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

[Honours Thesis]

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 behaviours of the algorithms are 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 the sense that 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. ES is flexible and able to solve many optimization problems from two aspects, variation and selection. Firstly, using a stochastic variation from mutation (random sampling of new directions) and recombination (combine the selected mutations) can introduce new unbiased information that may help explore the search space via generating new offspring. Secondly, search using a population of candidate solutions is more robust under moderate noise in multi-modal optimizations, as opposed to some classical search methods like quasi-Newton. Besides, applying a selection on the population can extract potential good step information that may help solve the optimization problem.

Various attempts have been made to reduce the cost by extracting information obtained from points evaluated in previous iterations, such information yields insights into better 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 a much lower cost compared with using the exact objective function. Despite the computation saving of applying surrogate modelling, the estimated objective function value may contain a model bias that can affect the step size being adapted and the direction selected accordingly. Therefore, surrogate modelling is helpful if the computational saving in using the true objective function outshines the potential poor step size and biased direction resulted from the inaccurate surrogate estimation of the candidate solution.

Some of the commonly used surrogate models include but are not limited to Polynomial Regression (PR, Response surface), Gaussian Process (GP, Kriging), neural networks and support vector machine (SVM), a comprehensive survey can be found by Jin [8] and Loshchilov [11]. Most recent works on surrogate model assisted ES considered sophisticated algorithms. These algorithms are heuristic in nature and the step behaviour of the algorithm are not always well interpreted. In this context, a simple model for surrogate models can be helpful in understanding the surrogate behaviour, leading to potential modification to surrogate update or parameter-setting. Recent paper in surrogate assisted EAs by

Kayhani and Arnold [9] analyzes surrogate assisted (1+1)-ES using simple model for surrogate models on simple test functions where the surrogate estimate is modeled using a noisy estimate of the true objective function and the step size behaviour is clear interpreted. As a natural sequence, we investigate the surrogate assisted $(\mu/\mu, \lambda)$-ES using the same surrogate model and following a similar analysis. Since the $(\mu/\mu, \lambda)$-ES generates a population of candidate solutions where the surrogate model can be potentially more fully exploited compared with the (1+1)-ES, it is interesting how much ES is to benefit from the surrogate and the resulting step behaviour as well as the model error would be affected.

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 [9] and exploit the potential benefit of using an extensive sampling with surrogate model assistance. The thesis is organized as follows: In Section 2 we give a brief review of related background and previous analysis that is needed later, in Section 3 we present the experimental result of the proposed local surrogate model-assisted $(\mu/\mu, \lambda)$-ES and study its behaviour on sphere functions. Based on the result, in Section 4, we first apply the well established cumulative step size adaptation (CSA) to the algorithm and present the results. Given the experimental result, we propose an algorithm that is a cross between (1+1)-ES and $(\mu/\mu, \lambda)$-ES where the performance on several test functions are recorded followed by a discussion and future work in Section 5.

2 RELATED WORK

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. A subset of selected offspring is refereed to as the parental population for the next iteration.

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

2.1.1 $(\mu/\rho^{+}_{,}\lambda) - ES$ [6].

A single iteration of the general ES in shown in Alg. 1. Assume a parental population X with size μ, the number of parents for recombination (in offspring generation) ρ and the offspring generated in each iteration Y with size 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 individuals from offspring population Y to update the parental population, no individual from past parental population can be chosen (i.e. only λ individuals are considered for selection). X_{selected} is defined as population ready for recombination after selection that follows

$$X_{\text{selected}} = \begin{cases} \text{select\_best}(\mu, X \cup Y, fX \cup fY), & \text{comma-selection} \\ \text{select\_best}(\mu, Y, fY), & \text{plus-selection,} \end{cases} \tag{1}$$

Algorithm 1 The $(\mu/\rho^{+}\lambda) - ES$

Initialize $N, \rho, \mu, \lambda \in N_{+}$
Initialize parental population $X = \{x_i : i = 1, 2, ..., \mu\}$
Evaluate X using objective function, yielding $fX = \{f(x_i) : i = 1, 2, ..., \mu\}$
while not terminate() **do**
  for $i = 1, 2, ..., \lambda$ **do**
    Generate standard normally distributed $z_i \in \mathbb{R}^N$
    if comma-selection **then**
      $X_{selected} = select\_best(\mu, X, fX)$
    else if plus-selection **then**
      $X_{selected} = select\_best(\mu, X \cup Y, fX \cup fY)$
    end if
    $x_{centroid} = recombine(select\_random(\rho, X_{selected}))$
    $y_i \leftarrow x_{centroid} + \sigma z_i$
    Evaluate y_i, yielding $f(y_i)$
  end for
  $Y = \{y_i : i = 1, 2, ..., \lambda\}$
  $fY = \{f(y_i) : i = 1, 2, ..., \lambda\}$
  $X = select\_best(\mu, Y, fY)$
end while

where fX, fY are objective function value for each individuals in population X and Y, select_best(μ, X, fX) selects the best μ individuals from X according to their fitness recorded in fX i.e. select_best$(\mu, X, fX) = \{x_{i;\lambda} : 1 \leq i \leq \mu\}$ where $f(x_{i;\lambda}) < f(x_{j;\lambda})$, $1 \leq i < j \leq \lambda$.

After selection, recombination is performed, meaning ρ individuals are randomly chosen from X_{selected} and recombined to generate the new offspring. There are two common recombination approaches, intermediate recombination and weight recombination. Intermediate recombination simply takes the average of ρ randomly selected individuals from X_{selected} where the point obtained after recombination is referred to as the centroid x_{centroid}, while weighted recombination uses a weighted average of ρ selected individuals from X_{selected} where the weights are normalized and directly related to individuals' fitness ranking. The centroid x obtained from recombination is defined as

$$x_{\text{centroid}} = \begin{cases} \frac{1}{\rho}\sum_{i=1}^{\rho} x_i, x_i \in \text{select\_random}(\rho, X_{\text{selected}}) & \text{intermediate recombination} \\ \sum_{i=1}^{\rho} x_i w_i x_i, x_i \in \text{select\_random}(\rho, X_{\text{selected}}) & \text{weighted recombination} \end{cases}, \tag{2}$$

where select_random(ρ, X) randomly selects ρ individuals from X without replacement, $0 \leq w_i \leq 1, 1 \leq i \leq \rho$ is a normalized weight related to the fitness of corresponding x_i that has $\sum_{i=1}^{\rho} w_i = 1$ and $0 \leq w_i < w_j \leq 1, f(x_i) > f(x_j)$.

We denote the parental population $X = \{x_1, x_2, ..., x_\mu\}$ and offspring generated in this iteration $Y = \{y_1, y_2, ..., y_\lambda\}$ where $x_i, y_j \in \mathbb{R}^N$ for $i = 1, ..., \mu$ and $j = 1, ..., \lambda$. When generating one offspring, a standard normally distributed mutation vector $z_i \in \mathbb{R}^N$ is generated and added to the centroid with a step size parameter $\sigma \in \mathbb{R}$ and we have

$$y_i = x_{\text{centroid}} + \sigma z_i, 1 \leq i \leq \lambda \tag{3}$$

where z_i represents the mutation, x_{centroid} is the centroid obtained after recombination in each iteration.

Here we consider two special cases for the general ES, namely (1+1)-ES and $(\mu/\mu, \lambda)$-ES. The (1+1)-ES ($\mu = \rho = \lambda = 1$ with plus-selection) generates a single offspring $y = x + \sigma z$ in each generation ($x = x_{\text{centroid}}$, the parental population $Y = \{y\}$), the fitness of y is evaluated and compared with its parent x. The parent population X is updated iff. the

offspring is superior to its parent i.e. $f(y) < f(x)$. Whereas the $(\mu/\mu, \lambda)$-ES ($\mu = \rho$ with comma-selection) generates λ offspring with offspring population $Y = \{y_i : y_i = x_{\text{centroid}} + \sigma z_i, 1 \leq i \leq \lambda\}$, the parental population X is updated by selecting the best μ individuals from Y i.e. $X = \text{select\_best}(\mu, Y, fY)$ where $fY = \{f(y_i) : y_i \in Y\}$.

2.1.2 *Step size adaptation.*

The 1/5th Success Rule. The 1/5th success rule is a basic step size control for ES. The step size is adapted according to the success rate of generating a good offspring i.e. an offspring y with $f(y) < f(x_{\text{centroid}})$ in the case of (1+1)-ES $x_{\text{centroid}} = x$. If the success rate is lower than 1/5, the step size is decreased, otherwise increased. The 1/5 is chosen by Rechenberg [14] after obtaining the optimal success rate (i.e. achieving the largest fitness gain per iteration) for corridor function and quadratic sphere function to be ≈ 0.184 and ≈ 0.270 respectively for $N \to \infty$.

Cumulative Step-Size Adaptation. The step size of $(\mu/\mu, \lambda)$-ES is commonly adapted using cumulative step size adaptation (CSA) proposed by Ostermeier et al [12]. For a strategy with ideally adapted step size, each step should be uncorrelated. If the consecutive steps are negatively correlated, the step size should be decreased. In contrast, if the consecutive 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. Following the recombination in Section 2.1.1

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

$$p^{(g+1)} \leftarrow (1-c)p^{(g)} + \sqrt{\mu c(2-c)}z_{\text{step}}, \tag{4}$$

where $0 < c \leq 1$ helps retain the history information (in generation (g)) and pass that to the evolution path in the next generation $(g+1)$, $\sqrt{\mu c(2-c)}$ is a normalization constant that updates the evolution path via the information obtained in this generation and z_{step} is the direction vector obtained by averaging the directions of ρ randomly chosen individuals after selection is applied. The z_{step} follows

$$z_{\text{step}} = \begin{cases} \frac{1}{\rho}\sum_{i=1}^{\rho} z_{i;\lambda}, & \text{comma-selection} \\ \frac{1}{\rho}\sum_{i=1}^{\rho} z_{i;\lambda+\mu} & \text{plus-selection} \end{cases}, \tag{5}$$

where $y_{i;\lambda} = x_{\text{centroid}} + \sigma z_{i;\lambda}$ with $f(y_{i;\lambda}) < f(y_{j;\lambda}), 1 \leq i < j \leq \lambda$ and $x_{\text{centroid}} + z_{i;\lambda} \in X, x_{\text{centroid}} + z_{i;\lambda+\mu} \in X \cup Y$.

The step size is adapted

$$\sigma^{(g+1)} \leftarrow \sigma \exp\left(\frac{c}{d}\left(\frac{\|p^{(g+1)}\|}{E\|N(0,I)\|} - 1\right)\right), \tag{6}$$

where under random selection and given $p^{(0)} \sim N(o, I)$, the expected length of the search path $p^{(g+1)}$ can be approximated as $E\|N(0,I)\| \approx \sqrt{n}(1 - 1/4n + 1/21n^2)$. In Section 4, we use the well established parameters for CSA from Hansen [5] that follows

$$\begin{cases} c = (\mu+2)/(N+\mu+5) \\ d = 1 + 2\max\left(0, \sqrt{(\mu-1)/(N+1)-1}\right) + c. \end{cases} \tag{7}$$

2.1.3 *Analyzing ES.*

To understand the behaviour EAs, we first introduce analyzing ES on simple test functions where the step behaviours of the algorithm are more likely to be understood. Then proceed to the analysis on noisy sphere where the same analysis can be used to model the surrogate assisted $(\mu/\mu.\lambda)$-ES. Specifically, the $(\mu/\mu.\lambda)$-ES is first analyzed on quadratic sphere and then noisy sphere that models the ideal performance of surrogate model assisted $(\mu/\mu.\lambda)$-ES.

On Sphere Function. Decomposition of z, first proposed by Rechenberg [15] can be used to study the expected step size of the strategy. Vector z can be decomposed as a vector sum $z = z_1 + z_2$, where z_1 is in the direction of the negative gradient of the objective function $\nabla f(x)$ with z_2 orthogonal to z_1. We have z_1 standard normally distributed, while $\|z_2\|^2$ χ^2-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 euclidean 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

$$\begin{aligned}
\delta &= \frac{N}{2R^2}\left(f(x) - f(y)\right) \\
&= \frac{N}{2R^2}(x^T x - (x + \sigma z_{\text{step}})^T (x + \sigma z_{\text{step}})) \\
&= \frac{N}{2R^2}(-2\sigma x^T z_{\text{step}} - \sigma^2 \|z_{\text{step}}\|^2) \\
&\overset{N\to\infty}{=} \sigma^* z_{\text{step},1} - \frac{\sigma^{*2}}{2},
\end{aligned} \tag{8}$$

where $z_{\text{step},1}$, the component of z_{step} pointing to the negative gradient of $f(x)$ and $\overset{N\to\infty}{=}$ denotes the convergence of the distribution $\|z_{\text{step}}\|^N/N = 1$.

On Noisy Sphere Function. The sphere is considered noisy when the fitness evaluation is inaccurate and the objective function on a fixed point may vary in a certain range in different objective function calls. The following uses the analysis and modelling proposed by Arnold and Beyer [3].

The objective function value on noisy sphere can be modeled by adding a Gaussian random variable with mean equals the true objective function value and some variance referred to as noise strength σ_ϵ. The noisy estimate of a candidate solution x follows $f_\epsilon(x) = f(x) + \sigma_\epsilon z_\epsilon$ where $z_\epsilon \in \mathbb{R}$ is a standard normally distributed random variable that randomize the noise generated. By further introduces $\sigma^*_\epsilon = N\sigma_\epsilon/(2R^2)$, the normalized fitness noise [2] and replace the accurate objective function evaluation with the noisy estimate, the normalized fitness advantage of y on noisy sphere when $n \to \infty$ in Eq. (8) is

$$\begin{aligned}
\delta_\epsilon &= \frac{N}{2R^2}\left(f_\epsilon(x) - f_\epsilon(y)\right) \\
&= \frac{N}{2R^2}(x^T x - (x + \sigma z_{\text{step}})^T (x + \sigma z_{\text{step}}) + \sigma_{x,\epsilon} z_{x,\epsilon} - \sigma_{y,\epsilon} z_{y,\epsilon}) \\
&= \frac{N}{2R^2}(-2\sigma x^T z_{\text{step}} - \sigma^2 \|z_{\text{step}}\|^2 + \sigma_\epsilon z_\epsilon) \\
&\overset{N\to\infty}{=} \delta + \sigma^*_\epsilon z_\epsilon,
\end{aligned} \tag{9}$$

the term $\sigma_{x,\epsilon} z_{x,\epsilon}, \sigma_{y,\epsilon} z_{y,\epsilon}$ denote the added noise for the recombined parent x and the offspring y respectively and w.l.o.g. assuming independence $\sigma_{x,\epsilon} z_{x,\epsilon}, \sigma_{y,\epsilon} z_{y,\epsilon}$ can be viewed as a vector sum denoted as $\sigma_\epsilon z_\epsilon$. By substituting the noise term, we get the simplified Eq. (9).

The expected value of the normalized change in objective function value

$$\begin{aligned}\Delta &= -\frac{N}{2} E\left[\log f_\epsilon(y) - \log f_\epsilon(x)\right] \\ &= -\frac{N}{2} E\left[\log \frac{f_\epsilon(x^{(t+1)})}{f_\epsilon(x^{(t)})}\right],\end{aligned} \tag{10}$$

where $x^{(t)}$ is the centroid of parental population (recombined parent) in timestamp t, the equation is normalized in terms of dimensionality.

In each iteration, λ offspring are evaluated, the objective function evaluation per iteration is λ (for $(\mu/\mu, \lambda)$-ES), therefore the normalized progress rate when dimensionality $N \to \infty$ is equation (7) from [3]

$$\eta = \frac{1}{\lambda} E[\Delta] \approx \frac{1}{\lambda}\left(\frac{\sigma^* c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2}} - \frac{(\sigma^*)^2}{2\mu}\right), \tag{11}$$

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

$$c_{\mu/\mu,\lambda} = \frac{\lambda - \mu}{2\pi}\binom{\lambda}{\mu}\int_{-\infty}^{\infty} e^{-x^2}\left[\Phi(x)\right]^{\lambda-\mu-1}\left[1-\Phi(x)\right]^{\mu-1} \mathrm{d}x, \tag{12}$$

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

2.2 Surrogate Model

Surrogate model is a computational model constructed based on the data evaluated using true objective function. The surrogate acts as am approximation to the true objective function that is costly in most cases, so that the objective function estimation using the surrogate model although inaccurate can be achieved at vanishing cost.

The surrogate model can be applied to EAs as an approximate fitness to accelerate the evolution process [13]. Despite the computational saving when using a surrogate model, issues can occur when the surrogate built leads to a false optima (i.e. the optima does not exist in the true objective function). This can leads to potential divergence and unstable optimization path where the convergence property of the ES may not be well preserved.

Gaussin Process (GP). The GP is defined using the notation of slides by Arnold : let $f(x)$ be an unkown scalar functuoon and $x \in \mathbb{R}^{\mathbb{N}}$ is a point in an $N-$dimensional space. Evaluate f at K data points $X = (x_1, x_2, ..., x_n)$ yields function values $f = (f(x_1), f(x_2), ..., f(x_n))$. We want to predict new function values $f(X_*)$ of a test set X_* with size n_*.

The vector of known function values and the predicted value (f, f_*) are jointly normally distributed with mean (μ, μ_*) and covariance matrix

$$\begin{bmatrix} K & K_* \\ K_*^T & K_{**} \end{bmatrix} \tag{13}$$

GP model

Illustrate GP in 1D

short but copy and bunch of refs

3.2 Surrogate assisted EAs, Other algorithms uses GP models

2.2.1 *Surrogate model management.*

One remedy in prevention of the false optima introduced by the surrogate model is to use surrogate model management that controls the choice of evaluation between true objective function and the surrogate model. Surrogate model management, according to Jin [7], can be divided into three categories, namely, individual-based, generation-based, and population-based. In individual-based surrogate, some of the individuals within a generation are evaluated using the surrogate model. Others using the true objective function. Whereas the generation-based surrogate model, all the offspring in one generations are evaluated using either the surrogate model or the true objective function. In the case of population based management, the population is divided into some subpopulations where each subpopulation has its own surrogate model. The surrogate model only evaluates the offspring generated in the population it belongs to.

2.2.2 *Surrogate model approach.*

Two approaches will be discussed. The first approach uses an EA to optimize a surrogate model and the second uses a surrogate model to assist EA.

The first one proposed by Buche et al. [4] uses an EA optimized surrogate that is individual-based where a fraction of individuals in the generation is evaluated depending on the surrogate model error.

The other approach , ensuring the model accuracy and the optimal can be found. The other approach proposed by Kern et al. [10] throughly evaluate the surrogate model to find a relative precise optima where the surrogate model is updated whenever the optimal solution is inconsistent in two consecutive generations, i.e. new points are evaluated and added to the training set for the surrogate model.

3 ANALYSIS

To understand the potential implications of using surrogate models in EAs with varying population size, in this section, we use a simple model for the use of a surrogate model. Specifically, we propose an EA that, in each iteration, a population of λ offspring Y are generated and then evaluated by the surrogate model instead of true objective function followed by a intermediate recombination based on the inaccurate surrogate estimate where a true objective function evaluation is made for the centroid of the selected offspring, referred to as the recombined parent x. Offspring generation is the same as is discussed in $(\mu/\mu, \lambda)$-ES. The major difference in analysis between $(\mu/\mu, \lambda)$-ES on noisy sphere and $(\mu/\mu, \lambda)$-ES with surrogate model assistance on sphere lies in the evaluation of recombined parent x where the former applies a noisy estimate and the latter uses a true objective function call. Therefore, the normalized fitness advantage of y over x is

$$\begin{aligned}
\delta_\epsilon &= \frac{N}{2R^2}\left(f(x) - f_\epsilon(y)\right) \\
&= \frac{N}{2R^2}(x^T x - (x + \sigma z_{\text{step}})^T (x + \sigma z_{\text{step}}) - \sigma_{y,\epsilon} z_{y,\epsilon}) \\
&= \frac{N}{2R^2}(-2\sigma x^T z_{\text{step}} - \sigma^2 \|z_{\text{step}}\|^2 + \sigma_\epsilon z_\epsilon) \\
&\overset{N\to\infty}{=} \delta + \sigma_\epsilon^* z_\epsilon,
\end{aligned} \tag{14}$$

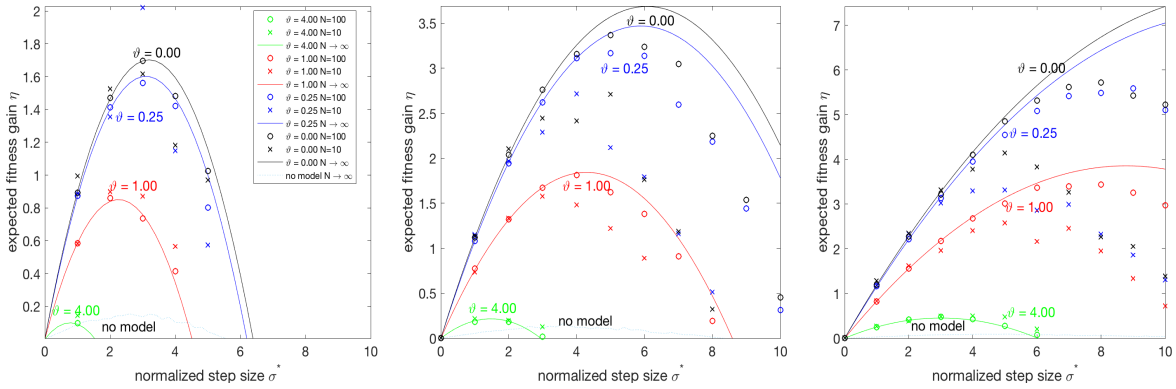

Fig. 1. The figures from left to right shows the expected single step behaviour of the surrogate model assisted $(\mu/\mu, \lambda)$-ES with unbiased Gaussian distributed surrogate error with $\lambda = 10, 20, 40$ respectively where $\mu = \lceil \lambda/4 \rceil$. The solid lines are the results obtained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

where by changing variable, the noise term $-\sigma_{y,\epsilon} z_{y,\epsilon}$ can be replaced by $\sigma_\epsilon z_\epsilon$ that gives exactly the same result as Eq. ïijĹ9). The analysis in Section 2.1.3 still holds. In this context, the noise-to-signal ratio ϑ can be interpreted as the measure of the surrogate model quality relative to step size of the algorithm. This analysis could be extend to biased surrogate models where the distribution mean is different from the exact objective function value[9].

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 without model assistance), therefore the normalized progress rate when dimensionality $N \to \infty$, by substituting λ with 1 in Eq. (11), the normalized progress rate is

$$\eta = \frac{1}{1} E[\Delta] \approx \frac{\sigma^* c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2}} - \frac{(\sigma^*)^2}{2\mu}, \tag{15}$$

To obtain the opt. expected fitness gain η_{opt} and its corresponding opt. normalized step size σ^*_{opt}, we take derivative of equation (15) over σ^* and get the following

$$\sigma^*_{opt} = \frac{\mu c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2}} \tag{16}$$

$$\eta_{opt} = \frac{\sigma^*_{opt} c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2}} - \frac{(\sigma^*_{opt})^2}{2\mu} \tag{17}$$

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. (15) (12). 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

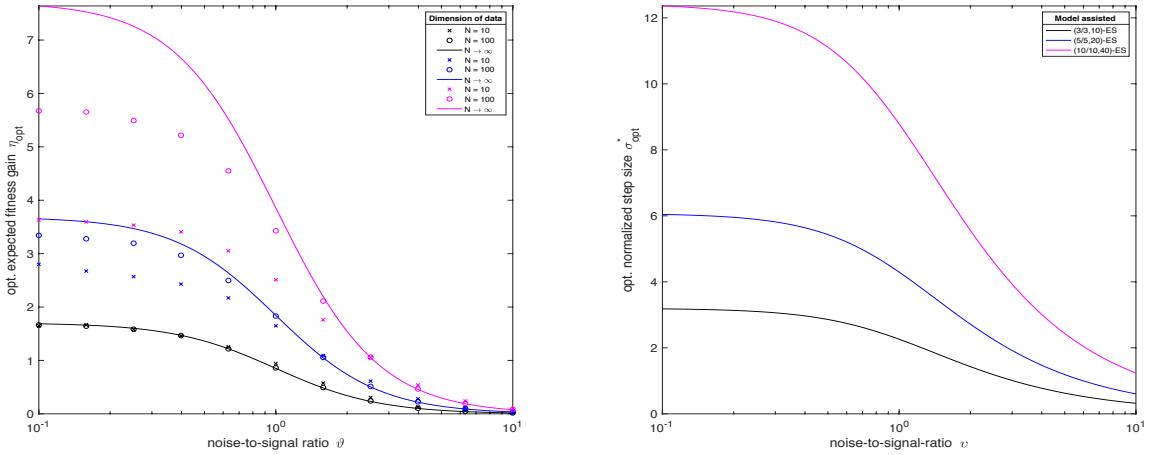

Fig. 2. Opt. expected fitness gain and corresponding opt. normalized step size of the surrogate model assisted $(\mu/\mu, \lambda)$-ES plotted against the noise-to-signal ratio. The line and dots with colour black, blue, magenta represent (3/3, 10)-ES, (3/3, 10)-ES, (3/3, 10)-ES The solid line represents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

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 result of the surrogate assisted (1+1)-ES [9] 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 (16) 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.

4 STEP SIZE ADAPTATION

4.1 Cumulative step size adaptation

Algorithm 2 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

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 step size adaptation described in Section 2.1.2 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. One signle iteration fo the surrogate model assisted $(\mu/\mu, \lambda)$-ES using CSA is shown in Alg. 2.

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 [16]) and quartic function [9] 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 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

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)

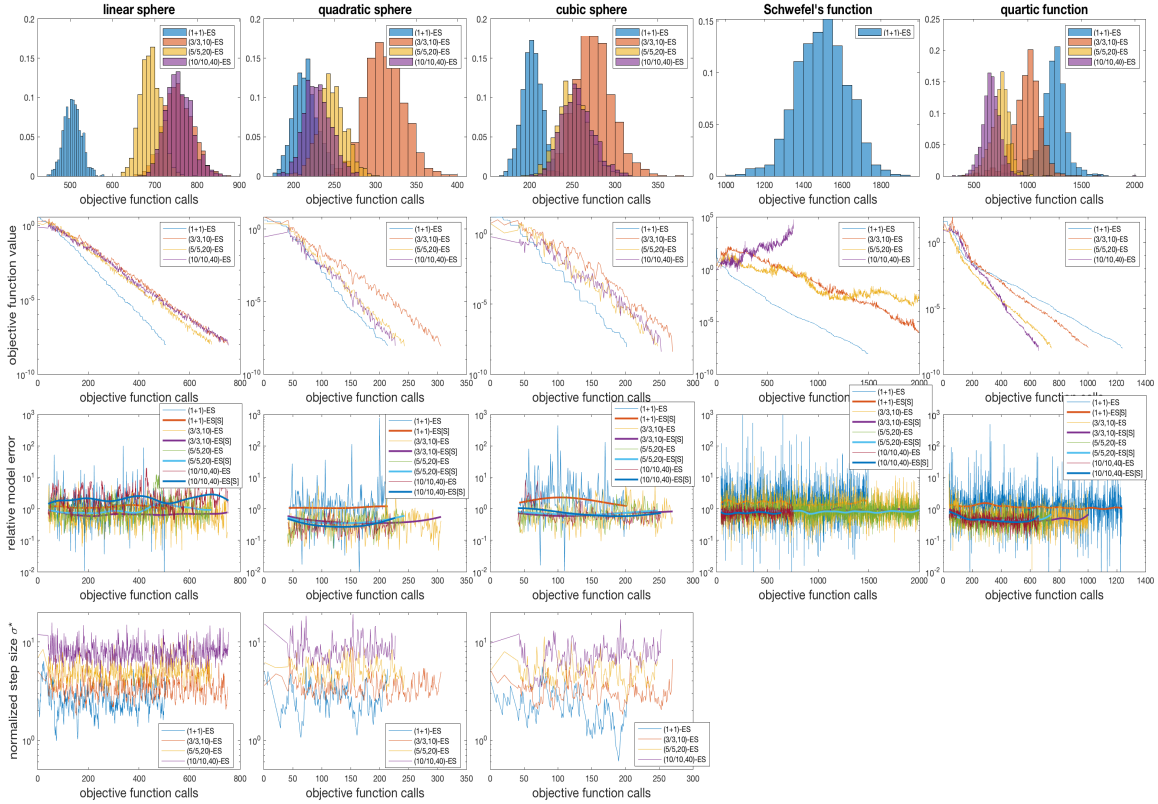

Fig. 3. Result obtained by adapting step size using CSA. Top row: Histogram showing the number of objective function calls needed to solve the five test problems. Second row: Convergence graphs for median runs. Third row: Relative model error obtained in median runs ([S] denotes the smoothed plot). Last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

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 Table 1. The result by Kayhani and Arnold [9] using surrogate assisted (1+1)-ES is also included for comparison. The

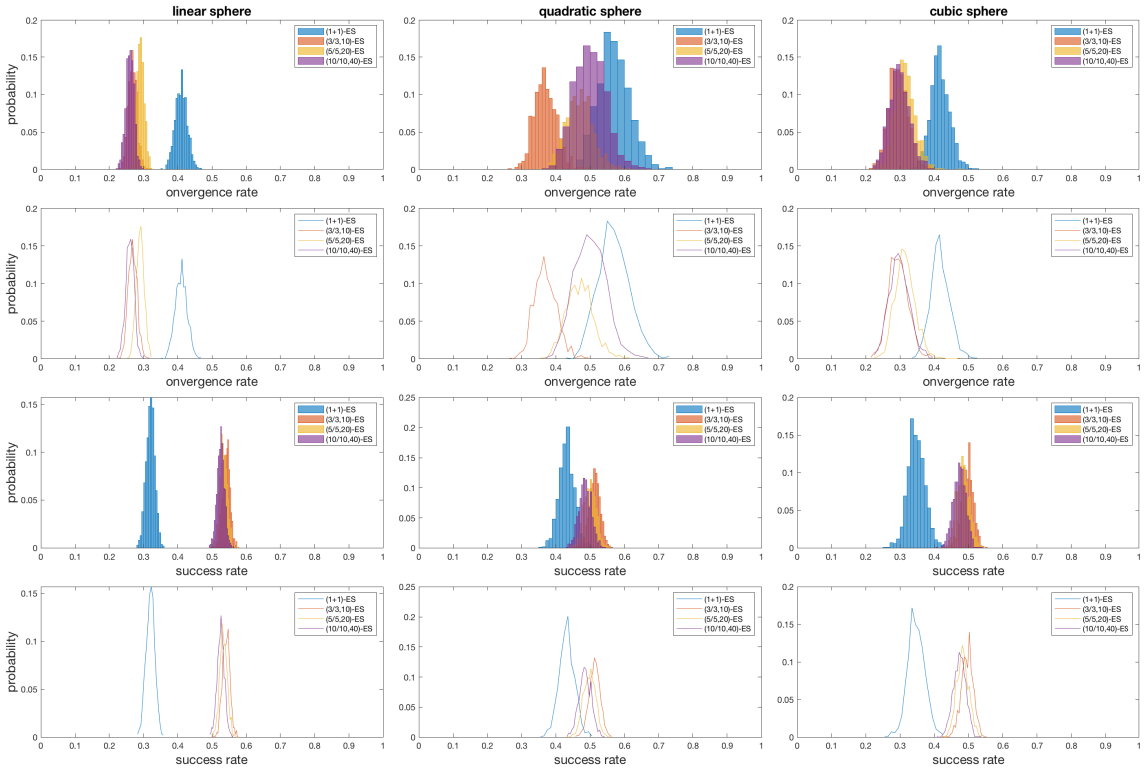

Fig. 4. Result obtained by adapting step size using CSA. The first two rows show the normalized convergence rate for each run plotted in histogram and normalized probability density function (pdf) respectively. The last two rows represent the success rate (proportion of good step size in each run) plotted in histogram and pdf respectively.1. cannot say histogram of convergence [logarithm of the equation, 2. duplicated fig.s 3. step size is properly adapted [CSA only works when

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 [9] 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

$$c = \begin{cases} -n \left[\log \left(\frac{f(x_{t+1})}{f(x_t)} \right) \right], & \text{linear sphere} \\ -\frac{n}{2} \left[\log \left(\frac{f(x_{t+1})}{f(x_t)} \right) \right], & \text{quadratic sphere} \\ -\frac{n}{3} \left[\log \left(\frac{f(x_{t+1})}{f(x_t)} \right) \right], & \text{cubic sphere.} \end{cases} \tag{18}$$

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.

4.2 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 [9], 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. 3 by adding a conditional statement comparing the fitness of the offspring obtained with its parent as is illustrated in line 15 Alg. 3. 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 3 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

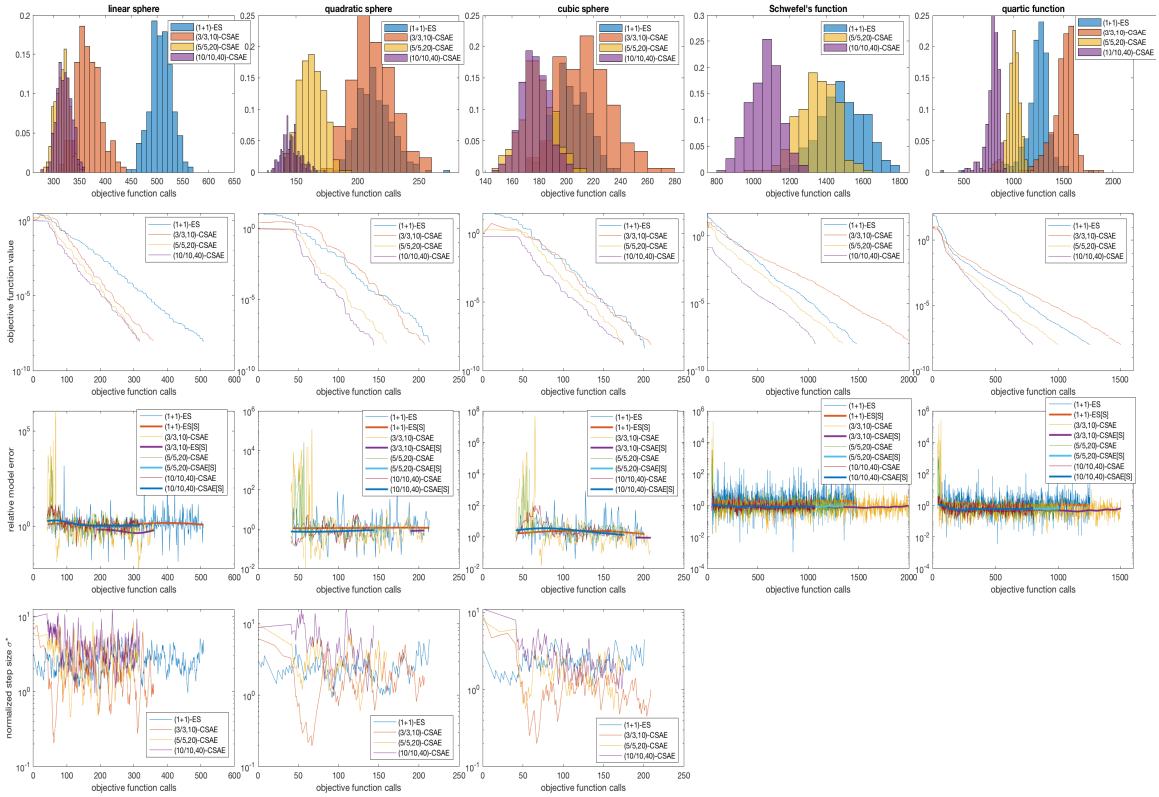

Fig. 5. Result obtained by adapting step size using CSA with emergency (denoted with CSAE). Top row: Histogram showing the number of objective function calls needed to solve the five test problems. Second row: Convergence graphs for median runs. Third row: Relative model error obtained in median runs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

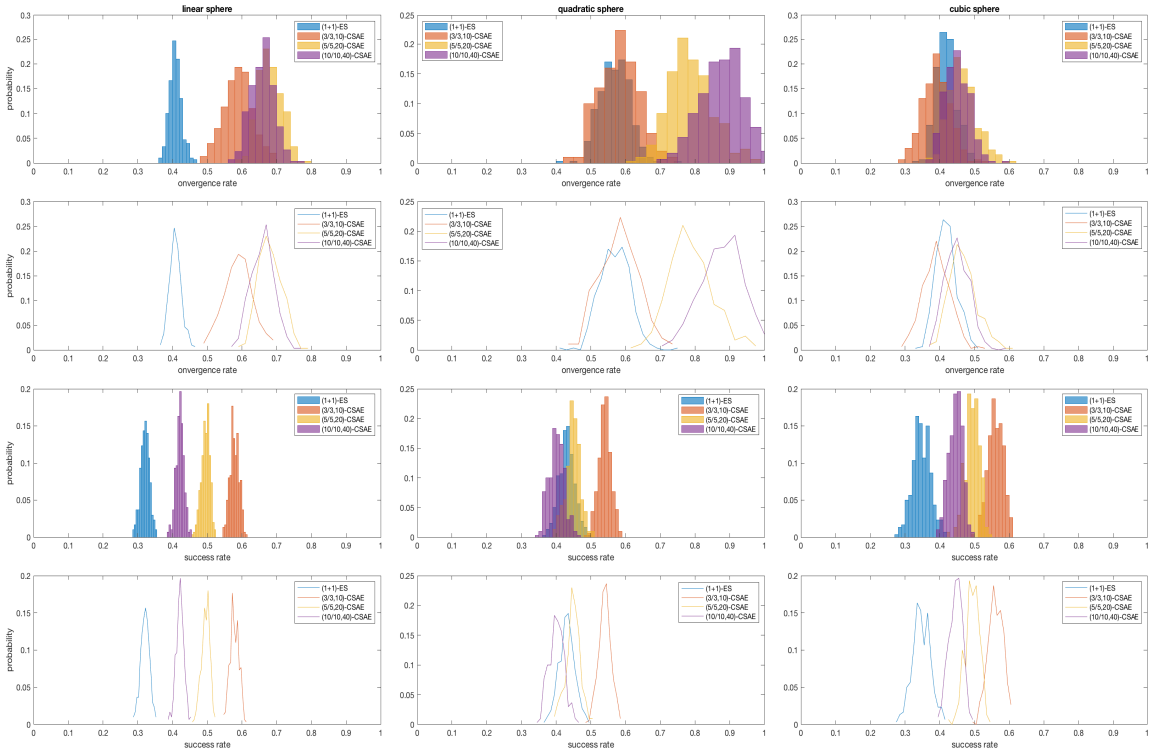

Fig. 6. Result obtained by adapting step size using CSA with emergency. The first two rows show the normalized convergence rate for each run plotted in histogram and normalized probability density function (pdf) respectively. The last two rows represent the success rate (proportion of good step size 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.

5 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.

[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. 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.

[4] 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.

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

[6] N. Hansen, D. V. Arnold, and A. Auger. 2015. Evolution strategies. In *Springer handbook of computational intelligence*. Springer, 871–898.

[7] Y. Jin. 2005. A comprehensive survey of fitness approximation in evolutionary computation. *Soft Comput.*, 9, 1, (Jan. 2005), 3–12.

[8] Y. Jin. 2011. Surrogate-assisted evolutionary computation: recent advances and future challenges. *Swarm and Evolutionary Computation*, 1, 2, 61–70.

[9] 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.

[10] S. Kern, N. Hansen, and P. Koumoutsakos. 2006. Local meta-models for optimization using evolution strategies. In *Parallel Problem Solving from Nature - PPSN IX*. Thomas Philip Runarsson, Hans-Georg Beyer, Edmund Burke, Juan J. Merelo-Guervós, L. Darrell Whitley, and Xin Yao, (Eds.) Springer Berlin Heidelberg, Berlin, Heidelberg, 939–948.

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

[12] 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.

[13] A. Ratle. 1998. Accelerating the convergence of evolutionary algorithms by fitness landscape approximation. In *Proceedings of the 5th International Conference on Parallel Problem Solving from Nature* (PPSN V). Springer-Verlag, London, UK, UK, 87–96. http://dl.acm.org/citation.cfm?id=645824.668750.

[14] I. Rechenberg. 1973. *Evolutionsstrategie : Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Number 15 in *Problemata*. Frommann-Holzboog, Stuttgart-Bad Cannstatt.

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

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