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 to 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 both the step size being adapted and the direction selected. Therefore, surrogate modelling is helpful if the computational saving in using the true objective function outshines the potential poor step size and biased direction resulting 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), and a comprehensive survey can be found by Jin [10] and Loshchilov [12]. Most recent works on surrogate model assisted ES consider 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. A recent paper in surrogate assisted EAs proposed by Kayhani and Arnold [11] analyzes surrogate assisted (1+1)-ES using a simple model for surrogate models on simple test functions where the surrogate estimate is modelled using a noisy estimate of the true objective function. The step size behaviour of the strategy on the test function is clearly 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 the model error would be affected and how much the ES is to benefit from the surrogate and the resulting step behaviour.

This thesis intends 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 [11] and to 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 the 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), a category of Evolutionary Algorithms (EAs), is a nature-inspired direct search method that addresses optimization problems by using stochastic variation and selection. In each iteration, new offspring are generated from the parental population through mutation, followed by a selection based on the fitness of the offspring. A subset of selected offspring is referred to as the parental population for the next iteration.

ESs are commonly used in black-box optimization where the search space $\mathbb{R}^N$ is N-dimensional, whereas the objective function value is 1-dimensional (in $\mathbb{R}$). We consider minimization of an objective function $f : \mathbb{R}^N \to \mathbb{R}$ that 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 $(\mu/\rho \overset{+}{,} \lambda)$ – ES [14].

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 ($\overset{+}{,}$) 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 populations can be chosen (i.e., only λ individuals are considered for selection). $X^{(g+1)}$ is defined as the parental population in generation $(g + 1)$ that follows

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

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

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

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

After selection, recombination is performed, meaning ρ individuals are randomly chosen from $X^{(g+1)}$ and recombined to generate the new offspring. There are two common recombination approaches, intermediate recombination and weighted recombination. Intermediate recombination simply takes the average of ρ randomly selected individuals from $X^{(g+1)}$ where the point obtained after recombination is referred to as the centroid $x_{\text{centroid}}^{(g+1)}$, while weighted recombination uses a weighted average of ρ selected individuals from $X^{(g+1)}$ where the weights are normalized and directly related to individuals' fitness ranking. The centroid $x_{\text{centroid}}^{(g+1)}$ obtained from recombination in generation $(g + 1)$ is defined as

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

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

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

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

where $z_i^{(g+1)}$ represents the mutation, $x_{\text{centroid}}^{(g+1)}$ is the centroid obtained after recombination in generation $(g+1)$.

Here we consider two special cases of 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 ($x = x_{\text{centroid}}$ and the parental population $Y = \{y\}$), and 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 $fY = \{f(y_i) : y_i \in Y\}$). A Covariance Matrix Adaptation Evolution Strategy (CMA-ES) adds a covariance matrix to a general ES where the mutation is generated from the covariance matrix that approximates the inverse of the Hessian and is updated in each generation. CMA-ES is very helpful in dealing with ill-conditioning (the condition number of the Hessian matrix is very large).

2.1.2 Step size adaptation.

The 1/5th Success Rule. The 1/5th success rule is a basic step size control for (1+1)-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 [18] 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 $N \longrightarrow \infty$.

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 [15]. 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.

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

Define the search path as

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

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 generation g and $z_{\text{step}}^{(g)}$ is the direction vector obtained by averaging the directions of ρ randomly chosen individuals after selection is applied. The $z_{\text{step}}^{(g)}$ follows

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

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

The step size is adapted

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

where under random selection and given $p^{(0)} \sim \mathcal{N}(0,I)$, the expected length of the search path $p^{(g+1)}$ can be approximated as $E\|\mathcal{N}(0,I)\| \approx \sqrt{N}(1 - 1/4N + 1/21N^2)$. In Section 4, we use the well established parameters for CSA from Hansen [8] 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 of EAs, we first introduce analyzing ES on simple test functions where the step behaviours of the algorithm are more likely to be understood, specifically on sphere function where . 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 defined as $f(x) = \sum_{i=1}^{N} x_i^2$ 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 [19] 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, $\|z_2\|^2$ χ^2-distributed with $N - 1$ degree of freedom and $\|z_2\|^2/N \overset{N\to\infty}{=} 1$ (see [5]). 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 $\sigma^* = N\sigma/R$. The normalized fitness gain of y over x given mutation z follows

$$\begin{aligned} \delta^*(z) &= \frac{N}{2R^2}\left(f(x) - f(y)\right) \\ &= \frac{N}{2R^2}(x^T x - (x + \sigma z)^T (x + \sigma z)) \\ &= \frac{N}{2R^2}(-2\sigma x^T z - \sigma^2\|z\|^2) \\ &\overset{N\to\infty}{=} \sigma^* z_1 - \frac{\sigma^{*2}}{2}, \end{aligned} \tag{8}$$

where z_1 is the component of z pointing to the negative gradient of $f(x)$ and $\overset{N\to\infty}{=}$ denotes the convergence of the distribution $\|z\|^N/N = 1$.

In the case of $(\mu/\mu, \lambda)$-ES, the mutation vector $z_{\text{step}} = \sum_{i=1}^{\mu} z_{i;\lambda}$ is the averaged z taken by the best μ candidate solutions. So the normalized fitness gain given z_{step} is

$$\begin{aligned}
\delta^*(z_{\text{step}}) &= \frac{N}{2R^2}(x^T x - (x + \sigma z_{\text{step}})^T (x + \sigma z_{\text{step}})) \\
&= \frac{N}{2R^2}(-2\sigma x^T z_{\text{step}} - \sigma^2 \|z_{\text{step}}\|^2) \\
&\overset{N \to \infty}{=} \sigma^* z_{\text{step},1} - \frac{\sigma^{*2}}{2\mu},
\end{aligned} \tag{9}$$

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

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 modelled by adding a Gaussian random variable with mean equals the true objective function value and some standard deviation referred to as noise strength σ_ϵ. The noisy estimate of a candidate solution y follows $f_\epsilon(y) = f(y) + \sigma_\epsilon z_\epsilon$ 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 gain of y given mutation z on noisy sphere when $n \to \infty$ in Eq. (9) is

$$\begin{aligned}
\delta_\epsilon^*(z) &= \frac{N}{2R^2}\left(f(x) - f_\epsilon(y)\right) \\
&= \frac{N}{2R^2}(x^T x - (x + \sigma z_{\text{step}})^T (x + \sigma z_{\text{step}}) + \sigma_\epsilon z_\epsilon) \\
&= \frac{N}{2R^2}(-2\sigma x^T z_{\text{step}} - \sigma^2 \|z_{\text{step}}\|^2 - \sigma_\epsilon z_\epsilon) \\
&= \delta(z) + \sigma_\epsilon z_\epsilon \\
&\overset{N \to \infty}{=} \sigma^*(z_1 + \vartheta z_\epsilon) - \frac{\sigma^2}{2},
\end{aligned} \tag{10}$$

where $\vartheta = \sigma_\epsilon^*/\sigma^*$ is the noise-to-signal ratio, defined to measure the noise level relative to the algorithm's step size, the term $+\sigma_\epsilon z_\epsilon$ denotes the added noise. Since z_ϵ is standard normally distributed, so is $-z_\epsilon$, by substituting $z_\epsilon = -z_\epsilon$ and adding the noise term to Eq. (9), we get the simplified Eq. (10).

The normalized progress rate when dimensionality $N \to \infty$ follows

$$\eta = \frac{1}{\lambda} E[\delta^*(z_{\text{step}})] \approx \frac{1}{\lambda} E\left[\sigma^*(z_{\text{step},1} + \vartheta z_\epsilon) - \frac{\sigma^2}{2\mu}\right] \tag{11}$$

where $1/\lambda$ denotes the number of offspring evaluated in each iteration i.e., the objective function evaluation per iteration for $(\mu/\mu, \lambda)$-ES is λ.

The expected value of $z_{\text{step},1}$ derived by Arnold [3] is

$$\begin{aligned} E\left[z_{\text{step},1}\right] &= E[\frac{1}{\mu}\sum_{i=1}^{\mu} z_{(i;\lambda),1}] \\ &= \frac{1}{\mu}\sum_{i=1}^{\mu} E[z_{(i;\lambda),1}] \\ &= \frac{1}{\mu}\sum_{i=1}^{\mu}\int_{-\infty}^{\infty} x p_{i;\lambda}(x)dx \\ &= \frac{c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2}} \end{aligned} \tag{12}$$

where $z_{(i;\lambda),1}$ is the component of $z_{(i;\lambda)}$ in the negative gradient direction of the objective function $f(x)$, $p_{k;\lambda}$ corresponds to the probability density function of $z_{(i;\lambda),1}$ that has $f(x+z_{i;\lambda}) < f(x+z_{j;\lambda})$, $1 \leq i < j \leq \lambda$, $c_{\mu/\mu,\lambda}$ is the $(\mu/\mu,\lambda)$-progress coefficient derived by Arnold and Beyer [1] that follows

$$c_{\mu/\mu,\lambda} = \frac{\lambda-\mu}{2\pi}\binom{\lambda}{\mu}\int_{-\infty}^{\infty} e^{-x^2}\left[\Phi(x)\right]^{\lambda-\mu-1}\left[1-\Phi(x)\right]^{\mu-1}\,\mathrm{d}x, \tag{13}$$

where Φ is the normal cumulative distribution function. The integral can be solved numerically.

Therefore by Eq. (11) (12), when dimensionality $N \longrightarrow \infty$ the normalized progress rate can be simplified as

$$\eta = \frac{1}{\lambda}E[\delta^*(z_{\text{step}})] \approx \frac{1}{\lambda}\left(\frac{\sigma^* c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2}} - \frac{(\sigma^*)^2}{2\mu}\right) \tag{14}$$

2.2 Surrogate Model

Surrogate models are computational models constructed based on the data evaluated using true objective function. The surrogate acts as an approximation to the true objective function that is costly in most cases; therefore, 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 [17]. Despite the computational saving when using a surrogate model, issues can occur when the surrogate built leads to a false optimum (the optimum does not exist in the true objective function), leading to potential divergence and unstable optimization path where the convergence property of the ES may not be well preserved [10]. Two approaches will be discussed that give potential solutions to address this issue by ensuring model accuracy. The first approach uses an EA to optimize a surrogate model, and the second uses a surrogate model to assist EA.

2.2.1 Gaussian Process.

A GP is a probabilistic model where the observations are in a continuous domain $\mathbb{R}^{\mathbb{N}}$. It is completely determined by its mean function

$$m(x) = E[f(x)], \tag{15}$$

commonly assumed to be zero, and a covariance function (a positive definite kernel) follows

$$\kappa(x, x\prime) = E[(f(x) - m(x))(f(x\prime) - m(x\prime))]. \tag{16}$$

$\forall x_i \in \mathbb{R}^N, i = 1, 2, ..., n$, the distribution of the function value $f(x_i), i = 1, 2, ..., n$ is jointly Gaussian with mean $\mu = (m(x_1), m(x_2), ..., m(x_n))$, and covariance matrix $\sum$ with entry $\sum_{ij} = \kappa(x_i, x_j)$.

Squared exponential kernel, a commonly used covariance function is applied in the context

$$\kappa(x, x\prime) = \exp\left(-\frac{\|x - x\prime\|^2}{2\theta^2}\right), \tag{17}$$

where θ is the length scale factor of the GP.

The GP is defined using the notation and content from [16] and [13]: let $f(x)$ be an unknown scalar function and $x \in \mathbb{R}^{\mathbb{N}}$ is a point in an $N-$dimensional space. Evaluating 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

$$\begin{bmatrix} K & K_* \\ K_*^T & K_{**} \end{bmatrix}, \tag{18}$$

where $K = \kappa(X, X), K_* = \kappa(X, X_*)$, and $K_{**} = (X_*, X_*)$.

By Bayes' rule, f_* is normally distributed with mean and covariance

$$\mu_* = K_*^T K^{-1} f \tag{19}$$

$$\sum_* = K_{**} - K_*^T K^{-1} K_* \tag{20}$$

2.2.2 Surrogate model assisted ES.

Surrogate model assisted by ES. Gaussian Process Optimization Procedure (GPOP) proposed by Buche et al. [7] uses an EA optimized surrogate where the surrogate is individual-based, meaning a fraction of individuals in each generation is evaluated depending on the surrogate model error.

A GP is constructed using a set of evaluated points from random sampling in previous iterations, a CMA-ES (briefly described in Section 2.1.1) is used for parameter searching in order to find the minimum of the GP prediction. To avoid false optima, a merit function is used to help explore new regions of the decision space, regarded as the fitness function for the CMA-ES. Predicted standard deviation is used as the merit function that follows

$$f_M(x) = \hat{f}(x) - \alpha\sigma_{\mathrm{GP}}(x), \tag{21}$$

where $\hat{f}(x)$ and $\sigma_{\mathrm{GP}}(x) = \sqrt{\sum_*}$ (considering a single input $x\prime$ in Eq. (20)) are the GP estimate and standard deviation for data x, $\alpha \geq 0$ balances the two terms via scaling the density measure where a larger α pushes the search harder into unexplored region. 4 merit functions with $\alpha = 0, 1, 2, 4$ are used and optimized in the context. Finally, the resulting minimum is added to the training set.

To approximate the true objective function with arbitrary precision, a local model is preferred because it uses a smaller number of training points and provides a more precise local approximation of the objective function, compared with a global model. A local model approximates the objective function within a limited region, as opposed to a global approximation that approximates the whole objective function using all evaluated points. In the local model, the new points sampled moves around the current best solution and is restricted to a neighbourhood of the evaluated current best point for model accuracy, which means only points within the region are considered reliable using surrogate estimates. In each iteration, N_C points are sampled. If more than half of the sampled points are not evaluated successfully (falls out the well-approximated region using GP), another $N_C/2$ points are generated using the $(2, 10) - CMAES$ and evaluated using the true objective function followed by a model update that chooses the best evaluated point x_{best} and re-builds the training set using the N_C closest points to x_{best} and N_R most recent successfully evaluated points. The GP parameter is then optimized by ES, each predicted optimum is found using the corresponding merit function where the unevaluated optimum is evaluated using the true objective function and the process repeats.

Experimental results show GPOP can be effective in solving unimodal functions and can achieve an average speed-up (number of objective function evaluations to solve the optimization problem within required precision using CMA divided by that of using GPOP) of 3 to 6 for quadratic sphere, Schwefel problem and Rosenbrock's function. It shows that the GPOP cannot determine the global minima for $n > 2$ given the large number of local minima and the high oscillation of the function value between minima. GPOP is problem-specific in terms of setting training set size The main problem with GPOP is its higher computational cost of the optimization procedure that scales $O(N^3)$ and $O(N)$ with training size and problem dimension respectively, therefore GPOP should be considered mainly for computational expensive problems.

ES assisted by surrogate model. The other approach, Local Metamodel Covariance Matrix Adaptation ES (lmm-CMA) proposed by Kern et al. [20] throughly evaluates the surrogate model (metamodel) to find a relative precise optimum predicted by the model where the model is updated whenever the ranking of the best μ solutions is inconsistent in two consecutive metamodel iterations.

An approximate ranking procedure can be applied to ensure the model accuracy without knowing the true ranking of the complete population. The proposed procedure acts as a ranking procedure in CMA-ES that ensures an effective ranking by evaluating a batch of individuals in each metamodel iteration where the model is updated once the model is regarded inaccurate. The CMA-ES used can be viewed as a $(\mu/\mu, \lambda)$-ES with a covariance matrix see Section 2.1.1. In ES generation g, λ offspring are generated where the best μ offspring $y_{1:\mu,\lambda}$ are selected according to the predicted fitness by the model $\hat{f}$. The ranking of the μ selected offspring is recorded as $\text{ranking}^{(i)}_{1-\mu;\lambda}$ in model iteration i and the best n_{init} selected individuals are evaluated using the true objective function for training set update. The selected λ individuals $y_{1:\mu,\lambda}$ are evaluated again by the model in the next model iteration $i + 1$, leading to a new ranking $\text{ranking}^{(i+1\prime)}_{1-\mu;\lambda}$ followed by a comparison. If $\text{ranking}^{(i)}_{1-\mu;\lambda} = \text{ranking}^{(i+1)}_{1-\mu;\lambda}$, the model is regarded as reliable, otherwise the next n_b best unevaluated points based on $\hat{f}$ are evaluated (by the true objective function) and added to training set, this iterates until all λ offspring in this generation has been evaluated or the number of model iteration exceeds a certain number.

The performance is evaluated using a local weighted regression (LWR) [6] as the metamodel for lmm-CMA. Lmm-CMA achieves a speed-up (number of objective function evaluates for the other strategy divided by lmm-CMA) of 5-8 and 2-3 compared with standard CMA-ES [9] on convex quadratic Schwefel function and Rosenbrock function respectively. The performances of lmm-CMA for the above two convex quadratic function also matches the GPOP described above for dimension $N \leq 4$ and is even better for larger dimension $N \geq 8$ given a potential less reliable Gaussian Process Regression. On noisy sphere, lmm-CMA gains a small advantage over CMA-ES but the advantage vanishes with increasing dimensionality. Despite the speed-up achieved compared with CMA-ES, constructing the metamodel for lmm-CMA takes computational complexity up to N^6, so that this algorithm should also be considered for computational expensive problems.

3 ANALYSIS

To understand the potential implications of using surrogate model assisted $(\mu/\mu, \lambda)$-ES with varying population sizes, in this section, we use the same simple model proposed by Kayhani and Arnold [11] for the use of a surrogate model. Specifically, we propose an EA that generates a population of λ offspring in each generation where the offspring are evaluated using the surrogate model instead of the true objective function. An intermediate recombination is performed based on the inaccurate fitness of the offspring estimated by the surrogate model; the resulting centroid x_{centroid} of the μ selected offspring is evaluated using the true objective function. The analysis for $(\mu/\mu, \lambda)$-ES using simple model is the same as that of the $(\mu/\mu, \lambda)$-ES on noisy sphere — in offspring ranking, the former uses a surrogate model to estimate offspring's fitness (achieved at vanishing cost, by the assumption in Section 2.2), and the latter uses a noisy estimate of offspring's fitness where in both cases, the fitness estimation are achieved the same way. The normalized fitness gain of $x_{\text{centroid}}^{(g+1)}$ over $x_{\text{centroid}}^{(g)}$ is

$$\begin{aligned}\delta^*_{GP} &= \frac{N}{2R^2}\left(f(x_{\text{centroid}}^{(g+1)}) - f(x_{\text{centroid}}^{(g)})\right)\\ &= \delta^*(z_{\text{step}}^{(g+1)})\\ &\overset{N\to\infty}{=} \sigma^*(z_{\text{step},1} + \vartheta z_\epsilon) - \frac{\sigma^2}{2\mu}. \end{aligned} \tag{22}$$

The analysis in Section 2.1.3 still holds which gives the same result as Eq. (10). In this context, the noise-to-signal ratio ϑ can be interpreted as the measure of the surrogate model quality relative to the step size of the algorithm.

Since the fitness of λ offspring generated are evaluated by the surrogate model at vanishing cost by earlier assumption. The number of objective function evaluation per iterations is one instead of λ (for $(\mu/\mu, \lambda)$-ES without model assistance), therefore the normalized fitness gain when dimensionality $N \to \infty$, by substituting λ with 1 in Eq. (10) is

$$\eta = \frac{1}{1}E[\delta^*_{GP}] \approx \frac{\sigma^* c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2}} - \frac{(\sigma^*)^2}{2\mu}, \tag{23}$$

wonder if the normalized progress rate in Eqn. (10) should also be changed to normalized fitness gain(for consistency with ref?)

To obtain the opt. expected fitness gain η_{opt} and its corresponding opt. normalized step size σ^*_{opt} over a fixed noise-to-signal ratio ϑ, by assuming independence of ϑ and σ^*, we take the derivative of equation (23) over σ^* and obtain

$$\sigma^*_{opt} = \frac{\mu c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2}} \tag{24}$$

$$\eta_{opt} = \frac{\sigma^*_{opt} c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2}} - \frac{(\sigma^*_{opt})^2}{2\mu} \tag{25}$$

The expected fitness gain is normalized regarding the population size λ (i.e., $\eta_{\text{opt}}/\lambda$ is used) for easy comparison. The normalized fitness gain against the normalized step size for $(\mu/\mu, \lambda)$-ES with population size $\lambda = 10, 20, 40$ and corresponding $\mu = 3, 5, 10$ are plotted in Fig. 1 from left to right respectively. The lines show the population-normalized results obtained from Eqs. (13), (14), (23), and Eqs. (4), (6), $\eta = E[\Delta]/p_{\text{eval}}$ (from surrogate assisted (1+1)-ES [11]). The dots represent the experimental results for unbiased Gaussian surrogate error for $N \in \{10, 100\}$ obtained by averaging 100 runs.

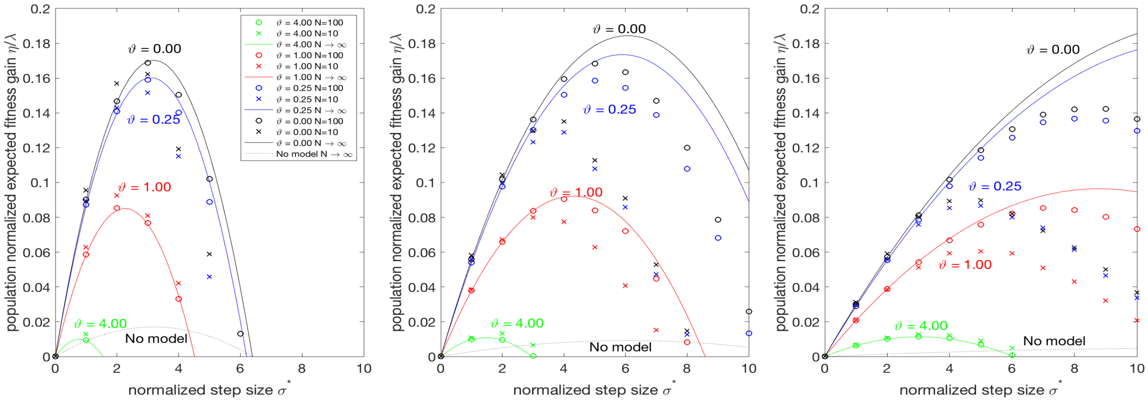

Fig. 1. The figures from left to right show the expected single step behaviour of the surrogate model assisted $(\mu/\mu, \lambda)$-ES with unbiased Gaussian distributed surrogate error for $\lambda = 10, 20, 40$ $(\mu = \lceil\lambda/4\rceil)$ respectively. The solid lines are the results obtained analytically when $N \longrightarrow \infty$; the dotted lines in the bottom of the figures show the relationship for corresponding $(\mu/\mu, \lambda)$-ES without surrogate model assistance (when $N \longrightarrow \infty$); the dots represent the experimental results for $N = 10$ (crosses) and $N = 100$ (circles).

It can be inferred from Fig. 1 that, for a fixed population size, the expected fitness gain decreases with an increasing noise-to-signal-ratio ϑ. When $\vartheta \longrightarrow \infty$, the surrogate model becomes useless, and the strategy becomes a random search. For moderate noise-to-signal ratio ϑ, the surrogate model assisted $(\mu/\mu, \lambda)$-ES can achieve much larger value for opt. expected fitness gain at a larger normalized step size compared with the surrogate model assisted (1+1)-ES. When $\vartheta = 1$, the maximal expected fitness gain achievable for (3/3, 10)-ES, (5/5, 20)-ES, and (10/10, 40)-ES are 0.8426, 1.841, and 3.856 with $\sigma^* = 2.258, 4.294$, and 8.769 respectively; for surrogate model assisted (1+1)-ES [11], the maximal fitness gain is 0.548 achieved at $\sigma^* = 1.905$. As for $(\mu/\mu, \lambda)$-ES and (1+1)-ES both without model assistance, the maximal fitness gains are 0.202, 0.170, 0.184, and 0.193 achieved at $\sigma^* = 1.224, 3.200, 6.080$, and 12.420 for $(1 + 1)$-ES, (3/3, 10)-ES, (5/5, 20)-ES, and (10/10, 40)-ES respectively. The value of maximal fitness achievable for $(\mu/\mu, \lambda)$-ES without surrogate model assistance grows as λ increases and gradually approaches 0.202, which asymptotically equals to that of the $(1 + 1)$-ES [4]. From the above analysis, surrogate assisted $(\mu/\mu, \lambda)$-ES does benefit from using a larger population, namely, the expected maximal fitness gain increased by a factor of λ after the surrogate model is applied; and an observable improvement in opt. fitness gain achieved at a larger normalized step size compared with the surrogate model assisted (1+1)-ES. For $\vartheta = 0$ (the surrogate model models the objective function exactly), from Eqn. (24) we can obtain the maximal expected fitness gain for surrogate assisted $(\mu/\mu, \lambda)$-ES 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 of using a growing population size, it is important to note the analytical result derived when $N \longrightarrow \infty$ 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 $N \longrightarrow \infty$ for $(\mu/\mu, \lambda)$-ES with $\lambda = 10, 20, 40$ and (1+1)-ES, all including cases with and without surrogate model assistance. The optimal expected fitness gain is also measured experimentally for $n \in \{10, 100\}$.

The speed-up is defined as the median number of objective function evaluations used to solve the test problems (when reaching the termination criteria) using one strategy divided by that of using the surrogate model assisted ES,

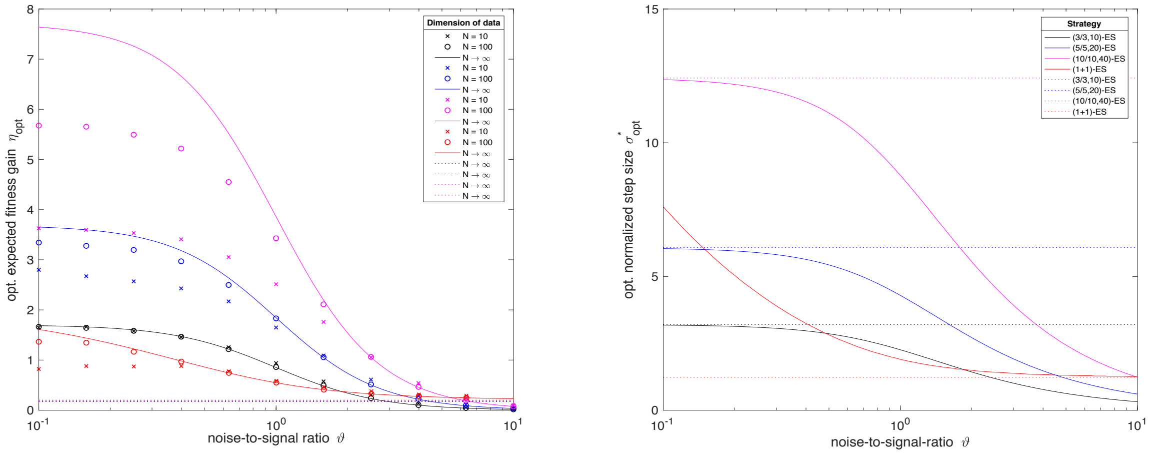

Fig. 2. Opt. expected fitness gain and corresponding opt. normalized step size of the surrogate model assisted $(\mu/\mu, \lambda)$-ES and (1+1)-ES plotted against the noise-to-signal ratio. Colour black, blue, magenta and red represent the result for (3/3, 10)-ES, (5/5, 20)-ES, (10/10, 40)-ES and (1+1)-ES respectively.The solid line represents the results obtained analytically when $N \longrightarrow \infty$. The dots are the experimental result for N = 10 (crosses) and N = 100. The dotted lines show the optimal values for the (3/3, 10)-ES, (5/5, 20)-ES, (10/10, 40)-ES and (1+1)-ES without surrogate model assistance.

Table 1. Speed-ups for a small noise-to-signal ratio ($\vartheta = 0.1$) observed in analysis

	Different Evolution Strategy used for comparison			
speed-up	(1+1)-ES	(3/3, 10)-ES	(5/5, 20)-ES	(10/10, 40)-ES
$\texttt{speed-up}_{\texttt{self}}(N = 10)$	4.1	9.9	15.2	18.8
$\texttt{speed-up}_{\texttt{self}}(N = 100)$	6.8	9.9	18.1	29.4
$\texttt{speed-up}_{\texttt{self}}(N \longrightarrow \infty)$	8.0	10.0	20.0	40.0
$\texttt{speed-up}_{\texttt{model}}(N = 10)$		2.4	3.7	4.6
$\texttt{speed-up}_{\texttt{model}}(N = 100)$		1.5	2.7	4.4

both using the opt. normalized step size. For comparison purposes, we define two speed-ups, namely speed-up$_{\text{self}}$ and speed-up$_{\text{model}}$. speed-up$_{\text{self}}$ is defined as the number of objective function calls used for a surrogate model assisted ES divided by that of the ES without surrogate model assistance (specifically $(\mu/\mu, \lambda)$-ES if the ES is not specified); and speed-up$_{\text{model}}$ as the number of objective function calls needed for (1+1)-ES to solve the problem divided by that of $(\mu/\mu, \lambda)$-ES, both with surrogate model assistance. The results obtained are shown in Table 1.

For a finite-dimension e.g. $N = 10$, the speed-up$_{\text{self}}$ achieved for small noise-to-signal ratio ($\vartheta = 0.1$) appears to be around nine and ten for a population size $\lambda = 10$, fifteen and sixteen for $\lambda = 20$, eighteen and nineteen for $\lambda = 40$; whereas for (1+1)-ES, the speed-up$_{\text{self}}$ tops out between four and five. When $N = 100$, the speed-up$_{\text{self}}$ for small noise-to-signal ratio are between nine and ten, eighteen and nineteen, and twenty-nine and thirty for (3/3, 10)-ES, (5/5, 20)-ES, and (5/5, 40)-ES respectively; as for (1+1)-ES, the speed-up$_{\text{self}}$ with a small noise-to-signal ratio is between six and seven when $N = 100$. The speed-up$_{\text{self}}$ (for $(\mu/\mu, \lambda)$-ES) mounts with a growing population size λ, but it is worth noting that speed-up$_{\text{self}}$ $\overset{N\to\infty}{\longrightarrow}$ λ, and the gap in speed-up$_{\text{self}}$ between a finite dimension (fixed) and infinite dimension increases as λ grows (e.g., compared with $N \to \infty$, the speed-up$_{\text{self}}$ achieved when $N = 10$ decreases 1%, 24%, and 53% for $\lambda = 10, 20$, and 40 respectively).

When comparing (1+1)-ES and $(\mu/\mu, \lambda)$-ES both with surrogate model assistance, speed-up$_{\text{model}}$ for $(3/3, 10)$-ES, $(5/5, 20)$-ES, and $(5/5, 40)$-ES when $N = 10$ are between two and three, three and four, four and five respectively; whereas $N = 100$, the speed-up$_{\text{model}}$ decrease to a value between one and two, two and three, and four to five respective for $\lambda = 10, 20$, and 40. From the observed result, the speed-up for surrogate model assisted $(\mu/\mu, \lambda)$-ES is significant over both the surrogate model assisted (1+1)-ES and $(\mu/\mu, \lambda)$-ES without model assistance. It seems that the expected fitness gain for the surrogate assisted $(\mu/\mu, \lambda)$-ES will increase in line with population size λ, especially with a large dimensionality. Besides, the shrinking gap in speed-up$_{\text{model}}$ between different finite dimension over a growing population (i.e., the speed-up$_{\text{model}}$ for $N = 100$ approaches that of $N = 10$ as λ grows) also suggests using a larger λ for potential stable speed-up$_{\text{model}}$ compared with the surrogate assisted (1+1)-ES. Although the analytical results obtained for $N \to \infty$ are approximations in the finite-dimension and can be inaccurate but it gives an implication for using larger λ for potential larger fitness gain.

4 STEP SIZE ADAPTATION

4.1 Cumulative step size adaptation

Algorithm 2 Surrogate Model Assisted $(\mu/\mu, \lambda)$-ES (GP-mml-ES)

```
$c \leftarrow \frac{\mu+2}{N+\mu+5}$
$d \leftarrow 1 + 2\max(0, \sqrt{\frac{\mu-1}{N+1}} - 1) + c$
$E\,\|\mathcal{N}(0, I)\| \approx \sqrt{N}(1 - \frac{1}{4N} + \frac{1}{21N^2})$
$s^{(0)} \leftarrow 0$
while not terminate() do
    for $i = 1, 2, ..., \lambda$ do
        Generate standard normally distributed $z_i^{(g)} \in \mathbb{R}^N$
        Evaluate $x_{\text{centroid}}^{(g)} + \sigma^{(g)} z_i^{(g)}$ using the surrogate model, yielding $f_\epsilon(x_{\text{centroid}}^{(g)} + \sigma^{(g)} z_i)$
    end for
    $z_{\text{step}}^{(g)} \leftarrow \frac{1}{\mu} \sum_{i=1}^{\mu} z_{i;\lambda}^{(g)}$
    $x_{\text{centroid}}^{(g+1)} \leftarrow x_{\text{centroid}}^{(g)} + \sigma^{(g)} z_{\text{step}}^{(g)}$
    Evaluate $x_{\text{centroid}}^{(g+1)}$ using true objective function, yieding $f(x_{\text{centroid}}^{(g+1)})$
    Update surrogate model by adding $(x_{\text{centroid}}^{(g+1)}, f(x_{\text{centroid}}^{(g+1)}))$
    $s^{(g)} \leftarrow (1 - c)s + \sqrt{c(2 - c)\mu} z_{\text{step}}^{(g)}$
    $\sigma^{(g+1)} \leftarrow \sigma^{(g)} \exp\left(\frac{c}{d} \frac{\|s^{(g)}\|}{E\|\mathcal{N}(0, I)\|} - 1\right)$
    $g \leftarrow g + 1$
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 that 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 CSA described in Section 2.1.1 and 2.1.2, and exploit the potential insight it may offer. The strategy is evaluated

Table 2. Median test results for $(\mu/\mu,\ \lambda)$-ES without surrogate model assistance.

	Median number of objective function calls		
Test functions	(3/3, 10)-ES	(5/5, 20)-ES	(10/10, 40)-ES
linear sphere	3300	4809	8405
quadratic sphere	1694	2436	4182
cubic sphere	1166	1659	2788
Schwefel' s function	6259	8064	13325
quartic function	6600	8442	14637

Table 3. Median test results and speed-ups of GP-mml-ES.

	Median number of objective function calls (speed-up$_{\text{self}}$, speed-up$_{\text{model}}$)			
Test functions	(1 + 1)-ES	(3/3, 10)-ES	(5/5, 20)-ES	(10/10, 40)-ES
linear sphere	502	756(4.4, 0.66)	696(6.9, 0.72)	761(11.0, 0.66)
quadratic sphere	214	309(5.5, 0.69)	245(9.9, 0.87)	231(18.1, 0.93)
cubic sphere	205	273(4.3, 0.75)	249(6.7, 0.82)	253(6.7, 0.81)
Schwefel' s function	1503	2278(2.7, 0.66)	$+\infty(/)$	$+\infty(/)$
quartic function	1265	1000(6.6, 1.3)	746(11.3, 1.7)	667(21.9, 1.9)

using a Gaussian Process surrogate model in place of the simple model described in Section 3; one single iteration of the surrogate model assisted $(\mu/\mu, \lambda)$-ES using CSA (referred to as GP-mml-ES) is shown in Alg. 2 where $z_{i;\lambda}^{(g)}$ denotes the ith best mutation ranked according to the fitness of the corresponding offspring estimated by the surrogate model that has $f_\epsilon(x_{\text{centroid}}^{(g)} + \sigma^{(g)} z_{i;\lambda}^{(g)}) < f_\epsilon(x_{\text{centroid}}^{(g)} + \sigma^{(g)} z_{j;\lambda}^{(g)}), 1 \leq i < j \leq \lambda$.

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 [21]); and quartic function [11] 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 with the condition number of the Hessian at the optimizer exceeds 3,500, making it very hard to find the global optimum without adapting the shape of the mutation distribution. So we use the quartic function with $\beta = 1$ in the context where the corresponding condition number of the Hessian at the optimizer equals to 49.0. The values of global optima for all test function are zero. For each test problem, 100 runs are conducted for surrogate assisted $(\mu/\mu, \lambda)$-ES using $\lambda = 10, 20, 40$ and corresponding $\mu = \lceil \lambda/4 \rceil$ both with and without surrogate model assistance where a parental population size $\lambda = 10, 20, 40$ with are used. $(1 + 1)$-ES and

For surrogate model, we use Gaussian process with squared exponential kernel and the length scale parameter in the kernel is set proportional to both the step size of the ES and the square root of data dimension N. For simplicity, the length scale is set to $8\sigma\sqrt{N}$ as is used in the surrogate assisted (1+1)-ES [11]. In our experience, using a smaller or larger value by a factor of two i.e., 2, 4 and 16, 32 does not give a significant improvement, so we stick to 8 in the context.

The Gaussian process kernel is constructed using a training size of 40. The training set consists of the 40 most recent candidate solutions evaluated, so that the surrogate model approximates the local landscape of the objective function.

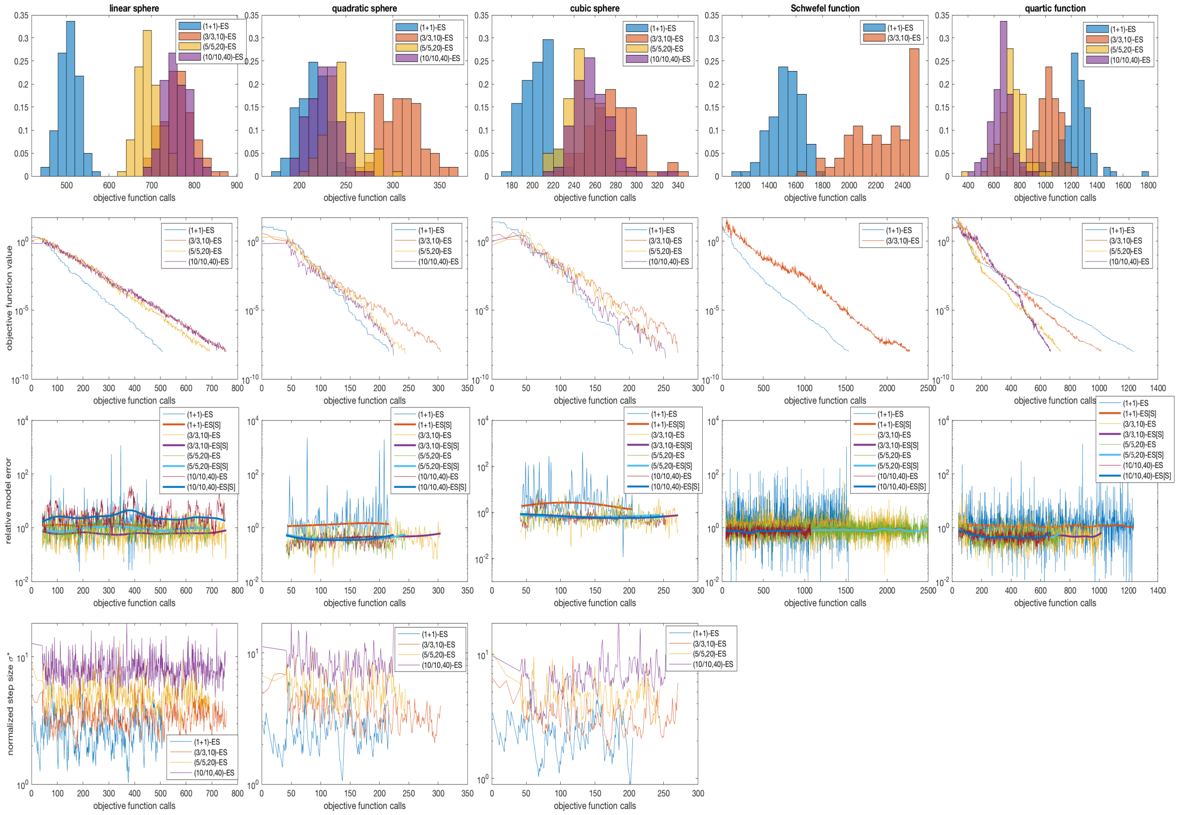

Fig. 3. Results for GP-mml-ES (with GP and CSA applied). Top row: Histograms showing the number of objective function calls needed to solve the five test problems. Second row: Convergence graphs observed in median runs. Third row: Relative model error obtained in median runs ([S] denotes the smoothed relative model error). Last row: Normalized step size measured in median runs for sphere functions.

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}.

Histograms 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 number of objective function calls required to solve each test problem for $(\mu/\mu, \lambda)$-ES without and with surrogate model assistance (the step size of both are adapted using CSA) is shown in Table 2 and Table 3 respectively; the result for surrogate assisted (1+1)-ES [11] is also included in Table 3 for comparison.

The two speed-ups speed-up$_{\text{self}}$ and speed-up$_{\text{model}}$ defined in Section 3 Analysis are used. The speed-up$_{\text{self}}$ on quadratic sphere for GP-mml-ES with $\lambda = 10, 20, 40 (\mu = \lceil \lambda/4 \rceil)$ are between five and six, nine and ten, eighteen and nineteen respectively; whereas the corresponding expected results obtained in Analysis are between nine and ten, fifteen and sixteen, eighteen and nineteen respectively. There is a gap between the expected performance (in Section 3) and the experimental results (after the GP surrogate model and CSA is in place) for GP-mml-ES. But the difference between expected values and observed values obtained through experiments seems to shrink as the population size λ grows. In terms of speed-up$_{\text{model}}$, the achieved values for GP-mml-ES on quadratic sphere are about 0.7, 0.9, and 0.9 for $\lambda = 10, 20, 40$ respectively, which is far from the expected maximal value of roughly 2.4 and 3.7 and 4.6 recorded in Table 1. There is about a factor of four in terms of the expected maximal speed-up$_{\text{model}}$ and the results obtained in table 3. The speed-up$_{\text{self}}$ observed for sphere functions are between four and five for GP-mml-ES with $\lambda = 10$, six and ten for GP-mml-ES with $\lambda = 20$, and six to nineteen for GP-mml-ES with $\lambda = 40$. It is interesting that only GP-mml-ES using $\lambda = 10$ solves the 10-dimensional Schwefel's function, the increasing population does not give an advantage for solving the problem comped with the (1+1)-ES with surrogate model assistance, but rather a trend for divergence. Overall, the performance of GP-mml-ES does not match that of the surrogate model assisted (1+1)-ES, specifically, speed-up$_{\text{model}}$ < 1 for all test functions except quartic function where the observed speed-up$_{\text{model}}$ are about 1.3, 1.7, and 1.9 for population size $\lambda = 10, 20, 40$ respectively. It seems a growing population can possibly bring larger speed-up$_{\text{model}}$ for quadric sphere and quartic function, meanwhile it appears contrary to the other three test functions, especially the Schwefel's function. The speed-up$_{\text{self}}$ observed in linear and cubic sphere is lower than that in quadratic sphere, indicating a potential less accurate Gaussian process based surrogate model in the two sphere functions mentioned compared with the latter.

The second row of Fig. 3 shows the convergence graphs observed in median runs. Linear convergence are achieved for all but the Schwefel's function, specifically the using GP-mml-ES with $\lambda = 20, 40$. Interestingly, using a larger population on Schwefel's function does not help achieve faster convergence compared to (1+1)-ES, but instead, slows the rate of convergence, and finally makes the strategy diverge. Relative model error for the median runs is shown in the third row of the figure. The relative model error in generation (g) is defined as

$$\begin{cases} \frac{\|f(y^{(g)})-f_\epsilon(y^{(g)})\|}{\|f(y)-f_{(}x)\|}, \text{for surrogate model assisted (1+1)-ES} \\ \sqrt{\frac{var(fY^{(g)}-fY_\epsilon^{(g)})}{var(fY^{(g)}-fX^{(g)})}}, \text{for surrogate model assisted } (\mu/\mu,\lambda)\text{-ES,} \end{cases} \tag{26}$$

where $x^{(g)}$ and $y^{(g)}$ are the parent and offspring in generation (g) for surrogate model assisted (1+1)-ES from Kayhani and Arnold [11]; $fY^{(g)} - fX^{(g)} = \{f(x_{\text{centroid}}^{(g)} + \sigma^{(g)} z_{\text{step}}^{(g)}) - f(x_{\text{centroid}}^{(g)}) : 1 \leq i \leq \lambda\}$, $fY^{(g)} - fY_\epsilon^{(g)} = \{f(x_{\text{centroid}}^{(g)} + \sigma^{(g)} z_{\text{step}}^{(g)}) - f_\epsilon(x_{\text{centroid}}^{(g)} + \sigma^{(g)} z_{\text{step}}^{(g)}) : 1 \leq i \leq \lambda\}$, and $var(fY^{(g)} - fY_\epsilon^{(g)})$ returns the variance of the λ differences between the true and estimated objective function values for the offspring generated in generation (g), and by taking square root over the division of two variances we can obtain the standard deviation similar to the case of surrogate model assisted (1+1)-ES. For easy comparison, 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.), interpreted as the a relative constant noise-to-signal ratio. The relative model error for all surrogate assisted ES in observed is approximately 1; but we notice the relative model error for GP-mml-ES is much smaller than that of the surrogate assisted (1+1)-ES, so is the variance of the relative model error. This is possibly the benefit of using a large population size. In linear sphere, the relative model error obtained by using GP-mml-ES with $\lambda = 40$ is significantly higher compared with the rest strategies, which may explain the decline in speed-up$_{\text{model}}$ from 0.72 to 0.66 after λ increased from 20 to 40. In the other problems, the relative model error for GP-mml-ES with different λ seem to stay at the same level. In the quadratic sphere, the relative constant noise-to-signal ratio is approximately 0.7 for GP-mml-ES with different λ; and according to Fig. 2, when $\vartheta = 0.7$, the opt. step size are between two and three, four and five, ten and eleven for $\lambda = 10, 20, 40$ respectively. The corresponding expected speed-up$_{\text{model}}$ for $\lambda = 10, 20, 40$ obtained by

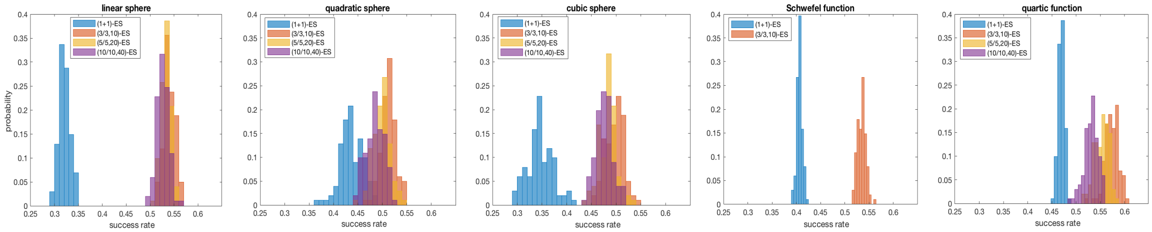

Fig. 4. Histograms of the success rate (proportion of good steps in a single run) for each test function using (1+1)-ES and GP-mml-ES with $\lambda = 10, 20, 40$.

using the opt. expected fitness gain from Fig. 2 when $\vartheta = 0.7$ are about 1.7, 3.1, and 4.3 for $\lambda = 10, 20, 40$ respectively. The bottom row in Fig. 3 shows the normalized step size over the number of objective function calls observed in the median runs in the three sphere functions tested. The GP-mml-ES achieves a larger normalized step size σ^* compared to the surrogate model assisted (1+1)-ES with σ^* growing in line with λ, which coincides with our knowledge that using a larger population size in $(\mu/\mu, \lambda)$-ES can give a larger step size i.e., in the extreme case when $\lambda \longrightarrow \infty$, we obtain the gradient at the current point (parent x). Given the relative constant noise-to-signal is 0.7 (i.e., $\vartheta = 0.7$) for quadratic sphere, we can obtain the opt. expected normalized step size σ^* from Fig. 2 and compare that with the experimental result obtained in the bottom row of Fig. 3. The obtained σ^* in experiment for each GP-mml-ES with different λ is very close to that of the analytical result shown in Fig. 2, where the former ranges from two to four, four to six, nine to eleven for $\lambda = 10, 20, 40$ respectively, and the latter has discussed above. The results indicates that the step size of the GP-mml-ES is appropriately adapted by CSA.

Despite the potentially well-adapted step size, there is a big gap between the expected speed-up$_{\text{model}}$ discussed above and the experimental result illustrated in Table 3. The reasons behind the resulting experimental results are not yet clear. Assuming the step size is well-adapted, the natural question to ask is what is the quality of the direction (z_{step}) used. If the resulting direction is poor (i.e., taking the direction cannot improve the fitness), the individual cannot gain any fitness even with an optimally adapted step size. On the other hand, taking a good step with a properly adapted step size can help generate better offspring (offspring with larger fitness gain compared with its parent). We further plot the histogram of success rate (the proportion of a good step size in each run) for all test problems in Fig. 4 by replicating 100 runs for each test problems. Overall, the success rates for GP-mml-ES are higher compared with the surrogate model assisted (1+1)-ES, where the latter achieves about 0.32, 0.42, 0.35, and 0.4 for linear sphere, quadratic sphere, cubic sphere, Schewefel's function, and quartic function respectively. Whereas the success rates for GP-mml-ES with $\lambda = 10, 20, 40$ are at about the same level in all five test problems, specifically, ranging form 0.45 to 0.55 in three sphere functions, about 0.53 in Schwefel's function, and between 0.5 and 0.6 in quartic function. Recombination and selection over a larger population size in $(\mu/\mu, \lambda)$-ES do help improve the chance of making a good step compared with the $(1 + 1) - ES$. It is interesting that the success rate of an increasing λ first increases and then decreases. The benefit of using a larger step size in terms of sampling a good direction in the context of a GP surrogate model is not yet clear. The overall success rate of GP-mml-ES approximately 0.5, meaning the strategy makes a bad step every other step.

4.2 Safeguard of plus-selection

Algorithm 3 Surrogate model assisted ES with plus-selection (GP-cross-ES)

```
$c \leftarrow \frac{\mu+2}{N+\mu+5}$
$d \leftarrow 1 + 2\max(0, \sqrt{\frac{\mu-1}{N+1}} - 1) + c$
$E\|\mathcal{N}(0,I)\| \approx \sqrt{N}(1 - \frac{1}{4N} + \frac{1}{21N^2})$
$s^{(0)} \leftarrow 0$
$D \leftarrow 0.72$
**while** not terminate() **do**
    **for** $i = 1, 2, ..., \lambda$ **do**
        Generate standard normally distributed $z_i^{(g)} \in \mathbb{R}^N$
        Evaluate $x_{\text{centroid}} + \sigma^{(g)} z_i^{(g)}$ using the GP surrogate model, yielding $f_\epsilon(x_{\text{centroid}} + \sigma^{(g)} z_i^{(g)})$
    **end for**
    $z_{\text{step}}^{(g)} \leftarrow \frac{1}{\mu}\sum_{i=1}^{\mu} z_{i;\lambda}^{(g)}$
    $y \leftarrow x_{\text{centroid}} + \sigma^{(g)} z_{\text{step}}^{(g)}$
    Evaluate $y$ using true objective function, yielding $f(y)$
    Update surrogate model by adding $(y, f(y))$
    **if** $f(x_{\text{centroid}}) < f(y)$ (occurrence of a bad step) **then**
        $\sigma \leftarrow \sigma D$
    **else**
        $x_{\text{centroid}} \leftarrow y$
        $s^{(g)} \leftarrow (1-c)s^{(g)} + \sqrt{c(2-c)\mu} z_{\text{step}}^{(g)}$
        $\sigma^{(g+1)} \leftarrow \sigma^{(g)} \exp\left(\frac{c}{d} \frac{\|s^{(g)}\|}{E\|\mathcal{N}(0,I)\|} - 1\right)$
    **end if**
    $g \leftarrow g + 1$
**end while**
```

The GP-mml-ES gives a success rate approximately equals to 0.5 for all population sizes used in all test functions. It comes natural to ask, how much we are to benefit if we can avoid or simply reject those bad steps. A recent paper in surrogate model assisted ES considers (1+1)-ES [11] where the step size of the strategy is successfully adapted based on the success rate of a good step size. The mentioned step size adaptation mechanism considers decreasing the step size in two cases. In the first case, the estimated fitness of the offspring (using the surrogate model) is inferior to the true fitness of its parent; as for the second case, the true fitness of the offspring evaluated is inferior to that of its parent. Using a similar idea, we propose a surrogate model assisted ES that is a cross between (1+1)-ES and $(\mu/\mu, \lambda)$-ES (referred to as GP-coress-ES) and a new step size adaptation mechanism using CSA and the safeguard of plus-selection.

The safeguard of plus-selection is used to reject the poor steps (resulted from a bad direction that cannot improve the fitness of the current parent). If the offspring is inferior to its parent, meaning the step generated in this generation is bad, the safeguard of plus-selection simply rejects the inferior offspring. The proposed GP-cross-ES and the step size adaptation are shown in Alg. 3: in each generation, one offspring $y = x_{\text{centroid}} + \sigma^{(g)} z_{\text{step}}^{(g)}$ is evaluated using the true objective function and added to the training set to update GP; the true fitness of the offspring $f(y = x_{\text{centroid}} + \sigma^{(g)} z_{\text{step}}^{(g)})$ is compared to that of its parent $f(x_{\text{centroid}})$. If the fitness of the offspring is inferior to its parent, indicating the resulted step $z_{\text{step}}^{(g)}$ is poor in this generation, the offspring y is discarded and the step size is decreased by a factor of D using an idea similar to the step size adaptation of the surrogate model assisted (1+1)-ES [11]; the bad step information is not

Table 4. Median test results and speed-ups of GP-cross-ES.

	Median number of objective function calls (speed-up$_{\text{self}}$, speed-up$_{\text{model}}$)			
Test functions	$(1+1)$-ES	$(3/3, 10)$-ES	$(5/5, 20)$-ES	$(10/10, 40)$-ES
`linear sphere`	502	367(9.0, 1.4)	316(15.1, 1.6)	321(26.3, 1.6)
`quadratic sphere`	212	211(8.0, 1.0)	164(14.9, 1.3)	146(29.0, 1.5)
`cubic sphere`	202	213(5.5, 0.96)	176(9.4, 1.3)	178(15.8, 1.4)
`Schwefel's function`	1503	2070(3.0, 0.73)	1355(6.0, 1.1)	1051(12.7, 1.4)
`quartic function`	1250	1511(4.4, 0.83)	1016(8.3, 1.3)	796(18.4, 1.6)

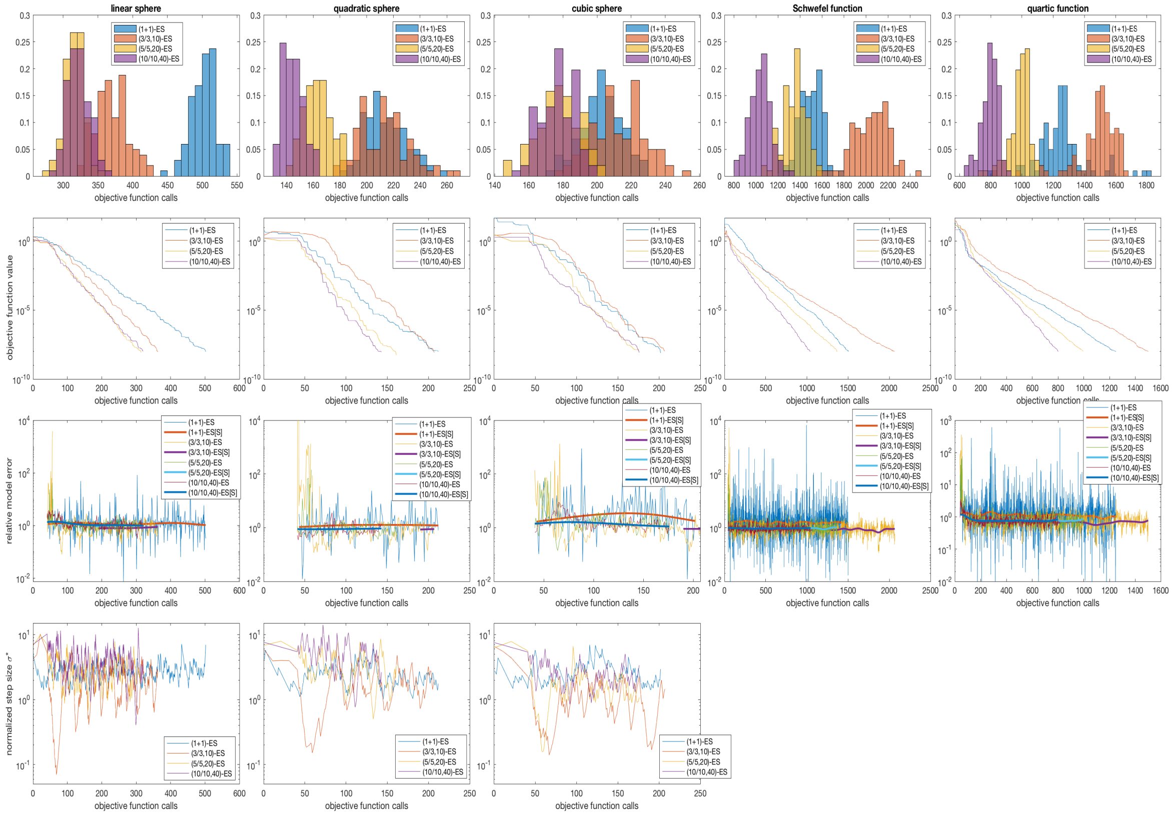

Fig. 5. Results of GP-cross-ES obtained by applying the safeguard of plus-selection . Top row: Histogram showing the number of objective function calls needed to solve the five test problems. Second row: Convergence graphs observed in median runs. Third row: Relative model error obtained in median runs ([S] denotes the smoothed plot). Last row: normalized step size measured in median runs for sphere functions.

added to the evolution path $s^{(g)}$ because we want to build an evolution path that is based on the good step information of previous generations. Otherwise (the step made is good), we update the parent by setting $X_{\text{centroid}} = y$ and update the step size accordingly using CSA.

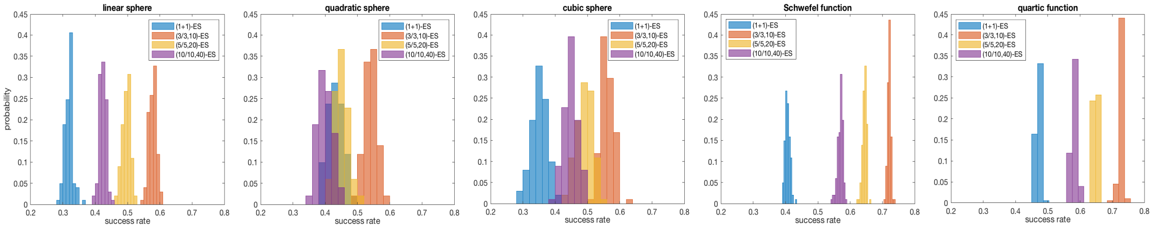

Fig. 6. Result obtained by applying the safeguard of plus-selection. The first two rows show the normalized convergence rate for each run plotted in histogram and normalized probability density function (pdf) respectively. The last two rows represent the success rate (proportion of good step size in each run) plotted in histogram and pdf respectively.

To test the proposed GP-cross-ES and the step size adaptation mechanism, we use the same test functions and generate the corresponding Fig.s in Section 4.1. The number of objective function evaluations used in median runs and the corresponding speed-ups are reported in Table 4. The overall performance of GP-cross-ES improves as the population size increases, and outperforms the surrogate model assisted (1+1)-ES [11] in all test functions after $\lambda \geq 20$. The speed-up$_{\text{model}}$ for all three population sizes tested is around 1.5 for the linear sphere, between 1 and 1.5 for both the quadratic sphere and cubic sphere, about 0.7 to 1.4 for the Schwefel's function, and between 0.8 and 1.6 for the quartic function. The linear sphere does not seem to benefit much from a growing population size and the speed-up$_{\text{model}}$ seems to plateau after $\lambda \geq 20$; whereas in the quadratic sphere, Schwefel's function, and quartic function, the speed-up$_{\text{model}}$ does not show a clear sign of plateau. This observation also coincides with the speed-up$_{\text{self}}$ observed on the quadratic sphere, Schwefel's function, and quartic function where the growth in speed-up is proportional to the growth in population size. For example, in the Schwefel's function, the speed-up$_{\text{self}}$ doubles form 3.0 to 6.0 then approximately doubles to 12.7 after the population size doubles from 10 to 20 and finally 40. The histograms of objective function calls that are needed to solve each test problem (first row of Fig. 5) also illustrate the decrease in objective function calls needed to solve each test problem as the population size grows. Convergence graphs for the medium run are shown in the second row of Fig. 5. Linear convergence appears to be achieved in all runs. The third row reports the relative model error defined in Section 4.1. It is interesting that the relative model error for proposed GP-cross-ES (in any case of λ being used) is still lower than that of the surrogate model assisted (1+1)-ES. One possible explanation for a larger variance in the relative model error observed for the surrogate model assisted (1+1)-ES (compared with the GP-mml-ES and GP-cross-ES) is the small population size (only a single offspring is generated), whereas the GP-mml-ES and GP-cross-ES generate λ offspring in each generation. The bottom row in Fig. 5 shows the normalized step size observed in median runs for the three sphere functions used. The normalized step size σ^* in all runs are approximately between 0.5 and 10, where the strategy using plus-selection loses the benefit of a potential larger step size as opposed to the comma-selection (e.g., the σ^* for the GP-mml-ES is significantly higher than that of the GP-cross-ES). Another observation is the decreasing trend in the variance of relative model error with a growing λ for GP-cross-ES, which indicates the possibility of a potential good step resulting from using a large population size in selection and recombination. An illustration of the continuous occurrence of bad steps that can result in a continuous decrease of σ^* is shown in the first 80 objective function calls in linear sphere for (3/3,10)-ES (GP-cross-ES with $\lambda = 10, \mu = 3$). But the occurrence or the length (the number of objective functions such situations take up) of such situations is reduced as λ increases.

Similarly, the histograms of success rates (the proportion of a good step size in each run) for all test problems are plotted in Fig. 6 by replicating 100 runs for each test problem. The success rate for GP-cross-ES is overall larger than that of the surrogate model assisted (1+1)-ES with the exception of the quadratic sphere in which the success rate of GP-cross-ES with $\lambda = 10$ and 20 being the same level as surrogate model assisted (1+1)-ES. For all other test functions, the success rate for GP-cross-ES decreases as λ increases. In both the linear sphere and cubic sphere, the success rate for GP-cross-ES ranges from 0.4 to 0.5, 0.45 to 0.55, and 0.55 to 0.65 for $\lambda = 10, 20,$ and 40 respectively; whereas surrogate model assisted (1+1)-ES being about 0.3 to 0.35. The success rate observed in the Schwefel's function and quartic function are between 0.55 and 0.61; 0.62 and 0.65; and 0.70 and 0.75 for GP-cross-ES with $\lambda = 10, 20,$ and 40 respectively, and ranges from 0.55 to 0.62 for surrogate model assisted (1+1)-ES. Even if the success rate reduces with a lager λ for GP-cross-ES, from the sharper slope, compared with GP-mml-ES, in the convergence plots, we can infer that the step is regarded as "good" by the GP-cross-ES has a potential better quality than that of GP-mml-ES. That is, the fitness gain obtained by taking the step is larger than that of the GP-mml-ES. By using a larger λ, the GP-cross-ES with proposed step-size adaption mechanism takes "good" steps that can bring potential larger fitness gain compared with the GP-mml-ES. One possible explanation of the better performance is the quality increase of taking the good steps outweighs the reduce in probability regarding the occurrence of a good step .

5 CONCLUSIONS

In this thesis, we apply the simple model [11] of unbiased Gaussian distributed noise for surrogate modelling approach. By using this approach, we analyzed the behaviour of surrogate model assisted $(\mu/\mu, \lambda)$-ES on the quadratic sphere function. Based on the analysis and the observations of the GP-mml-ES, we proposed GP-cross-ES and a corresponding step size adaptation mechanism. 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.

In future work, we will study the behaviour of surrogate assisted CMA-ES using the same analysis that can potentially handle ill-conditioned problems. Further goals include length scale adaptation mechanism in the Gaussian Process and surrogate model accuracy control that can possibly further reduce the gap between the expected analytical results and the experimental results.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my parents. I would also like to share my my supervisor Dr. Dirk V. Arnold for the continuous support during the past two semesters, for his patience, motivation, and passion that motivates me all the time both in study and beyond.

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] Hans-Georg Beyer. 1995. Toward a theory of evolution strategies: on the benefits of sex—the (mu/ mu, lambda) theory. *Evol. Comput.*, 3, 1, (Mar. 1995), 81–111. http://dx.doi.org/10.1162/evco.1995.3.1.81.

[5] H.-G. Beyer. 2013. *The theory of evolution strategies.* Springer Science & Business Media.

[6] A. W. Moore C. G. Atkeson and S. Schaal. 1997. Locally weighted learning for control. In *Lazy learning*. Springer, 75–113.

[7] N. N. Schraudolph D. Buche 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.

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

[9] N. Hansen and S. Kern. 2004. Evaluating the cma evolution strategy on multimodal test functions. In *International Conference on Parallel Problem Solving from Nature.* Springer, 282–291.

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

[11] 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.

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

[13] K. P. Murphy. 2012. *Machine Learning: A Probabilistic Perspective*. The MIT Press.

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

[15] 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.

[16] C. E. Rasmussen. 2004. Gaussian processes in machine learning. In *Advanced lectures on machine learning*. Springer, 63–71.

[17] 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.

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

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

[20] N. Hansen S. Kern 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.

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