Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

COMBINED GLOBAL AND LOCAL METHOD FOR STOCHASTIC SIMULATION OPTIMIZATION WITH AN AGLGP MODEL

Qun Meng
Szu Hui Ng

Department of Industrial & Systems Engineering
National University of Singapore
SINGAPORE

ABSTRACT

Surrogate methods, motivated from expensive black box simulations, are efficient approaches to solve stochastic simulation optimization problems. However, estimating an appropriate surrogate model can still be computationally challenging when the data size gets large. In this paper, we propose a new optimization algorithm based on a previously proposed Additive Global and Local Gaussian Process model (AGLGP). This algorithm leverages the global and local features of an AGLGP model and can automatically switch between a global search (for a promising region) and a local search (within the promising region). The algorithm proceeds by globally narrowing down the search space sequentially, but allows it to escape from the current search region. We numerically illustrate the mechanism behind the algorithm in an example.

1 INTRODUCTION

In practice, simulation models are widely built to model and analyze the behavior of real systems, and can evaluate the system behavior with different sets of parameters. When tasked with an optimization problem, simulations can be adopted iteratively to evaluate the performance of the model when simulation is run using parameters generated by optimization algorithms. Given an objective function $f(\theta)$, where the design parameters θ have a feasible region Θ, a general target of global optimization is to find the optimal solution θ^* that satisfies $f(\theta^*) = \min_{\theta\in\Theta} f(\theta)$. In this paper, we focus on a continuous objective function f with a continuous feasible region on a compact subset of $\mathbb{R}^d$. Embedded with stochastic noise, $f(\theta)$ cannot be obtained directly, but rather sample observations with noise $y(\theta,w)$ can be observed, where w denotes the stochastic effects. Here we are interested in the stochastic optimization problem where $f(\theta) = E(y(\theta,w))$.

To solve such problems, there are generally two classes of approaches: direct methods and surrogate-based methods. Popular direct methods include adaptive random search such as COMPASS (Hong and Nelson 2006), nested partition (NP) methods (Shi and Ólafsson 2000), and other heuristic methods such as genetic algorithm and simulated annealing approaches. These direct methods are applied directly into the simulator and the search is conducted on the simulator. COMPASS provides an efficient method to find local optimal solutions in a stochastic environment. NP systematically partitions the feasible region into subregions, assesses the potential of each region, and then concentrates the computational effort in the most-promising region. Some direct methods have been shown to be globally convergent, but they typically require a large number of simulations to obtain adequately good solutions.

Surrogate methods estimate a surrogate model (also known as *metamodel, or response surface model*) with few simulations in the search procedure, to quickly evaluate the performance at any given point in the domain space without the need to run the simulator at every potential point. Such methods provide the information of the entire surface for better identifying the points for further simulation. The Gaussian Process (GP) model has been used in various types of black-box optimization problems. A review of global optimization algorithms for deterministic computer models has been conducted by Jones (2001).

The Efficient Global Optimization (EGO) algorithm (Jones et al. 1998) is a popular and widely used algorithm that determines the next evaluation point by maximizing the expected improvement (EI) from the current optimal solution. Huang et al. (2006), Picheny et al. (2013), and Quan et al. (2013) further extended this framework to stochastic simulation with homogeneous or heterogeneous variance.

However, in GP-based optimization algorithms like EGO, the estimation of the GP model requires an inversion of an $n \times n$ covariance matrix in each iteration, which might require a significant amount of computational effort if the number of evaluation points is large. The matrix inversion in GP estimation can also encounter numerical problems if the evaluation points are close to each other, resulting in an ill-conditioned matrix. This can happen often in global optimization problems, when the algorithm exploits solutions around potential good points. Rasmussen and Williams (2006) discusses various approximation schemes, but these techniques are complicated and can be computationally intensive so that they are not suitable for iterative search algorithms with model fitting in each iteration.

In this paper, we focus on optimization problems where the number of evaluations observed can get large (growing to thousands for simulations that run relatively fast in minutes) as iterations progress to find the optimal point. In such cases, traditional GP-based search algorithms can become computationally expensive and unable to handle the problem. However, a global surrogate-based approach is still attractive as it can provide global information about the response to better guide the exploration and improve convergence (Stephens and Baritompa 1998). Sun et al. (2014) proposed a Gaussian-based Search (GPS) algorithm which built the sampling distribution based on a quickly constructed GP model to overcome some of the computational expense; but it is mainly designed for discrete optimization problems with small variances. More sophisticated and efficient estimation schemes are required for heterogeneous noisy simulations.

Moreover, the EGO algorithm and its extensions are faced with an expensive optimization problem to find the next evaluation point when the domain space is large (as it requires evaluating the expected improvement for all possible points). Regis and Shoemaker (2007) introduce the Stochastic Response Surface (SRS) method that simplifies this optimization problem by carefully generating a small set of random candidate points to evaluate in each iteration. They show that the algorithm is convergent without needing to solve an expensive optimization problem at each step.

In this paper, we propose a combined global and local optimization (CGLO) algorithm that applies the Additive Global and Local Gaussian Process (AGLGP) model (Meng and Ng 2015), which is an additive surrogate that consists of a global model that captures the global trend and local models that capture the local residuals. Based on the global model, a promising local area is selected in the global search stage, and an adaptive local search algorithm that incorporates the local model helps to search in the local area. To reduce the computational burden from optimizing the search criteria, the SRS method is adopted in both the global and local search stages. With the fast estimated AGLGP model, the CLGO spends less time for model estimation in each iteration, while with a two-stage searching criterion, it systematically searches promising regions with a global stage and then searches more deeply into the region with the local stage. The balance of the two in an iterative manner provides more opportunities to exploration and exploitation both locally and globally than a single stage/single criterion approach. So the CGLO significantly improves the efficiency of traditional GP-based optimization algorithm when the data size gets large. The rest of the paper is organized as follows. Section 2 gives an overview of our approach and then reviews the background of the components of the approach: the AGLGP model, the expected improvement criterion we adopt and the SRS method. In Section 3, we provide the details of the combined global and local search algorithm. Numerical results are reported in section 4, followed by conclusions in section 5.

2 BASICS AND NOTATIONS

The general framework of CGLO algorithm is shown in Figure 1. Here we adopt the AGLGP model as a surrogate for optimization, leveraging its global model for a global search and its local model for a local search. We also adopt the SRS method with the expected improvement as the searching criterion for both the global search and local search stages. We define the notations used for the AGLGP model and

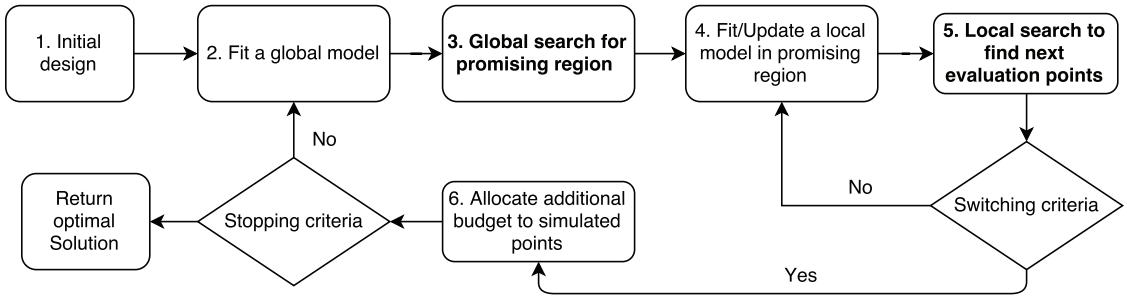


Figure 1: General framework of CGLO.

algorithm development in Table 1. Since we are focusing on stochastic simulations, multiple replications can be allocated to each evaluation point x, and the average of replications will be taken as the observation at each x, i.e., $y(x) = \sum_{w=1}^{r} y_w(x)/r$.

Table 1: Parameter Definitions.

Parameter	Definition	Param	Definition
$\mathbf{R}_k, k = 1,...,K$	Non-overlapping local region	$\mathbb{X}_\Omega$	Complete domain space $\cup_{k=1}^{K}\mathbf{R}_k$
$\mathbf{x} = (x_1,...,x_n)$	n evaluation points	$\mathbf{y}$	Simulation observations at $\mathbf{x}$
$x_i = (x_{i1},...,x_{id})$	d-dimension evaluation point	$\mathbf{x}_g$	m inducing points
$x_g^i = (x_g^{i1},...,x_g^{id})$	d-dimension inducing point	$\mathbf{y_g}$	Latent global observations at $\mathbf{x}_g$
$\mathbf{x_l^k} = (x_l^1,...,x_l^{r_k})$	r_k evaluation points in local region $\mathbf{R}_k$	$\mathbf{y_l^k}$	Latent local residuals at $\mathbf{x_l^k}$
Ω_g	n_c candidate points for global search	x_g^0	Point selected by global search
Ω_l	n_l candidate points for local search	x^*	Point selected by local search
T	Total budget size	n_0	Number of initial designs
r_{min}	Minimum replications for a new point	$B_{t,s}$	Budget for search step
B_t	Budget allocated in iteration t	N_i	Budget for evaluation point x_i
$B_{t,a}$	Budget for allocation step	$B_{t,b}$	Budget for optimal solution

2.1 AGLGP Model

The AGLGP model (Meng and Ng 2015) assumes that the stochastic simulation response can be modeled as a realization of a random process

$$y(x) = f(x) + \varepsilon(x) = f_{global}(x) + \sum_{k=1}^{K} w_k f_{local}^k(x) + \varepsilon(x), \quad w_k = \begin{cases} 1, & x \in \mathbf{R}_k \\ 0, & x \notin \mathbf{R}_k \end{cases} \tag{1}$$

where $f(x)$ is the deterministic mean function of the stochastic response. The random noise $\varepsilon(x)$ has a normal distribution $\varepsilon(\mathbf{x}) \sim N(0, \sigma_\varepsilon^2(x))$, which is independent and identically distributed across replications and uncorrelated at different locations. The mean function $f(x)$ can be further decomposed to a global model $f_{global}(x)$, which models the global trend, and K local models with each local model $f_{local}^k(x)$ modeling the residual process that is unexplained by $f_{global}(x)$ in local region $\mathbf{R}_k$, where $\cup_{k=1}^{K}\mathbf{R}_k = \mathbb{X}_\Omega$. To capture the global trend, a set of m inducing points are used (where $m \ll n$), so the selection of inducing points is crucial for good estimation of the model. The GP models $f_{global}(x)$ and $f_{local}^k(x)$ are assumed to be piece-wise independent.

Figure 2 illustrates the general idea of the AGLGP model, which is an additive combination of a global model and local models. To fit the model, the entire space is divided into several nonoverlapping local regions via classification techniques like Support Vector Machine (SVM) (as shown in Figure 3).

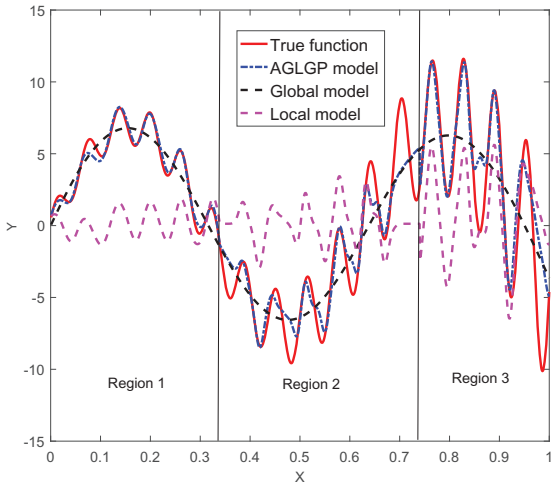


Figure 2: AGLGP model.

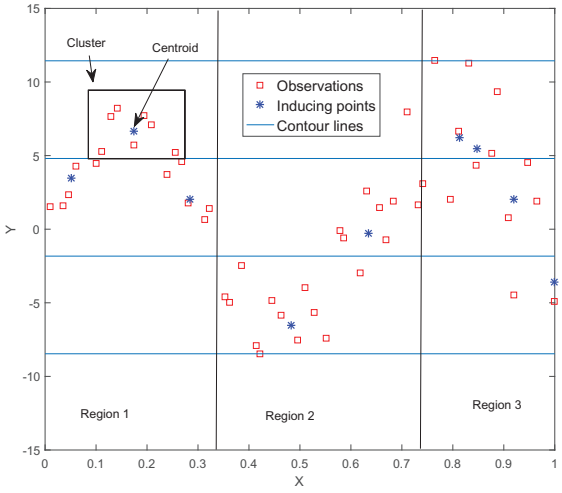


Figure 3: Inducing points and local regions.

To obtain inducing points to fit the global model, within each local region, we further separate points into clusters based on a set of equally spaced contour lines from the minimum observation value to the maximum observation value. This avoids clustering points that have large variability together. The purpose of this clustering of similar observations (in both x and y space) is to obtain a cluster centroid which can reasonably represent the observations in the cluster and be a representative point (called an *inducing point*) of the cluster. With a good spread of evaluation points in the domain, the inducing points (although fewer) will also have a good spread in the entire space and will be able to capture the global trend.

As the global model $f_{global}(x)$ is a latent process modeling only the inducing points, it is not reasonable to include noise in the global model. It is assumed that $f_{global}(x)$ can be modeled by a deterministic GP model with a mean β_0 and covariance $\sigma^2 r_g(x_i - x_j, \boldsymbol{\theta})$, where σ^2 is the variance of the global component and $r_g(\cdot)$ is the correlation structure with a sensitivity parameter $\boldsymbol{\theta}$. Given the set of inducing points and the global evaluations $\mathbf{y}_g$, the best linear unbiased global predictor can then be written as

$$\widehat{y}_{global}(x) = \beta_0 + \mathbf{g}'\mathbf{G}_\mathbf{m}^{-1}(\mathbf{y}_g - \mathbf{1}'\beta_0), \tag{2}$$

where $\mathbf{g} = (g(x - x_g^i), \cdots, g(x - x_g^m))$, $\mathbf{G_m}$ is an $m \times m$ covariance matrix with ijth element $g(x_g^i - x_g^j)$. The global predictor interpolates $\mathbf{y}_g$ since $\widehat{y}_{global}(\mathbf{x}_g^j) = \beta_0 + \mathbf{e_i}'(\mathbf{y}_g - \mathbf{1}'\beta_0) = y_g^i$. With the fitted global model, the global predictors at $\mathbf{x}$ are $\widehat{\mathbf{y}}_{global} = (\widehat{y}_{global}(x_1), ..., \widehat{y}_{global}(x_n))$. The residuals, which include both the residuals from the signal function and the random noise, are then obtained by $\mathbf{y_l} = \mathbf{y} - \widehat{\mathbf{y}}_{global}$ and modeled by another stochastic GP model $y_{local}(x) = \sum_{k=1}^{K} w_k f_{local}^k(x) + \varepsilon(x)$, where $f_{local}^k(x) \sim N(0, \tau_k{}^2 r_k(x_l^i - x_l^j, \boldsymbol{\alpha}_k))$, and $\boldsymbol{\alpha}_k$ is the sensitivity parameter for the local model. We assume that the residual process is correlated within a local region while independent across the regions, so different correlation functions are allowed in different regions. This enables the flexibility to capture nonstationarity in the process. Given K local regions $\mathbf{R} = (\mathbf{R_1}, ...\mathbf{R_K})$, the local predictor is given by

$$\widehat{y}_{local}(x) = \mathbf{l_k}'(\mathbf{L_k} + \boldsymbol{\Sigma_\varepsilon})^{-1}\mathbf{y_l^k}, \quad \forall x \in \mathbf{R}_k, \tag{3}$$

where $\mathbf{l_k} = (l_k(x - x_l^i), ...l_k(x - x_l^{r_k}))$ and $\mathbf{L_k}$ is covariance matrix with (jh) element $l_k(x_l^h - x_l^j)$, $\forall x_l^j, x_l^h \in \mathbf{x_l^k}$. The quantities $\boldsymbol{\Sigma_\varepsilon} = diag(\hat{\sigma}_\varepsilon^2(x_l^1), ..., \hat{\sigma}_\varepsilon^2(x_l^{r_k}))$ and $\hat{\sigma}_\varepsilon^2(x_l^i)$ can be estimated from the sample variance. So the overall AGLGP predictor can be expressed by

$$\widehat{y}(x) = \widehat{y}_{global}(x) + \widehat{y}_{local}(x) = \beta_0 + \mathbf{g}'\mathbf{G}_\mathbf{m}^{-1}(\mathbf{y}_g - \mathbf{1}'\beta_0) + \mathbf{l_k}'(\mathbf{L_k} + \boldsymbol{\Sigma_\varepsilon})^{-1}\mathbf{y_l^k}, \quad \forall x \in \mathbf{R}_k. \tag{4}$$

As $\mathbf{y_g}$ and $\mathbf{y_l^k}$ are latent processes that cannot be observed directly, the predictive distribution of any input x can be derived by integrating out the random variable $\mathbf{y_g}$ and $\mathbf{y_l^k}$, see Meng and Ng (2015). In Section 3, we will discuss an alternative sequential estimation method that first estimates the global model and its parameters from the marginal likelihood of the global model and then estimates the local model and its parameters from the predictive residuals. This sequential approach is computationally faster and can be better incorporated into the optimization algorithm.

2.2 Expected Improvement

The EGO algorithm (Jones et al. 1998) sequentially selects evaluation points based on the EI criterion, which balances the trade-off between exploration (searching the entire space for regions that have not been explored before) and exploitation (searching around the current optimal solution for better solutions). At a point x_0, the EI function is defined as

$$EI(x_0) = E[\max(y_{min} - y(x_0), 0)].$$

where y_{min} is the observed best function value while $y(x_0)$ is a random variable that models the uncertainty of the function value at x_0, where $y(x_0) \sim N(\widehat{y}(x_0), \widehat{s}^2(x_0))$ with predictive mean $\widehat{y}(x_0)$ and predictive variance $\widehat{s}^2(x_0))$. The point that maximizes the EI function will be selected for next evaluation either due to its low predictive mean (for exploitation) or due to its high predictive variance (for exploration). Huang et al. (2006) proposed an augmented EI (AEI) function for stochastic simulation systems. The AEI accounts for the influence of the random noise $\varepsilon(x)$ and considers selecting and adding replications on the existing evaluation points other than searching for new points.

A two-stage sequential optimization (TSSO) algorithm (Quan et al. 2013) further extended the EGO framework to the heterogeneous case, which comprises of a search stage to determine the next evaluation point and an allocation stage to evaluate the best optimal solution by running simulations at each evaluated point according to Optimal Computing Budget Allocation (OCBA) strategy. The search criterion is defined by the modified EI function

$$mEI(x_0) = E(\max[y_{min} - z(x_0), 0]). \tag{5}$$

where $z(x_0)$ is a normal random variable with mean given by MNEK predictor (Yin et al. 2011) and variance given by spatial prediction uncertainty. By ignoring predictor uncertainty caused by random variability, the search stage can focus on new points with a low predicted response or a high spatial uncertainty.

2.3 Stochastic Response Surface Method

Regis and Shoemaker (2007) introduced a Stochastic Response Surface (SRS) method that simplifies the optimization problem of finding the next evaluation point. The SRS method follows the same framework with the EGO algorithm, i.e., simulate at initial designs and iteratively fit or update a response surface model, select the next evaluation point, run simulations at the selected point. Differently, SRS selects the next evaluation point only among the set of candidate points without having to evaluate all possible solutions. This sampling framework is shown to have nice convergence properties and has wide adaptability. It can be applied to various metamodel based optimization approaches and search criteria (like the EI, the probability of improvement). The probability distribution such as uniform or normal distributions can be applied to generate candidate points.

3 COMBINED GLOBAL AND LOCAL OPTIMIZATION WITH AGLGP MODEL

In this section, we explain in detail the development of the CGLO algorithm that incorporates the global and local characteristics of the AGLGP model. From Figure 1, we know that each iteration of the algorithm comprises of a global search stage (step 2,3) and a local search stage (step 4,5,6). Local search here emphasizes the search in the promising local region. The terminology is not to be confused with the search for a local solution to the optimization problem.

The global search stage separates the whole space into several local areas and applies a modified EI-type function to identify which area has a small number of evaluated points or a low global trend while the local search stage explores and exploits in the promising area with *mEI* (see Equation (10)). So they combine to realize the exploration and exploitation of the whole space, balancing the efforts between the current local area and the other unexplored local areas in a stochastic environment. The algorithm switches between these two stages adaptively to quickly identify the promising regions. In the following subsections, we will explain in detail each search stage.

3.1 Global Search Stage

In the global search stage, the CGLO algorithm focuses on narrowing down the entire space to a promising local area based on an estimated global model, which incorporates the latent variables $\mathbf{y}_g$ in equation (2). Here we describe an alternative method to estimate the predictive global and local models sequentially. To determine the global predictive mean and variance, we first derive the marginal likelihood can be derived as

$$p(\mathbf{y}|\mathbf{x}_g) = N(\beta_0, \mathbf{G_{nm}}\mathbf{G_m^{-1}}\mathbf{G_{mn}} + \mathbf{\Lambda} + \mathbf{\Sigma_\varepsilon}),$$

where $\mathbf{\Lambda} = diag(\mathbf{G_n} - \mathbf{G_{nm}}\mathbf{G_m^{-1}}\mathbf{G_{mn}})$. The global predictive mean and variance at a point x can then be shown to be

$$\hat{y}_g(x) = \beta_0 + \mathbf{g}'\mathbf{Q_m^{-1}}\mathbf{G_{mn}}(\mathbf{\Lambda} + \mathbf{\Sigma_\varepsilon})^{-1}(\mathbf{y} - \mathbf{1}'\beta_0), \tag{6}$$

$$\hat{s}_g^2(x) = [G_{nn} - \mathbf{g}'\mathbf{G_m^{-1}}\mathbf{g}] + \mathbf{g}'\mathbf{Q_m^{-1}}\mathbf{g}. \tag{7}$$

where $\mathbf{Q_m} = \mathbf{G_m} + \mathbf{G_{mn}}(\mathbf{\Lambda} + \mathbf{\Sigma_\varepsilon})^{-1}\mathbf{G_{nm}}$. With the assumption of a deterministic global model, equation (2) interpolates at the inducing points to give the global trend $\mathbf{y}_g$. We observe that although the global model is assumed deterministic, the global predictors can only be estimated through the noisy observations $\mathbf{y}$ (Equation 6), and this noise is reflected in the mean squared prediction error in equation (7). The term in brackets in (7) is the mean square error of a deterministic GP model, which takes zero at inducing points, but the second term in the right hand-side is positive, reflecting how the intrinsic noise inflates the mean squared error. In this global stage, we adopt the SRS framework in the evaluation of the search criterion.

3.1.1 Generation of Candidate Points

To generate the candidate points to evaluate the EI search criterion at, we first define the vector Ω_g as the set of candidate points, which includes n_c items inside. The following condition will be imposed on the selection of candidates. [C1] The components in Ω_g have to be equally spread out across the entire space and be at least equal to the number of local regions such that there exists at least one candidate point in each local region. Possible choices to spread these points include the Latin Hypercube design and the Treed partitioning approach that divides up the entire space by making binary splits on the value of each single variable (Breiman et al. 1984). Based on numerical tests, we find that about $n_c = 10k$ points is an reasonable choice for k dimensional problems.

For each candidate point in a local region $\mathbf{R}_k$, the local area is a subset of the local region and defined as the neighborhood of this candidate point x_i, i.e., $\mathbb{B}_i = \{x \in \mathbf{R}_k : ||x - x_i|| \leq ||x - x_j|| \forall x_j \in \Omega_g \cap \mathbf{R}_k\}$. The local area where the most promising point lies will be selected as the the most promising area. If there is one candidate point in each local region, the local areas are defined the same as local regions. In Section 3.1.2, we will explain how the most promising point is selected.

3.1.2 Global Expected Improvement

We propose to use the EI-type function due to its ability to balance between exploration and exploitation across the entire space. When the EI function based on the global model is applied to identify the most

promising point, note that it only reflects the information of inducing points as $\hat{s}_g^2(x)$ can only reflect the spatial uncertainty given the inducing points instead of the true observations. Hence, applying the EI function alone is not sufficient to reflect the density and spread of the observations.

With the global model, when a point has high spatial uncertainty, it indicates that there are few or no inducing points in the area of the point. When there are no inducing point, which indicates no observations, exploration should be done in that area. When there are few inducing points, this implies two possibilities on the distribution of observations. If the few inducing points is due to few observations in that area, it is worth to explore there, while if the few inducing points is due to a large number of observations (with small variability among them) in that area being very similar and hence clustered together requiring only one inducing point to closely represent them, there is little need to further explore in that area as it has many observations that are all close together there already. To avoid expending budget in areas that already have dense observations, we propose a modified global EI function for the global search,

$$gEI(x) = E\{\max(y_{gmin} - y_g(x), 0)\} \cdot \frac{1}{1+e^{n_i/t-5}}. \tag{8}$$

where y_{gmin} is the lowest predicted global quantity among the current inducing points. The quantity $y_g(x)$ is a normal random variable with mean $\hat{y}_g(x)$ and variance $\hat{s}_g^2(x)$. It consists of the EI function on the global model and a penalty factor designed to account for the diminishing return of selecting the same candidate point as the observations become more dense around it. Together, this criterion is able to explore and exploit the areas with few inducing points and few observations. We select the promising point x_g^0 by maximizing the *gEI* from the *candidate points* Ω_g.

To identify a promising candidate point that has few observations around, the density of the observations within its neighborhood has to be evaluated. The penalty factor has n_i, the number of neighbors of each candidate point x_c^i, and this decreases as n_i increases. Specifically, neighbors of $x_c^i \in \mathbf{R}_k$ is defined as $\mathbb{B}(x_c^i) = \{x \in \mathbf{x_l^k} : ||x - x_c^i|| < ||x - x_c^j||, \forall x_c^j \in \Omega_g \cap \mathbf{R}_k, j \neq i\}$, with each observation x classified to its nearest candidate point in the same local region.

By introducing the penalty factor, the *gEI* can jump quickly between local areas and better explore the entire space. The user-defined parameter t controls the steepness of the function, which represents how fast the gEI decreases with the increase of neighbor observations. Given n_i, the larger the t, the larger the factor. The factor approaches zero when $n_i/t = 10$, so a possible choice for t is $t = MAX/(10n_c)$, where MAX is the maximum number of evaluated points allowed given the total budget, and n_c is the number of global candidate points.

3.2 Local Search Stage

Here we provide the details on the local search. Once the candidate point x_g^0 is selected by the global search, the local area in which x_g^0 lies will be selected as the promising area and a local model (equation 3) is constructed in the local region $\mathbf{R}_k$ where it lies to capture the local residuals in this region. The local model is estimated only with the data $y_l(x_i) = y(x_i) - \hat{y}_g(x_i), \forall x_i \in \mathbf{R}_k$. Then the overall model can be expressed as

$$\hat{y}(x) = \hat{y}_g(x) + \hat{y}_l(x) = \beta_0 + \mathbf{g'Q_m^{-1}G_{mn}}(\mathbf{\Lambda + \Sigma_\varepsilon})^{-\mathbf{1}}(\mathbf{y} - \mathbf{1'}\beta_0) + \mathbf{l_k}'(\mathbf{L_k} + \mathbf{\Sigma_\varepsilon})^{-\mathbf{1}}\mathbf{y_l^k}, \forall x \in \mathbf{R}_k. \tag{9}$$

In this stage, we search deeply into the promising area based on the overall model followed by extra evaluations are conducted in this selected local region. Similar to TSSO (Quan et al. 2013), the local search stage can be further divided into a search step and an allocation step. In the search step, the mEI in Equation (10) is used to sequentially select a number of new points until a switching criterion for another iterative global search is satisfied. After the search step, additional simulation replications are intelligently distributed among all evaluated points. By ignoring predictor uncertainty caused by random variability, the mEI function in the search step assumes that the observations are made with infinite precision so the

same point is never selected again. This allows the algorithm to quickly escape from a local optimum in the region. The random variability is further managed in the allocation step.

In the notation below, we will use B_t to denote the number of replications allocated to the local search stage at iteration t, which can be separated into the searching budget $B_{t,s}$ for the search step and the allocation budget $B_{t,a}$ for the allocation step, i.e.,$B_t = B_{t,s} + B_{t,a}$. The detailed framework and budget required for each step are described in the next subsections.

3.2.1 The Search Step

In this step, we again follow the SRS framework to evaluate the search criterion. As the search step continues to select new evaluation points until a switching criterion is satisfied, the number of points selected in the search step n_t is dynamically decided in each iteration. Each new evaluation point is simulated with r_{min} replications, and the budget $B_{t,s} = n_t \times r_{min}$ is also dynamically decided based on n_t. The candidate points Ω_l have to be uniformly generated within the neighborhood of x_g^0, which is defined as $\mathbb{B}_0 = \{x \in \mathbf{R}_k : ||x - x_g^0|| \leq ||x - x_j|| \, \forall x_j \in \Omega_g \cap \mathbf{R}_k \text{ and } x_j \neq x_g^0\}$, and then mEI is maximized among Ω_l.

$$mEI(x) = E(\max[y_{min} - z(x), 0]), x \in \Omega_l. \tag{10}$$

where y_{min} is the predicted response at the sampled points in the promising region with the lowest sample mean, and $z(x)$ is a normal random variable with mean given by the AGLGP predictor and variance given by spatial prediction uncertainty $\hat{s}_z^2 = L_{nn} - \mathbf{l}' \mathbf{L}_\mathbf{n}^{-1} \mathbf{l}$. Since the mEI function ignores the predictive uncertainty caused by the random noise, the same point will never be selected again once simulated. Each time when a new evaluation point is simulated, the local model and y_{min} will be updated accordingly and a new set of candidate points Ω_l will be generated. The evaluated points in the local region $\mathbf{R}_k$ will be reclustered and the inducing points get updated. The clustering technique is easily implemented.

Switching Criterion The switching criterion dynamically decides whether we continue with the local search in the current local region or return to the global search. It is reasonable to continue searching in the current local region if the new observations and the updated clusters do not make much changes to the location of inducing points, and hence to the global model. However, even if the inducing points do not change much, the evaluations of the gEI function among the global candidate points can change when either the EI changes or the penalty factor reduces. When these occur, a switch back to the global search stage can identify a different promising region.

So the switching criterion proposed here is defined by the increase or the change of location in the inducing points or a maximum number of new evaluated points n_{max}. If the criterion is satisfied, we jump out of the local search step for a new iteration of the global search. This reduces the number of iterations between the global and local search. The refitting of the global model is only done when the global model or the gEI changes. By defining the switching criterion, we also limit the budget exhausted in this step by $n_{max} \times r_{min}$. If the remaining budget is smaller than the limit ($T - \sum_{i=1}^{t-1} B_i < n_{max} \times r_{min}$), additional simulations should focus on the allocation step for finding the optimal solution.

3.2.2 The Allocation Step

Since the search step focuses on searching for new evaluation points in the local region, the allocations step is dedicated to distribute additional simulations among the evaluated points for better evaluation of the noisy observations. For every evaluated points, more simulation replications can be allocated for two purposes. Firstly, to improve the fit of the global model for a more efficient global search, more simulation replications should be allocated to improve the clustering and the estimation of inducing points. Secondly, to improve the local model estimation and the local optimal value (i.e., global optimal in the local region), more simulation replications are needed to improve the observations at the evaluated points. Here, we dynamically decide the budget allocated at this step by taking the maximum of budget allocated to each point that satisfies both the purposes.

To improve the clustering and estimation of inducing points within a local region, the effect of the random noise on the clustering technique has be to evaluated. An evaluation point might get wrongly clustered based on the defined contour lines. Given the contour lines, we define that a point is wrongly clustered if its noisy observation falls into a different cluster from its best cluster, defined as the cluster where its true mean value (first two terms of equation 1) falls into. To decide which evaluation point is likely to be wrongly clustered and how much budget should be allocated to bring it to its best cluster, an evaluation criterion related to the distance to the contour lines and the noise variance is used. It is explained by equation (11) and (12) below.

An observation with a large noise variance or a small distance to the contour lines has a high chance of being wrongly clustered. The noise variance only decreases at the rate of $1/\sqrt{r}$ when averaging over r replications, but we can be at least approximately 99.7% confident that one point x_i is rightly clustered if

$$\Delta_i \geq 3\hat{\sigma}_\varepsilon^2(x_i)/\sqrt{N_{i,v}+r_i}. \tag{11}$$

where Δ_i is the distance to the nearest contour line, $\Delta_i = \min_{y\in\mathbb{Y}} |\bar{y}(x_i) - y|$, and $\hat{\sigma}_\varepsilon^2(x_i)$ is the sample variance estimated through r_i replications, and $N_{i,v}$ is the additional simulations required for improving the cluster where the evaluation point x_i lies. Hence, $N_{i,v} \geq (3\hat{\sigma}_\varepsilon^2(x_i)/\Delta_i)^2 - r_i$. As we don't want to allocate too much budget without the updated information, the evaluation budget allocated to each evaluated point in this step is limited by a small number v_{min} (usually r_{min}). Specifically,

$$N_{i,v} = \begin{cases} \min\{\lceil (3\hat{\sigma}_\varepsilon^2(x_i)/\Delta_i)^2 \rceil - r_i, v_{min}\} & \text{for} \quad (3\hat{\sigma}_\varepsilon^2(x_i)/\Delta_i)^2 > 1, \\ 0 & \text{for} \quad (3\hat{\sigma}_\varepsilon^2(x_i)/\Delta_i)^2 \leq 1. \end{cases} \tag{12}$$

For the second purpose of improving the local model estimation and the local optimal value, additional replications are distributed with the goal of maximizing the probability of the correct selection of the local optimum. OCBA provides a rigorous way of identifying the evaluated point with the best response. eTSSO (Pedrielli and Ng) proposed to dynamically decide the simulation effort for selecting the correct solution. More budget should be allocated for evaluation when the noise is high, and when the noise is lower, more budget is allocated only when the overall model fitting is good enough. Each region is initialized with a equal budget $B_{1,b}^k = B_0, k = 1, \cdots, K$, and in each iteration we increase the budget based on the noise level and the budget that has already been exhausted in the selected promising region.

$$B_{t,b}^k = \begin{cases} B_{t-1,b}^k \left(1 + \frac{\hat{\sigma}_\varepsilon^2}{\hat{s}_z + \hat{\sigma}_\varepsilon^2}\right) & k = k* \\ B_{t-1,b}^r & k \neq k^* \end{cases}, \quad B_{t,b} = B_{t,b}^{k^*} \tag{13}$$

The factor in equation (13) is to balance between the noise at the current set of evaluation points (exploitation) against the predicted noise in unexplored areas in this local region. So $\hat{s}_z^2$ is defined by the predictive mean square error of the latest point selected by the search step given the updated local model and $\hat{\sigma}_\varepsilon^2$ is the estimate of the sample variance at the point to which the OCBA procedure assigns the largest number of replications since it indicates the potential to exploit among the evaluation points. Assuming that each point x_i having a sample mean $\bar{y}_i$ and a sample variance $\sigma_\varepsilon^2(x_i)$, according to Theorem 1 provided by (Chen and Lee 2010), the Approximate Probability of Correct Selection (APCS) can be asymptotically maximized when the available budget tends to infinity and when

$$\frac{N_{i,b}}{N_{j,b}} = \left(\frac{\sigma_\varepsilon(x_i)/\Delta_{b,i}}{\sigma_\varepsilon(x_j)/\Delta_{b,j}}\right)^2, i,j \in \{1,2,...r_{k^*}\}, i \neq j \neq b \tag{14}$$

$$N_b = \sigma_\varepsilon(x_b)\sqrt{\sum_{i=1,i\neq b}^{n} \left(\frac{N_i}{\sigma_\varepsilon(x_i)}\right)^2}$$

where $N_{i,b}$ is the number of simulations allocated to the point x_i, and $\Delta_{b,i} = \bar{y}_i - \bar{y}_b$ with $\bar{y}_b$ denoting the lowest sample mean in the current region. At the end of the allocation stage, the evaluated point with the lowest sample mean will be selected as the global optimal in this region.

So finally we have $N_i = max(N_{i,v}, N_{i,b})$ to satisfy the two purposes of improving the estimation of inducing points and improving the global optimal in the local region. $B_{t,a} = \sum_{j=1}^{r_i} N_j$. We have to notice that the allocation budget required is dynamically decided in each iteration. If the remaining budget $A = T - \sum_{i=1}^{t-1} B_i - B_{t,s} < B_{t,a}$, it is reasonable to allocate the remaining budget to find the best optimal solution by OCBA since the additional budget for validating the global model only works for next iteration of the global search.

3.3 An Algorithm Overview

The proposed global and local search are sequentially executed in each iteration. We initialize a space-filling design n_0 before the iterative search and evaluation. The algorithm can be formulated as follows.

Step 1: *(Initialization)* Run a size n_0 space filling designs, with r_{min} replications allocated to each point. Total initial replications $B_0 = n_0 r_{min}$. $\mathbf{t} = \mathbf{0}$.

Step 2: *(Validation of overall model)* Fit an AGLGP response model to the set of sample means, and use cross validation to ensure that the AGLGP prediction is satisfactory.

While the available replications $A = T - \sum_{i=0}^{t} B_i > 0$, $t = t + 1$

Step 3: *(Global Search)* Search for the location x_0^g that maximizes the gEI criterion Equation (8) given the set of candidates Ω_g, and the local region with x_0^g is selected for local search.

Step 4.1: *(Local Search Step)* $n_t = 1$. **While** $n_t <= n_{max}$ and $A > 0$,
(Fit/Update a local model) Fit or update the local model in the selected local region.
(Randomly Generate Candidate Points) Randomly Generate candidate points Ω_l.
(Select the Next Function Evaluation Point) Search for the location $x_{n_t}^*$ that maximizes the *mEI* criterion Equation (10) from the candidate points Ω_l and evaluate at $x_{n_t}^*$ with r_{min} replications.
(Re-cluster points in the selected local region) **Break Step 4.1** if the number of inducing points increases. Else, set $n_t = n_t + 1$, $A = A - r_{min}$. **end**

Step 4.2: *(Local Allocation Step)* Decide the budget $B_{t,a}$ based on equation (12) and equation (13). If $A > B_{t,a}$, allocate $B_{t,a}$ replications among evaluated points in the local region. Set $A = A - B_{t,a}$ Else, Use OCBA in equation (13) to allocated A replications among sampled designs. **end**

Step 5: *(Return the Optimal Solution)* Return the point with the lowest sample mean.

We note that the number of local regions is an important factor for the accuracy of AGLGP model, and the fitting of the AGLGP model drives the performance of CGLO algorithm. Based on some numerical evidence, we recommend to choose the number of local regions such that each local region has 200 to 600 evaluation points. If the number of the local regions is too large, the CGLO algorithm might waste budget for exploiting an unpromising area due to the coarsely estimated AGLGP model. Conversely if the number is too small, the estimation of AGLGP model will require significant computation time.

4 NUMERICAL STUDY

In this section, we test the following function as an example

$$\max_{0 \le x_1, x_2 \le 100} g(x_1, x_2) = 10 \cdot \frac{sin^6(0.05\pi x_1)}{2^{((x_1-90)/50)^2}} + 10 \cdot \frac{sin^6(0.05\pi x_2)}{2^{((x_2-90)/50)^2}} \tag{15}$$

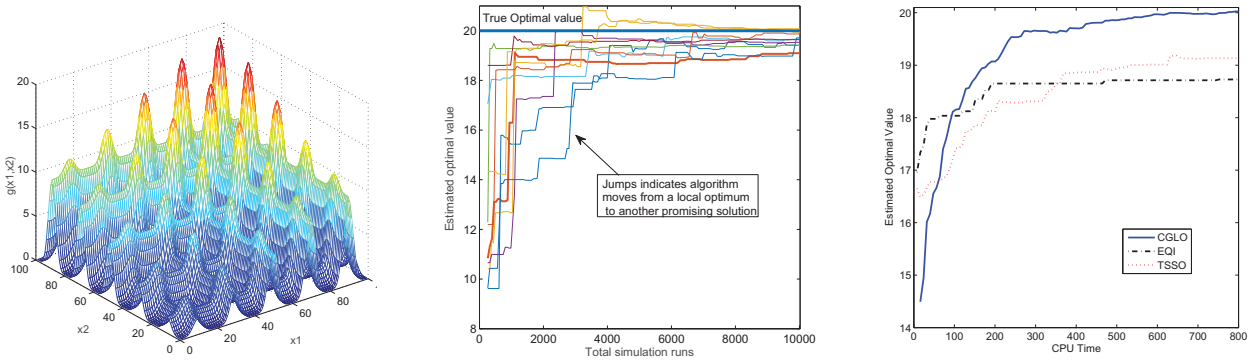
The same example is used by Sun et al. (2014). The function g has a global optimal of $g(90,90) = 20$, and the second best local optimal is $g(70,90) = g(90,70) = 18.95$ (see Figure 4a). We introduce a noise term that is normally distributed with mean 0 and variance $\sigma_\varepsilon^2(x_1,x_2) = 3(1+x_1/100)^2(1+x_2/100)^2$. This problem requires an efficient balance between exploration and exploitation with limited budget because it

has 25 local optimums and the difference between the largest and the second largest local optimal values is quite small. By introducing a large noise variance at the optimal area, it becomes more difficult to solve.

Figure 4b shows 10 macro-replications of the algorithm, with a total simulation budget $T = 10^4$ (approximately 800 evaluated solutions) in each macro replications. We initialize 20 points with Latin Hypercube designs with a minimum of 10 simulations for each point. It shows that in each macro-replication, the estimated optimal value keeps jumping from one solution to a better one over simulation runs. The algorithm is shown to have a good scheme for global optimization since it can successfully escape from local optimums until it get close to the global optimal area.

Results also show that over the 10 macro-replications, the algorithm can find a optimal solution that is very close to the true optimal solution [90,90], and a optimal value that is close the the true optimal value 20. The average Euclidean distance to [90 90] is 0.7173 and the average of optimal value is 19.633. The noise standard deviation around [90 90] can be as high as 1.97, but the evaluation scheme in the algorithm can intelligently allocate more replications to distinguish the global optimal area from its neighboring sub-optimal areas.

Although this example is a small example (with only 800 evaluation points), the results are promising. The average total time taken to run the algorithm (excluding simulation time) as a ratio of the total simulation time is 0.18. This indicates that the CGLO can efficiently deal with more expensive simulations or potentially deal with a larger size of data.



(a) (b) (c)

Figure 4: Estimated optimal value for test function $g(x_1, x_2)$.

To compare the CGLO with other GP-based algorithms, we apply a two stage sequential optimization (TSSO) algorithm (Quan et al. 2013) and a expected quantile improvement (EQI) algorithm (Picheny et al. 2013) to solve the problem. Figure 4c shows the average performances of all three algorithms over 30 macro-replications. We can see the CGLO is significantly better than TSSO and EQI with the same CPU time. Note that the TSSO and EQI are built on the GP model, the estimation of which is much slower than the AGLGP model as iteration goes on. The global search criterion of CGLO algorithm conducts a faster explorative search that allows jump out of a locally optimal area before sufficiently exploiting the local area by the local search.

5 CONCLUSION

This paper proposes a combined global and local search algorithm based on the AGLGP model. The scheme systematically searches promising regions with a global stage and then searches more deeply into the region with the local stage. We then derive an allocation strategy that intelligently allocates budget to the evaluated points for the purpose of improving the metamodel fit and estimating the optimal solution. We also study their numerical performance. For future work, we will compare efficiency with other adaptive random search algorithms and GP-based search algorithms and theoretically derive its asymptotic behavior.

REFERENCES

Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. *Classification and Regression Trees*. Wadsworth.

Chen, C.-H., and L. H. Lee. 2010. *Stochastic Simulation Optimization: An Optimal Computing Budget Allocation*. World scientific.

Hong, L. J., and B. L. Nelson. 2006. "Discrete Optimization via Simulation Using COMPASS". *Operations Research* 54 (1): 115–129.

Huang, D., T. T. Allen, W. I. Notz, and N. Zeng. 2006. "Global Optimization of Stochastic Black-Box Systems via Sequential Kriging Meta-Models". *Journal of Global Optimization* 34 (3): 441–466.

Jones, D. 2001. "A Taxonomy of Global Optimization Methods Based on Response Surfaces". *Journal of Global Optimization* 21 (4): 345–383.

Jones, D., M. Schonlau, and W. Welch. 1998. "Efficient Global Optimization of Expensive Black-box Functions". *Journal of Global Optimization* 13 (4): 455–492.

Meng, Q., and S. H. Ng. 2015. "An Additive Global and Local Gaussian Process Model for Large Data Sets". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 505–516. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Pedrielli, G., and S. H. Ng. "Two Stage Sequential Optimization Approach: Generalization and Asymptotic Properties". *under review*.

Picheny, V., D. Ginsbourger, Y. Richet, and G. Caplin. 2013. "Quantile-Based Optimization of Noisy Computer Experiments With Tunable Precision". *Technometrics* 55 (1): 2–13.

Quan, N., J. Yin, S. H. Ng, and L. H. Lee. 2013. "Simulation Optimization via Kriging: a Sequential Search Using Expected Improvement with Computing Budget Constraints". *IIE Transactions* 45 (7): 763–780.

Rasmussen, C. E., and C. K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. MIT Press.

Regis, R. G., and C. A. Shoemaker. 2007. "A Stochastic Radial Basis Function Method for the Global Optimization of Expensive Functions". *INFORMS Journal on Computing* 19 (4): 497–509.

Shi, L., and S. Ólafsson. 2000. "Nested Partitions Method for Global Optimization". *Operations Research* 48 (3): 390–407.

Stephens, C., and W. Baritompa. 1998. "Global Optimization Requires Global Information". *Journal of Optimization Theory and Applications* 96 (3): 575–588.

Sun, L., L. J. Hong, and Z. Hu. 2014. "Balancing Exploitation and Exploration in Discrete Optimization via Simulation Through a Gaussian Process-Based Search". *Operations Research* 62 (6): 1416–1438.

Yin, J., S. Ng, and K. Ng. 2011. "Kriging Metamodel with Modified Nugget-effect: The Heteroscedastic Variance Case". *Computers & Industrial Engineering* 61 (3): 760–777.

AUTHOR BIOGRAPHIES

QUN MENG received the B.E. degree in industrial engineering and logistics management from Shanghai Jiao Tong University, Shanghai, China, in 2012. She is currently working towards the Ph.D. degree at the Department of Industrial and Systems Engineering, National University of Singapore. Her research interests are in the area of simulation optimization. Her email address is mengqun@u.nus.edu.

SZU HUI NG is an Associate Professor in the Department of Industrial and Systems Engineering at the National University of Singapore. She holds B.S., M.S., and Ph.D. degrees in Industrial and Operations Engineering from the University of Michigan. Her research interests include computer simulation modeling and analysis, design of experiments, and quality and reliability engineering. She is a member of IEEE and INFORMS and a senior member of IIE. Her email address is isensh@nus.edu.sg and her web page is http://www.ise.nus.edu.sg/staff/ngsh/index.html.