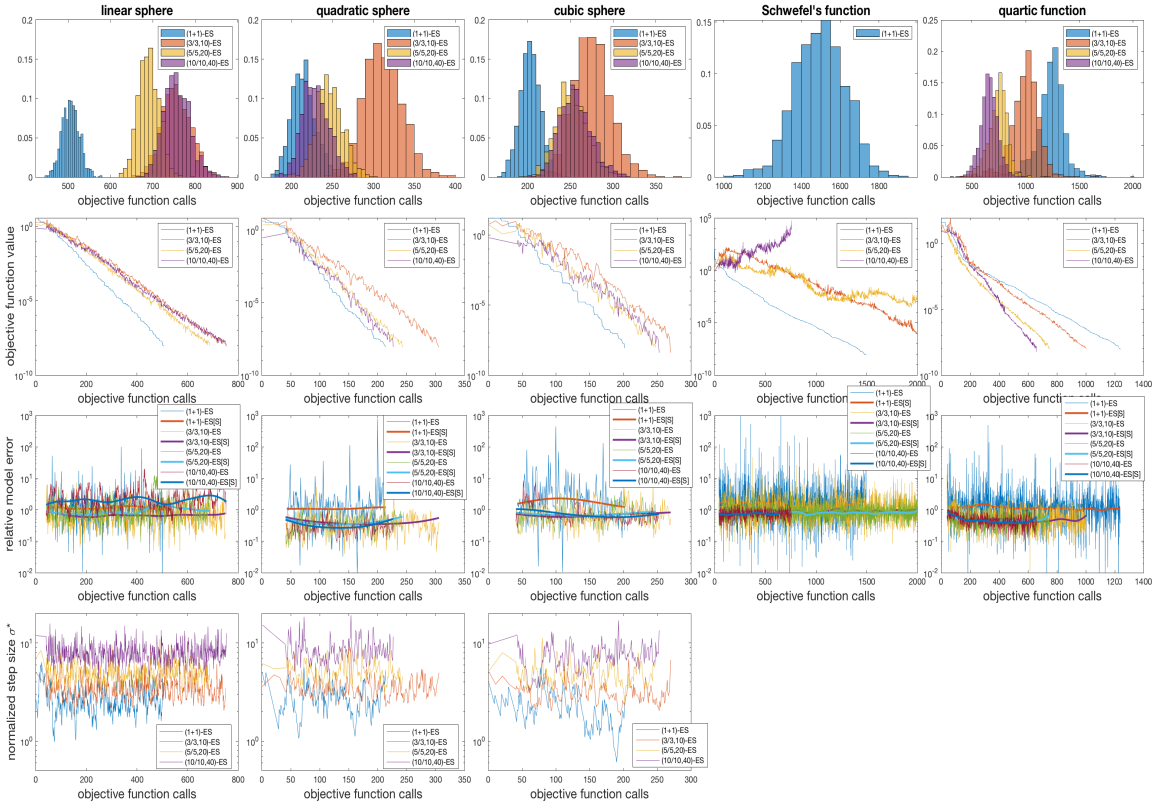

Fig. 3. Result obtained by adapting step size using CSA. Top row: Histogram showing the number of objective function calls needed to solve the five test problems. Second row: Convergence graphs for median runs. Third row: Relative model error obtained in median runs ([S] denotes the smoothed plot). Last row: normalized step size measured in median runs. modified the code to make it more compact, look fine?

equals to 49.0. The value of global optima for all test functiosn is zero. For each test problem, 1000 runs are conducted both for surrogate assisted (1 + 1)-ES and surrogate assisted $(\mu/\mu, \lambda)$-ES where a parental population size $\lambda = 10, 20, 40$ with $\mu = \lceil \lambda/4 \rceil$ are used. For surrogate model, we use Gaussian process with squared exponential kernel and the length scale parameter in the kernel is set proportional to the square of the dimension and the step size of the ES. For simplicity, the length scale is set to $8\sigma\sqrt{n}$.

The Gaussian process kernel is constructed using a training size of 40. The training set consists of the most recent 40 candidate solutions evaluated, so that the surrogate model approximates the local landscape of the objective function. All runs are initialized with starting point sampled from a Gaussian distribution with zero mean and unit covariance matrix and initial step size $\sigma_0 = 1$. The termination criteria is defined as one solution achieves objective function value below 10^{-8}.

Histogram showing the number of objective function calls needed to solve the test problems within the required accuracy are represented in the first row of Fig. 3, the median objective function calls for each test problem is shown in

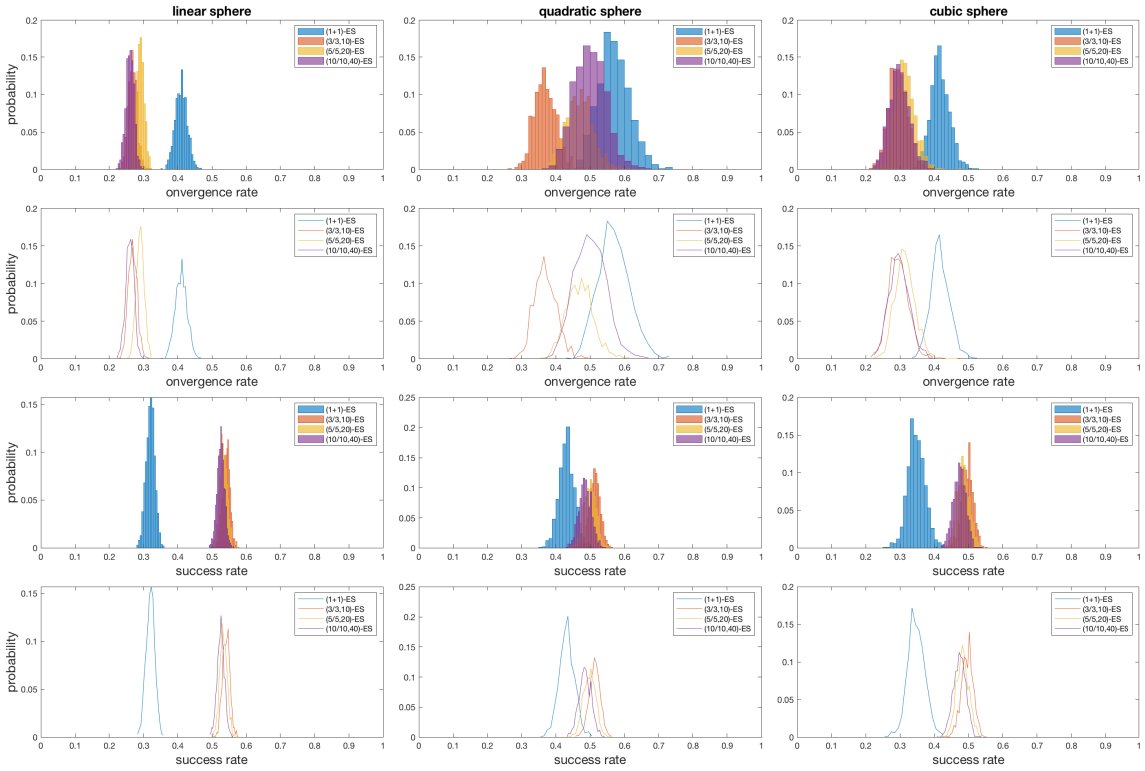

Fig. 4. Result obtained by adapting step size using CSA. The first two rows show the normalized convergence rate for each run plotted in histogram and normalized probability density function (pdf) respectively. The last two rows represent the success rate (proportion of good step size in each run) plotted in histogram and pdf respectively.

Table 1. The result of Arash and Dirk [17] using surrogate assisted (1+1)-ES is also included for comparison. The results of surrogate assisted $(\mu/\mu, \lambda)$-ES do not match the performance of surrogate assisted (1+1)-ES. The speed-up is defined as the median number of objective function evaluations used by the surrogate assisted (1+1)-ES [17] divided by the surrogate assisted $(\mu/\mu, \lambda)$-ES. Despite achieving a speed up between 1.2 and 1.9 for quartic function, the surrogate assisted $(\mu/\mu, \lambda)$-ES performs worse on sphere functions and even does not converge in Schwefel' s function. There is a trend in quadratic sphere and quartic function that the performance improves with a growing parental population, the number of objective function evaluations needed to solve linear sphere even increases after $\lambda > 20$.

The second row of Fig. 3 shows th convergence graphs observed in median runs. Linear convergence are achieved for all test functions despite the Schwefel's function using surrogate assisted $(\mu/\mu, \lambda)$-ES, interestingly, using a larger population does not help achieve a better convergence rate, but instead, makes the strategy diverge. Relative model error for the median runs is shown in the third row of the figure, defined as $\|f(y) - f_\epsilon(y)\|/\|f(y) - f(x)\|$ where x is the parent and y the offspring candidate solution for (1+1)-ES and $\mathrm{var}(f(y) - f_\epsilon(y))/\mathrm{var}(f(y))$, the variance of the difference between the surrogate estimate of λ candidate solutions and their true objective function values dvided by the variance of the true objective function values of λ candidate solutions for $(\mu/\mu, \lambda)$-ES. The relative model error is smoothed

logarithmically by convolution with a Gaussian kernel with window size 40 that is represented as the bold line in the centre of the plots (denoted [S] in the Fig.). This can be interpreted as the a relative constant noise-to-signal ratio. The relative model error for all surrogate assisted ES in this context is approximately 1 and according to the analysis in Section 3 should give a much larger speed up especially given a larger population. This may give indication that the step size is not appropriately adapted. The bottom row shows the normalized step size $\sigma^* = N\sigma/R$ for three sphere functions, where N,R are the dimension of data and distance to optimal respectively is the dimension. It coincides with the knowledge that using a population in offspring generation is possible for larger step size but the potential improvement is still yet clear.

There is a big gap between the analytical result obtained in Section 3 and the experimental result shown above. The relation between the expected fitness gain and population size is not yet clear. To better understand the relation, we plot the histogram and probability density function (pdf) for success rate (for a good step size) and normalized convergence rate for linear, quadratic and cubic sphere functions in Fig. 4. The normalized convergence rate are defined as follows

$$c = \begin{cases} -n \left[\log \left(\frac{f(x_{t+1})}{f(x_t)} \right) \right], & \text{linear sphere} \\ -\frac{n}{2} \left[\log \left(\frac{f(x_{t+1})}{f(x_t)} \right) \right], & \text{quadratic sphere} \\ -\frac{n}{3} \left[\log \left(\frac{f(x_{t+1})}{f(x_t)} \right) \right], & \text{cubic sphere.} \end{cases} \tag{11}$$

Histograms showing the normalized convergence rate for all runs are shown in the top row of Fig. 4, followed by its probability density function (pdf) in row 2. The last two rows in Fig. 4 shows the success rate plotted in histogram and pdf for all runs. Despite the relative large success rate for a good step size, the $(\mu/\mu, \lambda)$-ES with model assistance has a lower normalized convergence rate compared with (1+1)-ES with model assistance. In linear sphere, the normalized convergence rate between 0.2 for and 0.3 compared with 0.4 for (1+1)-ES partially explains the relative poor performance. Both success rate and normalized convergence rate for $(\mu/\mu, \lambda)$-ES do not vary much in terms of population size. It is interesting that the probability of a good step size for surrogate assisted $(\mu/\mu, \lambda)$-ES is approximately 0.5, indicating the strategy makes a bad step every other step.

5.2 Cumulative step size adaptation with emergency

Given a success rate approximately 0.48 for all population size in all sphere functions. It comes natural to ask, how much we are to benefit if we can avoid or simply reject those bad steps. Recent papers in surrogate model assisted ES consider (1+1)-ES [17], the step size of the strategy is successfully adapted based on the success rate of a good step size. The step size decreases if the estimated fitness of the offspring is inferior to the true fitness of its parent or the true fitness of the offspring evaluated is inferior to that of its parent. Applying a similar idea, we propose step size adaptation mechanism for the surrogate assisted $(\mu/\mu, \lambda)$-ES based on CSA that handles emergency. We define the emergency situation as an offspring generated is inferior to its parent, meaning the step size generated in this iteration is bad. Given the emergency, we decrease the step size by a factor of 0.72. The proposed step size adaptation using CSA with emergency is shown in Alg. 2 by adding a conditional statement comparing the fitness of the offspring obtained with its parent as is illustrated in line 15 Alg. 2. In each timestamp, one offspring (the centroid of the λ) is evaluated using the true objective function and its fitness is compared to its parent. If the fitness of the offspring is inferior to its parent, indicating the step size made is poor, the offspring is discarded and the step size is decreased. The bad step

Algorithm 2 Cumulative Step Size Adaptation with Emergency

$c \leftarrow \frac{\mu+2}{n+\mu+5}$
$d \leftarrow 1 + 2\max(0, \sqrt{\frac{\mu-1}{n+1}} - 1)$
$p \leftarrow 0$
$D \leftarrow 0.68$
while not terminate() **do**
  for $i = 1, 2, ..., \lambda$ **do**
    Generate standard normally distributed $z_i \in \mathbb{R}^N$
    $y_i \leftarrow x + \sigma z_i$
    Evaluate y_i using the surrogate model, yieding $\hat{f}(y_i)$
  end for
  $z = \frac{1}{\mu} \sum_{i=1}^{\mu} z_{i;\lambda}$
  $y = x + \sigma x$
  Evaluate y using true objective function, yieding $f(y)$
  Update surrogate modle
  if $f(x) < f(y)$ (Emergency) **then**
    $\sigma \leftarrow \sigma D$
  else
    $s \leftarrow (1-c)s + \sqrt{c(2-c)\mu z}$
    $\sigma \leftarrow \sigma \times \exp\left(\frac{c}{d} \frac{\|X\|}{E\|N(0,I)\|} - 1\right)$
  end if
end while

Table 2. Median test results (CSA with emergency).

	Median number of objective function calls (speed-up)			
Test functions	$(1+1)$-ES	$(3/3, 10)$-ES	$(5/5, 20)$-ES	$(10/10, 40)$-ES
`linear sphere`	505	364(1.4)	315(1.6)	322(1.6)
`quadratic sphere`	214	211(1.0)	162(1.3)	146(1.5)
`cubic sphere`	203	213(1.0)	177(1.1)	176(1.2)
`Schwefel' s function`	1496	2002(0.747)	1352(1.1)	1067(1.4)
`quartic function`	1244	1509(0.8)	987(1.3)	797(1.6)

size is not added to the evolution path since we want to build an evolution path based on the good step information of previous iterations.

To test the proposed the step size adaptation mechanism, we use same test functions and generate corresponding plots from Section 4.1. The number of objective function evaluations in median runs and the corresponding speed-up is represented in Table 2. The performance of surrogate assisted $(\mu/\mu, \lambda)$-ES improves as the population size increases, which is as expected. For a population size of 40, the speed up for sphere functions are 1.6, 1.5 and 1.2 for linear, quadratic and cubic respectively. It is notable that the speed-ups in linear sphere are between twice and three times before the emergency situation is proposed. For Schwefel' s function and quartic function, the strategy obtain a

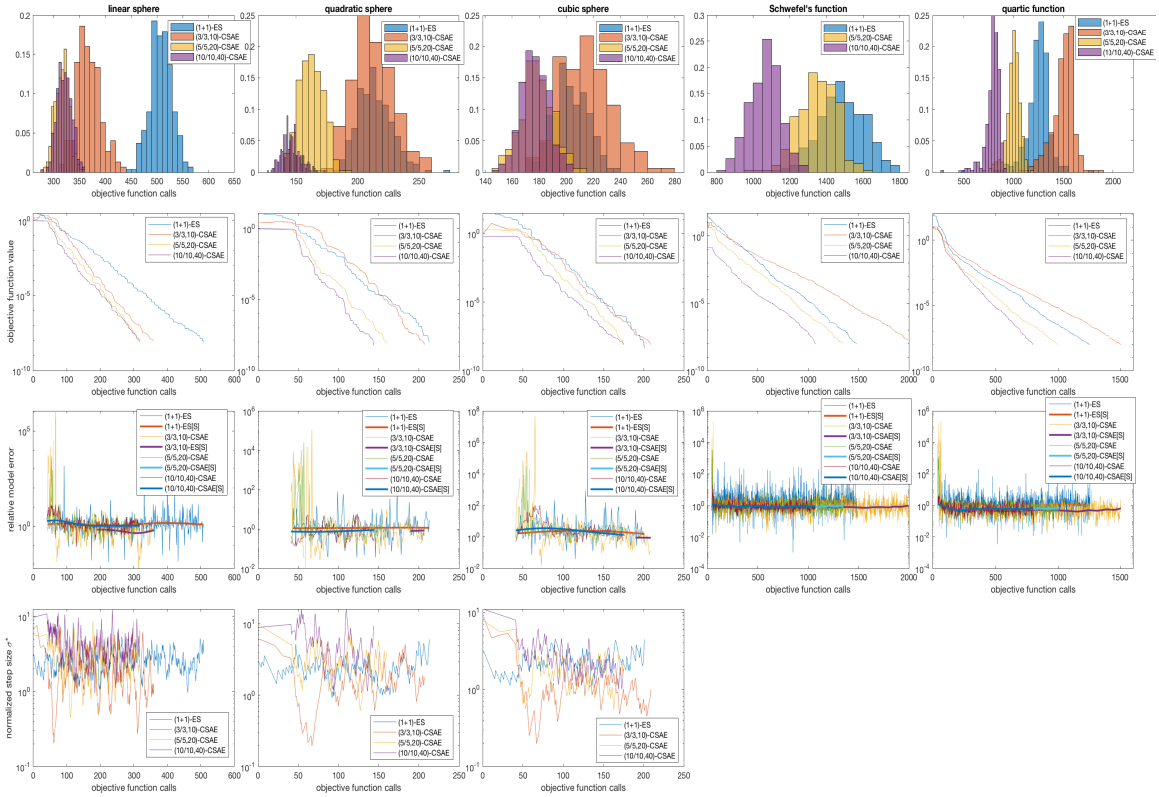

Fig. 5. Result obtained by adapting step size using CSA with emergency (denoted with CSAE). Top row: Histogram showing the number of objective function calls needed to solve the five test problems. Second row: Convergence graphs for median runs. Third row: Relative model error obtained in median runs ([S] denotes the smoothed plot). Last row: normalized step size measured in median runs.

convergence rate of 0.35 and 0.5 respectively for a population size from 10 to 20 with 0.3 for both test functions for a population size from 20 to 40. This is well illustrated in the histogram of objective function calls in first row of Fig. 5 that the objective function calls for all functions reduces with a growing population size. The convergence graphs in the second show that linear convergence is achieved for all strategies for all test functions. The third row shows the relative model error for the median runs described in Section 4.1, it is interesting that the relative model error for surrogate assisted $(\mu/\mu, \lambda)$-ES with different population size is actually higher after the step size is adapted using the CSA with emergency, previous result in Section 3.1 shows (1+1)-ES with model assistance has higher relative model error, but the value is really close after using the new step size adaptation. The last row in Fig. 5 shows the normalized step size, where the benefit of $(\mu/\mu, \lambda)$-ES is no longer obvious given the fact that we discard the inferior offspring, but it can be inferred that using a larger population size could reduce the variance in normalized step size. will add more accurate data for comparison, including the range of GP error

Histogram and probablity density function of normalized convergence rate and success rate are plotted in Fig. 6. The convergence rate for all population size grows significantly, almost doubled for all test functions despite a

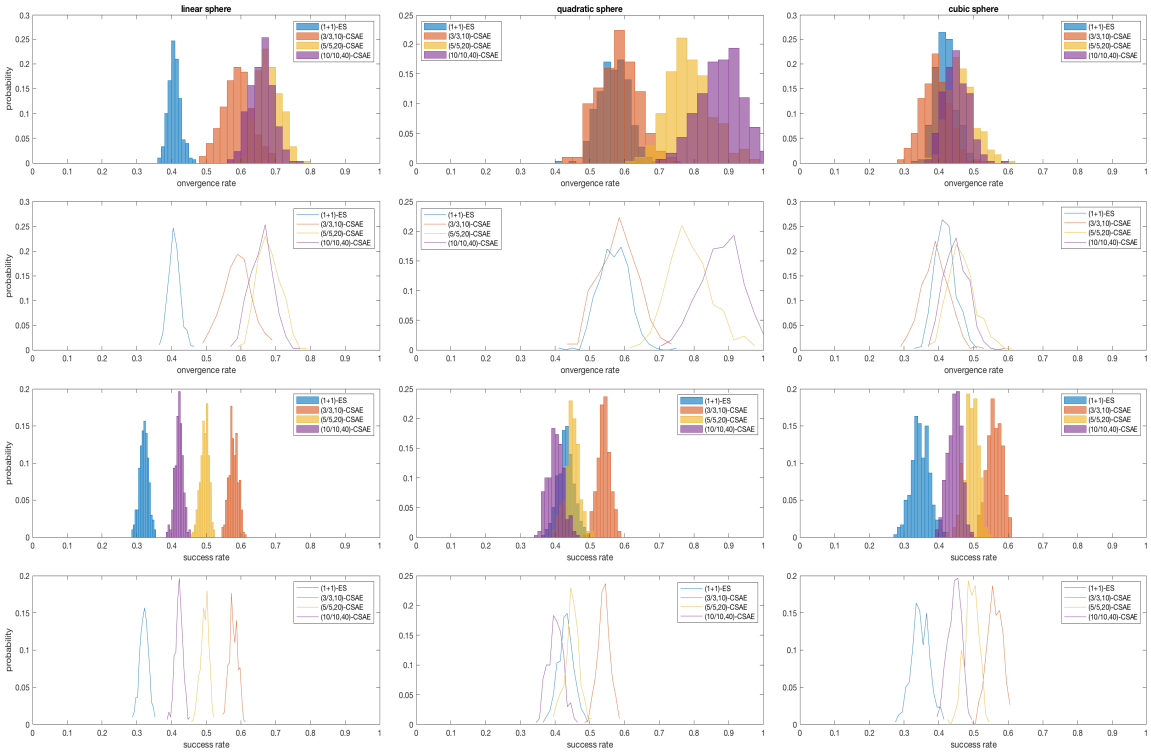

Fig. 6. Result obtained by adapting step size using CSA with emergency. The first two rows show the normalized convergence rate for each run plotted in histogram and normalized probability density function (pdf) respectively. The last two rows represent the success rate (proportion of good step size in each run) plotted in histogram and pdf respectively. is a problem but will be solved as more texts are added

slight decrease in success rate (can also be interpreted as one minus the rate when emergency happens). It makes sense that the CSA with emergency rejects bad steps so that the quality of each step taken improves and therefore larger normalized convergence rate. Using CSA with emergency with a large population suggests an improvement in normalized convergence rate but a slight decrease in success rate for sphere functions. There is a trade-off between the two and finding the optimal relation can be a future goal to work on.

6 CONCLUSIONS

In this paper, we used unbiased Gaussian distributed noise to model the surrogate model's hebaviour. By using this approach, we analyzed the behaviour of surrogate model assisted $(\mu/\mu, \lambda)$-ES on quadratic sphere functions. Based on the analysis and the observation using cumulative step size adaptation, we proposed a step size adaptation mechanism in terms of emergency for the surrogate model assisted $(\mu/\mu, \lambda)$-ES. The strategy is evaluated numerically using a set of test functions. It shows that the step size adaptation mechanism adapted the step size successfully in all runs especially for a potential large population.

In future work, we will study the behaviour of surrogate assisted CMA-ES using the same analysis. Further goals include length scale adaptation mechanism in the Gaussian process. surrogate model accuracy control, and online surrogate models that could possibly further reduce the gap between expected analytical result and experimental result.

REFERENCES

[1] D. V. Arnold and H. -G. Beyer. 2000. Efficiency and mutation strength adaptation of the (mu, mui, lambda)-es in a noisy environment. In *Proceedings of the 6th International Conference on Parallel Problem Solving from Nature* (PPSN VI). Springer-Verlag, London, UK, UK, 39–48. http://dl.acm.org/citation.cfm?id=645825.669117.

[2] D. V. Arnold and H. -G. Beyer. 2004. Performance analysis of evolutionary optimization with cumulative step length adaptation. *IEEE Transactions on Automatic Control*, 49, 4, (Apr. 2004), 617–622.

[3] D. V. Arnold and H.-G. Beyer. 2002. *Noisy optimization with evolution strategies*. Vol. 8. Springer Science & Business Media.

[4] Dirk V. Arnold and Hans-Georg Beyer. 2001. Local performance of the (mu/mu, lambda)-es in a noisy environment. In *Foundations of Genetic Algorithms 6*. W. N. Martin and W. M. Spears, (Eds.) Morgan Kaufmann, San Francisco, 127–141. http://www.sciencedirect.com/science/article/pii/B9781558607347500901.

[5] D. Buche, N. N. Schraudolph, and P. Koumoutsakos. 2005. Accelerating evolutionary algorithms with gaussian process fitness function models. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35, 2, (May 2005), 183–194.

[6] Y. Chen and X. Zou. 2014. Performance analysis of a (1+1) surrogate-assisted evolutionary algorithm. In *Intelligent Computing Theory*. V. Bevilacqua D. Huang and P. Premaratne, (Eds.) Springer International Publishing, Cham, 32–40.

[7] B Dunham, D Fridshal, R Fridshal, and JH North. 1963. Design by natural selection. *Synthese*, 15, 1, 254–259.

[8] K.C. Giannakoglou. 2002. Design of optimal aerodynamic shapes using stochastic optimization methods and computational intelligence. *Progress in Aerospace Sciences*, 38, 1, 43–76. http://www.sciencedirect.com/science/article/pii/S0376042101000197.

[9] Y. Jin, (Ed.) 2005. *Model assisted evolution strategies. Knowledge Incorporation in Evolutionary Computation*. Springer Berlin Heidelberg, Berlin, Heidelberg, 333–355. https://doi.org/10.1007/978-3-540-44511-1_16.

[10] N. Hansen. 2016. The cma evolution strategy: a tutorial. *arXiv preprint arXiv:1604.00772*.

[11] W. J. Hill and W. G. Hunter. 1966. A review of response surface methodology: a literature survey. *Technometrics*, 8, 4, 571–590. eprint: https://amstat.tandfonline.com/doi/pdf/10.1080/00401706.1966.10490404. https://amstat.tandfonline.com/doi/abs/10.1080/00401706.1966.10490404.

[12] M. Schoenauer I. Loshchilov and M. Sebag. 2012. Self-adaptive surrogate-assisted covariance matrix adaptation evolution strategy. In *Proceedings of the 14th annual conference on Genetic and evolutionary computation*. ACM, 321–328.

[13] W. J. Welch J. Sacks, T. J. Mitchell, and H. P. Wynn. 1989. Design and analysis of computer experiments. *Statist. Sci.*, 4, 4, (Nov. 1989), 409–423. https://doi.org/10.1214/ss/1177012413.

[14] Y. Jin. 2011. Surrogate-assisted evolutionary computation: recent advances and future challenges. *Swarm and Evolutionary Computation*, 1, 2, 61–70. http://www.sciencedirect.com/science/article/pii/S2210650211000198.

[15] Y. Jin, M. Olhofer, and B. Sendhoff. 2002. A framework for evolutionary optimization with approximate fitness functions. *IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION*, 6, 481–494.

[16] Y. Jin and B. Sendhoff. 2002. Fitness approximation in evolutionary computation - a survey. In *Proceedings of the 4th Annual Conference on Genetic and Evolutionary Computation* (GECCO'02). Morgan Kaufmann Publishers Inc., New York City, New York, 1105–1112. http://dl.acm.org/citation.cfm?id=2955491.2955686.

[17] A. Kayhani and D. V. Arnold. 2018. Design of a surrogate model assisted (1 + 1)-es. In *Parallel Problem Solving from Nature - PPSN XV - 15th International Conference, Coimbra, Portugal, September 8-12, 2018, Proceedings, Part I*, 16–28. https://doi.org/10.1007/978-3-319-99253-2%5C_2.

[18] I. Loshchilov. 2016. LM-CMA: an Alternative to L-BFGS for Large Scale Black-box Optimization. *Evolutionary Computation*, to appear.

[19] G. Andreas O. Andreas and H. Nikolaus. 1994. A derandomized approach to self-adaptation of evolution strategies. *Evol. Comput.*, 2, 4, (Dec. 1994), 369–380. http://dx.doi.org/10.1162/evco.1994.2.4.369.

[20] A. Ratle. 2001. Kriging as a surrogate fitness landscape in evolutionary optimization. *Artif. Intell. Eng. Des. Anal. Manuf.*, 15, 1, (Jan. 2001), 37–49. http://dx.doi.org/10.1017/S0890060401151024.

[21] I. Rechenberg. 1973. Evolutionsstrategie–optimierung technisher systeme nach prinzipien der biologischen evolution.

[22] H. -P. Schwefel. 1981. *Numerical Optimization of Computer Models*. John Wiley & Sons, Inc., New York, NY, USA.

[23] M. Smith. 1993. *Neural Networks for Statistical Modeling*. (1st ed.). Thomson Learning.

[24] H. Ulmer, F. Streichert, and A. Zell. 2003. Evolution strategies assisted by gaussian processes with improved pre-selection criterion. In *in IEEE Congress on Evolutionary Computation,CEC 2003*, 692–699.

[25] Z. Zhou, Y. S. Ong, P. B. Nair, A. J. Keane, and K. Y. Lum. 2007. Combining global and local surrogate models to accelerate evolutionary optimization. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37, 1, (Jan. 2007), 66–76.