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

Jingyun Yang and Dirk V. Arnold
Faculty of Computer Science, Dalhousie University
Halifax, Nova Scotia B3H 4R2
jingyun.yang@dal.ca,dirk@cs.dal.ca

ABSTRACT

Surrogate models have been widely used to assist evolutionary algorithms (EAs) to avoid unnecessary objective function evaluations. The cost is reduced by substituting the true objective function evaluation with a cheap but inaccurate estimate using the surrogate model. The surrogate model is built on th knowledge gained in previous iterations. Using surrogate assised (1+1)-ES for simple model and single steps have been studied, but the effect of actual inferior parent resulted from an inaccurate surrogate estimation and a corresponding poor step size are not well understood. We study the behaviour using a surrogate model assisted $(\mu/\mu, \lambda)$-ES using a population instead of a single offspring that. By comparing the bahaviour of the two, we propose a step size adaptation mechinisum and systematically evaluate the strategy for several test functiions.

KEYWORDS

$(\mu/\mu, \lambda)$-ES, Surrogate Model, Evolutionary algorithms(EAs), Gaussian Process

ACM Reference format:
Jingyun Yang and Dirk V. Arnold. 2018. Design of a Surrogate Model Assisted $(\mu/\mu, \lambda)$-ES. In *Proceedings of , , ,* 4 pages.
https://doi.org/

1 INTRODUCTION

Evolution strategies (ESs) have been widely utilized to solve optimization problems with an exceptional good performance in continuous solution space [11]. While in the case where objective function evaluation is costly, the time needed to compute the fitness of each offspring can take a considerable amount of time. One remedy is to use the surrogate model, a substitution of the true objective function built based on the information of previous candidate solutions that are evaluated by the true objective function. The surrogate model can give a more computationally efficient inaccurate estimate of the offsprings' fitness, replacing the expensive true objective function evaluation. The estimate using surrogate model yields insight into potential avoidance of poor step size. The optimization of the candidate solutions in each iteration also gives a better convergence rate.

, ,

ACM ISBN .
https://doi.org/

There are a range of surrogate models and a survey of the development can be found by Jin [8] and Loshchilov [12]. The surrogate model can do a pre-selection of the candidate solutions generated in each iteration with no computational cost. The effectiveness of surrogate assisted model is that it can effectively filter those undesired candidate solutions (reduce the population size) so that the unnecessary objective function evaluations can be avoided. It is the reduction of the population size by filtering undesired candidate solutions rather than the surrogate mode that saves the computational cost directly [10].

By using a similar idea to surrogated model assisted EAs, we do not use a local surrogate model to do the pre-selection in each iteration. Instead, we use it to optimize the candidate solutions obtained in each iteration by fitting a quadratic model in one dimension. By taking the lowest point of the quadratic model as the offspring of this iteration, we hope the fitness of the offspring obtained by the quadratic model would be better compared with the candidate solution directly generated from the parent. For better interpretation, we study the behavior of the proposed surrogate model assisted EAs using simple test functions, allowing comparisons with established basslines. The contributions of this paper are as follows: in Section 2, we give a brief review of related background, in Section 3, we propose a local surrogate model assisted (1+1)-ES, in Section 4, the model is evaluated on quadratic sphere following the same framework proposed by Kayhani and Arnold [10].

2 RELATED WORK

Using an approximate model to reduce computational cost can be traced back to 1960s [5]. Some successful surrogated models include but are not limitted to Polynomial Regression (PR, response surface methodology) [7], Gaussian Process (GP, Kriging models) [15], Artificial neural networks [16]. There are two types of surrogate models, global surrogate model and local surrogate model. ES using global surrogate model based on Kring was examined by Ratle [13]. Another ES using global surrogate model based on Artificial neural networks was constructed by Jin [9] which gives an imperial criterion on using the true objective function or the surrogate model to evaluate the offspring. Ulmer et al [17] and Buche et al [3] also applied GP as surrogate models in ES. But the performance of global surrogate models degrade as the dimension of the data increases, known as *curse of dimensionality.* Online local surrogate models [19] can be constructed using methods like radial basis function (RBF) [6] to replace the global surrogate model, where the surrogate model is updated online, giving a more accurate estimation compared with the global surrogate model.

Most recent works in surrogated assisted EAs uses a combination of different surrogate models to estimate the fitness strength of the candidate solutions. Zhou et al [19] proposed a hierarchical surrogate-assisted ES where a global surrogate model and a local surrogate model are integrated. The Global surrogate model uses GP and PR to estimate the global fitness of ES's search space, filtering the unpromising candidate solutions. Then, a local surrogate-assisted Lamarckian learning based on RBF is performed to search the promising candidate solutions.

There are various surrogate-assisted EAs integrating global and local surrogate models or using a combination of heuristics. These methods tend to be sophisticated for good performance, while few literatures have investigated the surrogated-assisted 1+1-ES. One exception is what Chen and Zou [4] proposed but yet incomplete in terms of two aspects. Firstly, it uses a linear surrogate that cannot give a precise estimate when coordinate transform is applied, the precondition to solve a generalized optimization problem [10]. Secondly, it does not include a step size adaptation mechanism. Besides that, Ulmer et al [18] proposed a Model Assisted Steady-State Evolution Strategy (MASS-ES), which is a $(\mu + \lambda)$-ES that is a (1+1)-ES when we set $\mu = \lambda = 1$. But the behavior of step size adaptation is unclear given the proposed conditions.

There is a wealth of literatures on solving black box optimization using (1+1)-ES on solving unimodal test problems for evaluation and the convergence property of convex functions. Arash et al [10] proposed a surrogated-assisted (1+1)-ES that investigates the acceleration and the step size adaptation behavior of the algorithm using GP based online local surrogate. In this model, the local surrogate model filters the undesired candidate solutions by comparing the objective function value of the parent with the GP estimate of the offspring. The candidate solution is evaluated using the true objective function if and only if its fitness evaluated by GP is superior to its parent. The surrogate model is updated whenever a new objective function call is made. The most recent offspring evaluated by true objective function is then added to the training set for Gaussian Process, replacing the oldest data point in the training set. The proposed GP based local surrogate gives a 3-time-speed-up compared with the usual (1+1)-ES on quadratic sphere. We want to construct a similar GP based local surrogate model and compare the result using the same test functions and analysis.

3 DESCRIPTION

The first few steps of the proposed algorithm are essentially a usual (1+1)-ES where the step size of the parent is adapted using one-fifth rule. In each iteration the parent $x \in R^n$ is the best offspring (with the lowest objective function value) obtained so far. The offspring in this iteration y is then generated by $y = x + \sigma z$ where z is standard normally distributed n-dimensional random vector and σ is the step size of the algorithm.

After K iterations using usual (1+1)-ES we can get the K candidate solutions x_{train} and their corresponding objective function values fx_{train}, referred as the training data for the surrogate model. The training data can be used to build a Gaussian Process (GP) model that makes an estimation of the true objective function value of the next iteration at a much lower cost. In each iteration, x_{train} and fx_{train} are updated using the most recent K candidate solutions and corresponding objective function values where the GP model is updated at the same time.

To get an offspring, we sample a standard normally distributed random vector $z \in R^n$ referred as a direction and the offspring generated in this iteration can be optimized in one dimension (the direction of z) by fitting a quadratic model using the value and the the corresponding objective function values of the parent, candidate solutions ($y_\pm$) taking the positive and negative directions $\pm z$) respectively. Then we choose the lowest point of the quadratic model as the offspring, the best candidate solution (lowest function evaluation for the quadratic model) in that direction. The basic MATLAB code is shown above (in Figure 1).

Since the estimate of GP can be precise in the neighborhoods of the parent solution, we use the GP estimates($f_\epsilon(y\pm)$) of the two candidate solutions taking positive and negative directions to save 2 true objective function evaluations at the cost of a small precision loss. The lowest point of the quadratic model is used as the offspring of this iteration, it is then evaluated with the true objective function where the offspring and its objective function value are added to x_{train} and fx_{train} respectively for GP model update. The next steps are the same as the usual (1+1)-ES where we choose the best of all candidate solutions as the parent for the next iteration.

4 ANALYSIS

To understand the potential implications of using surrogate models in EAs. The evaluation follows what Arash et al [10] did. In this section, we propose a simple model that, in each iteration, optimizes the candidate solution generated using the surrogate models. We first sample a random direction and generate two points $y_\pm$ by adding the positive and negative directions to the parent. The optimization is done by first fitting a quadratic model using the parent (evaluated by the true objective function) and the other two points (evaluated by the surrogate models where the inaccurate estimate can be achieved at vanishing cost), then choosing the lowest point of the quadratic model as the offspring of this iteration. For simplicity, we assume the error of using the GP estimate is a Gaussian random variable with mean coincides with the true objective function value of the candidate solution and some variance introduced as the nosie for the estimate. In this case, we can apply the analysis of evolution strategies with the presence of Gaussian noise (see [2] and references theorem). The analysis could be extended to biased surrogate models where the mean of the estimates' distribution is different from the exact objective function value. It is likely that analyzing the effect of non-Gaussian noise for the performance of ES could be achieved by using models with error distribution that is skew [1].

Considering the minimization of quadratic sphere $f : R^n \to R$ with $f(x) = x^T x$ where the local surrogate model assisted (1+1)-ES is applied, this section will use the surrogate model described above to replace the candidate solution generated in each iteration. We first consider a simple iteration of the strategy. In each iteration the strategy generates a single candidate solution $y = x + \sigma z$ as is described in Section 3 where σ is the step size parameter and x the parent (best candidate solutions obtained so far). Then we randomly sample standard normally distributed random vector $z \in R^n$ referred to as the direction vector and generate two points based on the direction vector $y_\pm = x \pm z$. We want to fit a quadratic model that optimizes y in direction z where the minimization of the quadratic model can be written as:

$$\text{Given } x \in R^n,\ z \in R^n \text{ find } \alpha_{opt} \in R\ s.t.\ min\{f(x+\alpha z)\}$$

We denote the quadratic model $g(\alpha) = a\alpha^2 + b\alpha + c = f(x+\alpha z)$ where the surrogate step size α essentially illustrates the signed distance we added to the given candidate solution y in direction z. For the three points described above, the strategy uses the surrogate model estimate $f_\epsilon(y_\pm)$ to replace the true objective function value of $y_\pm$ that goes like the following:

$$\begin{cases} g(1) = a + b + c = f_\epsilon(y_+) \\ y(-1) = a - b + c = f_\epsilon(y_-) \\ y(0) = c = f(x) \end{cases}$$

Solving the above system gives $\alpha_{opt} = -\frac{b}{2a} = -\frac{f_\epsilon(y_+) - f_\epsilon(y_-)}{f_\epsilon(y_+) - f_\epsilon(y_-) + 2f(x)}$ where $y = x + \alpha_{opt} z$ is the offspring obtained in this iteartion.

By assumption, $f_\epsilon(y_\pm)$ is a random variable with mean $f(y_\pm)$ and some standard deviation $\sigma_\epsilon > 0$ that can be rewritten as $y_\epsilon(y_\pm) = f(y_\pm) + \sigma_\epsilon$. Better surrogate models with small value of σ_ϵ can result more precise estimation of the true objective function. In the case where $\sigma_\epsilon = 0$, we obtain the estimation without Gaussian noise, which is essentially the same as we evaluate using the true objective function.

We use the decomposing of z proposed by Rechenberg [14] to analyze the expected step size of the strategy. Vector z could be decomposed as a vector sum $z = z_1 + z_2$, where z_1 is in the direction of the negative gradient of z, while z_2 orthogonal to z_A. We have z_1 standard normally distributed while $\|z_2\|^2$ χ-distributed with $n-1$ degree of freedom and $\frac{\|z_2\|^2}{n} \overset{n\to\infty}{=} 0$. The estimate and optimal α_{opt} of this iteration using surrogate model results the following where $z_\epsilon^\pm$ is standard normally distributed.

$$f_\epsilon(y_\pm) = (R \mp \sigma z_1) + \sigma\|z_2\|^2 + \sigma_\epsilon^\pm$$

$$\alpha_{opt} = \frac{4\sigma z_1 + \sigma_\epsilon(z_\epsilon^+ - z_\epsilon^-)}{2(2\sigma^2\|z\|^2 + \sigma_\epsilon(z_\epsilon^+ + z_\epsilon^-))}$$

The normalized fitness advantage of the offspring over its parent is denoted as $\delta = n(f(x) - f(y))/(2R^2)$ where $R = \|x\|$ is referred to as the distance from the parent to the global minima. By introducing the normalized step size $\sigma^* = n\sigma/n$ and estimated normalized fitness advantage $\sigma_\epsilon^* = n\sigma_\epsilon/(2R^2)$(the normalized fitness advantage using the surrogate model) we can simplify α_{opt}

$$\begin{aligned} \alpha_{opt} &= -\frac{-2\sigma^* z_1 + \sigma_\epsilon^*(z_\epsilon^+ - z_\epsilon^-)}{2(\sigma *)^2\|z\|^2/n + 2\sigma_\epsilon^*(z_\epsilon^+ - z_\epsilon^-)} \\ &\overset{n\to\infty}{=} -\frac{\sigma^* z_1 - \sigma_\epsilon^*/2(z_\epsilon^+ - z_\epsilon^-)}{(\sigma^*)^2 + \sigma_\epsilon^*(z_\epsilon^+ + z_\epsilon^-)} \end{aligned}$$

Simplify the expected fitness advantage of the offspring over its parent

$$\begin{aligned} \delta &= \frac{n}{2R^2}(x^T x - (x + \alpha_{opt}\sigma z)^T(x - (x + \alpha_{opt}\sigma z)) \\ &\overset{n\to\infty}{=} \alpha_{opt}\sigma^* z_1 - \frac{(\alpha_{opt}\sigma^*)^2}{2} \end{aligned}$$

where $z_1 = -x^T z$ is standard normally distributed random variable that coincides with z in the negative gradient direction and $\overset{n\to\infty}{=}$ denotes the convergence of the distribution. The equation could be further written as

$$\delta(x,y) \overset{n\to\infty}{=} x\sigma^* y - \frac{(x\sigma^*)^2}{2}$$

The probability desity of α_{opt} conditioned on z_1 obtained by Dirk is

$$\begin{aligned} &p_{\alpha|z1}(x|y) \\ &= 2De^{\frac{-E^2}{2(1+D^2)}} \frac{\left[(\sigma_\epsilon^* + 2y + 2\theta\frac{E}{1+D^2})(1 - 2normcdf(\beta)) + 4\theta\frac{De^{-\beta^2/2}}{\sqrt{2\pi}\sqrt{1+D^2}}\right]}{(1-2x)^2\theta\sqrt{2\pi}\sqrt{1+D^2}} \end{aligned}$$

where we have the following

$$\begin{cases} \theta = \frac{\sigma_\epsilon^*}{\sigma^*} \\ E = \frac{-2(y + x\sigma_\epsilon^*)}{1+2x}/\theta \\ D = \frac{1-2x}{1+2x} \\ \beta = \frac{(\sigma_\epsilon^*/2 + y) * \sqrt{1+D^2} + \theta E/\sqrt{1+D^2}}{D} \\ normcdf(\beta) : \text{cdf of N(0,1) evaluated at } \beta \end{cases}$$

In each iteration, one objective function call is made after the candidate solution is optimized using the quadratic model. The offspring y replacing its parent x if and only if $\delta > 0$. We write $p_{step} = Prob[\delta > 0]$ for the probability of the offspring replacing its parent

$$p_{step} = Prob[\delta > 0] = \int_{-\infty}^{\infty} p_{z_1}(y) \int_{-\infty}^{\infty} p_{\alpha|z1}(x|y)dxdy$$

Finally the expected value of the normalized change in objective function is

$$\Delta = \begin{cases} \delta & \text{if } \delta > 0 \\ 0 & \text{otherwise} \end{cases}$$

In each itertaion, it could be computed as

$$E[\Delta] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-y^2/2} p_{\alpha|z1}(x|y)\delta(x,y)dxdy$$

We define $\theta = \sigma_\epsilon^*/\sigma^*$ the noise-to-signal ratio to measure the quality of surrogate model relative to the algorithm's step size and the function to be integrated is denoted as

$$g(x, y, \sigma_\epsilon^*, \sigma^*) = 1/\sqrt{2\pi}e^{-y^2/2}p_{\alpha|z1}(x|y)\delta(x,y)$$

The expected fitness gain relative to normalized step size is plotted in figure 2. The dots show the corresponding values observed in experiments with unbiased Gaussian surrogate error for $n \in \{10, 100\}$ obtained by averaging 40 runs. We tried to compute the expected fitness gain ($n \to \infty$) over the normalized step size analytically, but feailed because the function to integrate does not seem to converge over y. We experimented with different values of σ_ϵ^* and θ for $g(x, y, \sigma_\epsilon^*, \sigma^*)$ but the value obtained is a negative constant along the y-axes when y exceeds a certain value for a range of x. An example of the plot for $g(x, y, \sigma_\epsilon^*, \sigma^*)$ over x and y with $\sigma_\epsilon^* = 1$ and $\sigma^* = 1$ is shwon in Fig.3 where two half-cylinder-shape along y-axes can be observed. $n \to \infty$ can be intepreted as a large step size with small noise-to-signal ratio.

It can be seen from Fig.2 that for a small noise-to-signal ratio, the evaluation rate increases with an increasing step size. While for a large noise-to-signal ratio, evaluation rate decreases with an increasing step size. The strategy makes progress one out of two steps for very small step size. The strategy becomes more "wise" in generating offsprings with a large step size and a relative small noise-to-signal ratio in a sense that the candidate solution optimized in this iteration is more close to the true lowest point of the quadratic model with a relative large space for candidate generation. In the case of zero noise-to-signal ratio, the candidate solution is deemed superior to its parent by choosing the best candidate solution in the direction sampled.

The expected fitness gain decreases as the step size increases (observed in Fig. 2 $\theta = 4$ green dots) which is contrary to the case where $\theta = 0.25$ and $\theta = 1$. One intepretation is that the strategy samples one direction in each iteration and the candidate solution is optimized in that direction using a quadratic model. The fitness gain of the candidate solution largely depends on the direction sampled and the accurancy of the estimation for the points fitting the quadratic model. The noise-to-signal ratio θ controls the precision of the optimal candidate solution estimtaed in the quadratic model relative to step size. A large θ means the estimation of the optimal candidate solution in the direction sampled is less accurate and therefore the evaluation rate decreases as the step size increases (the increasing step size also adds inaccurancy to the estimation of the optimal candidate solution).

The direction sampled also affects the model performance. We would like to optimize the candidate solution in a direction with a potentially large fitness gain. But random sampling makes the process uncertain and according to experimental results, chances are that we do not sample a good direction in most cases i.e. the imporvement for the parent is limitted. So that a selection mechanism for the (direction of) candidate solutions in each iteration should be further considered. One possible approach is to consider the difference (distance) of the candidate solution before and after optimization using the quadratic model. Or we could possibly try unit vectors in each dimension, optimize using the quadratic model and choose the direction with the largest absolute $-2a/b$ (surrogate step size) in the optimization using the quadratic model.

5 STEP SIZE ADAPTATION

6 CONCLUSIONS

In this paper, We proposed a local surrogate-assisted (1+1)-ES on a quadratic sphere function. The strategy uses a lcoal surrogate model to optimize the candidate solution obtained in each iteration. The performance is analyzed by adding different levels of Gaussin noise in the strategy.

For future work, we will work on a selection mechanism for candidate solutions deciding what candidate solutions to choose based on the fitness gain it may bring. Further, a step size adaptation mechanism for the surrogate model assisted (1 + 1)-ES should be considered.

REFERENCES

[1] D. V. Arnold and H. -. Beyer. 2006. A general noise model and its effects on evolution strategy performance. *IEEE Transactions on Evolutionary Computation*, 10, 4, (Aug. 2006), 380–391.

[2] Dirk V. Arnold and Hans-georg Beyer. 2002. Noisy Optimization with Evolution Strategies. Tech. rep. SIAM Journal on Optimization.

[3] 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.

[4] Yu Chen and Xiufen Zou. 2014. Performance analysis of a (1+1) surrogate-assisted evolutionary algorithm. In *Intelligent Computing Theory*. De-Shuang Huang, Vitoantonio Bevilacqua, and Prashan Premaratne, (Eds.) Springer International Publishing, Cham, 32–40.

[5] B. Dunham, D. Fridshal, R. Fridshal, and J. H. North. 1963. Design by natural selection. *Synthese*, 15, 1, (Jan. 1963), 254–259. https://doi.org/10.1007/BF00484855.

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

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

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

[9] Yaochu Jin, Markus Olhofer, and Bernhard Sendhoff. 2002. A framework for evolutionary optimization with approximate fitness functions. *IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION*, 6, 481–494.

[10] Arash Kayhani and Dirk V. Arnold. 2018. Design of a surrogate model assisted (1 + 1)-es.

[11] 2016. *Evolution strategies. Machine Learning for Evolution Strategies*. Springer International Publishing, Cham, 13–21. https://doi.org/10.1007/978-3-319-33383-0_2.

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

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

[14] 1973. *Evolutionsstrategie: optimierung technischer systeme nach prinzipien der biologischen evolution*. Problemata, 15. Frommann-Holzboog. https://books.google.ca/books?id=-WAQAQAAMAAJ.

[15] Jerome Sacks, William J. Welch, Toby J. Mitchell, and Henry P. Wynn. 1989. Design and analysis of computer experiments. *Statist. Sci.*, 4, 4, (Nov. 1989), 409–423. https://doi.org/10.1214/ss/1177012413.

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

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

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

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