Proceedings of the 2012 Winter Simulation Conference
C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, eds.

EFFICIENT DISCRETE OPTIMIZATION VIA SIMULATION USING STOCHASTIC KRIGING

Jie Xu

George Mason University
Fairfax, VA 22030, USA

ABSTRACT

We propose to use a global metamodeling technique known as stochastic kriging to improve the efficiency of Discrete Optimization-via-Simulation (DOvS) algorithms. Stochastic kriging metamodel allows the DOvS algorithm to utilize all information collected during the optimization process and identify solutions that are most likely to lead to significant improvement in solution quality. We call the approach Stochastic Kriging for OPtimization Efficiency (SKOPE). In this paper, we integrate SKOPE with a locally convergent DOvS algorithm known as Adaptive Hyperbox Algorithm (AHA). Numerical experiments show that SKOPE significantly improves the performance of AHA in the early stage of optimization, which is very helpful for DOvS applications where the number of simulations for an optimization task is severely limited due to a short decision time window and time-consuming simulation.

1 INTRODUCTION

In the past decade, there has been a fast growing body of literature on how to optimize the design of a system using a simulation model. We refer to such problems as Optimization via Simulation (OvS). See Fu (2002) and Fu, Glover, and April (2005) for a review of OvS. When the decision variables are discrete valued, such as the stocking levels of products in a multi-product inventory management problem, we refer to it as Discrete Optimization via Simulation (DOvS).

When there are only a few hundreds feasible solutions, ranking-and-selection algorithms can be applied to choose the best solution. Examples include the indifference-zone procedure of Nelson et al. (2001) and Kim and Nelson (2001), the Bayesian procedure of Chick and Inoue (2001), and the Optimal Computing Budget Allocation (OCBA) procedure of Chen et al. (2000). Most recently, Frazier (2012) proposed an indifference-zone procedure for more than 15,000 alternatives. However, in a typical DOvS problem, the feasible solution space often includes millions and even billions of feasible solutions and thus ranking-and-selection procedures are not directly applicable.

Adaptive random search has been the dominant paradigm for designing DOvS algorithms when the solution space is large. Most existing DOvS algorithms focus on asymptotic global convergence, including the stochastic ruler algorithm of Yan and Mukai (1992), the simulated annealing algorithm of Alrefaei and Andradóttir (1999), and the nested partitions algorithm of Shi and Ólafsson (2000) and Pichitlamken and Nelson (2003). These algorithms essentially have to visit every solution to guarantee global convergence and lack the efficiency necessary to solve real world problems.

Another class of adaptive random search DOvS algorithms guarantees convergence to a local optimal solution. Andradóttir (1995) proposed a locally convergent algorithm for one-dimensional DOvS problems. The COMPASS algorithm of Hong and Nelson (2006) and the AHA algorithm of Xu, Hong, and Nelson (2011) also belong to this class of DOvS algorithms, but they are not restricted to one dimension. By focusing on finding a local optimum, these algorithms can efficiently search the solution space and achieve good finite-time performance. Xu, Hong, and Nelson (2010) proposed a framework for locally convergent DOvS algorithms with a global search phase as a "multi-start" mechanism for COMPASS/AHA and developed a software package known as Industrial Strength COMPASS (ISC) (www.iscompass.net). Numerical ex-

periments showed that ISC has finite time performance comparable to a popular commercial DOvS solver while providing theoretical guarantees that commercial products lack.

Despite the initial success of ISC, DOvS applications are still limited to small-scale problems where a large number of simulations can be expended for one optimization. However, in many important real world problems, the limited decision time window and the large-scale simulation model mean that only a small number of simulations (e.g., within 1000) are possible for each optimization, which is far from enough for any existing DOvS algorithm to achieve any real progress, let alone converge to an optimal solution. In such context, a decision maker is more interested in the efficiency of a DOvS algorithm, i.e., the algorithm is able to very quickly find solutions with good quality using a small number of simulation replications, than the algorithm's eventual convergence to an optimal solution.

In this paper we propose to use a global metamodeling technique known as stochastic kriging (Ankenman, Nelson, and Staum 2010) to improve the efficiency of adaptive random search DOvS algorithms. At a high level, we construct a stochastic kriging global metamodel using previously visited solutions and design points spread across the search space. Compared to polynomial type response surface models, a stochastic kriging metamodel can quantify prediction uncertainty globally and leverage all information collected during the optimization process. Using the metamodel, we choose solutions that are more likely to lead to improvement to be simulated and thus improve optimization efficiency. We call this approach Stochastic Kriging for OPtimization Efficiency (SKOPE). SKOPE can be easily integrated with existing DOvS algorithms such as COMPASS and AHA to improve their efficiency and at the same time maintain their local convergence property. We report an implementation using AHA because of its simplicity and excellent empirical performance. Stochastic kriging has been successfully used in estimating expected shortfall in a nested simulation paradigm (Liu and Staum 2010). As to the best of our knowledge, this is the first time stochastic kriging is used for DOvS.

Kriging has been used previously to optimize stochastic black-box systems (Huang et al. 2006). Our work is fundamentally different from Huang et al. (2006) in several significant ways. First, their kriging model does not consider the intrinsic noise in simulation and they assumed that the stochastic noise is IID across all solutions, which is not true in most simulation models. Second, their algorithm dose not offer any convergence guarantee. Finally, the way we use the kriging metamodel is simpler and better captures prediction uncertainty than their method, which involves a nonlinear search for the maximum expected improvement value over all solutions.

The Gaussian Process-based Search (GPS) (Sun, Hong, and Hu 2011) is in a similar spirit as ours. GPS constructs a Gaussian process model and uses it as a sampling distribution to search the solution space. The sampling distribution takes into account both model uncertainty and simulation noise and balance exploring less visited solution space and improving the simulation estimates of already visited solutions. GPS does not require correlation matrix inversions as stochastic kriging does, which may be numerically unstable. However, GPS lacks the MSE-optimal predictive power of stochastic kriging. Therefore, it may not be able to identify good solutions as efficiently as stochastic kriging. GPS also has to work with a complex sampling distribution. In comparison, SKOPE is designed to facilitate rapid progress in the early stage of DOvS for applications where simulation is very time-consuming and thus the computation cost of matrix inversion is negligible.

The rest of the paper is organized as follows. In Section 2, we provide a brief introduction to AHA and stochastic kriging. In Section 3, we discuss the design issues of SKOPE and how it can be integrated with AHA, and prove the local convergence property. In Section 4, we report numerical experiments comparing the performance of AHA with and without SKOPE. We give conclusions and discuss future research directions in Section 5.

2 BACKGROUND

2.1 AHA for DOvS

In a DOvS problem, we want to find a D-dimensional integer decision variable $\mathbf{x}$ to optimize the performance of a system modeled via stochastic simulation. We will work on minimization problems in this paper. The performance of the system $G(\mathbf{x})$ is a random variable and can be independently sampled by running stochastic simulations. We assume that the sample mean of the simulation observations is a strongly consistent estimator of $g(\mathbf{x}) = \mathrm{E}[G(\mathbf{x})]$, as stated in Assumption 1.

Assumption 1 For all $\mathbf{x} \in \Theta$,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} G_j(\mathbf{x}) = g(\mathbf{x}) \quad \text{w.p. 1.}$$

Formally, we want to solve the problem

$$\text{Minimize} \quad g(\mathbf{x}) = \mathrm{E}\left[G(\mathbf{x})\right] \quad \text{subject to} \quad \mathbf{x} \in \Theta = \Phi \cap \mathscr{Z}^D, \tag{1}$$

where Φ is convex and compact and $\mathscr{Z}^D$ denotes the D-dimensional integer lattice.

Following Hong and Nelson (2006), we define a *local minimum* of Problem 1 as follows:

Definition 1 Let $\mathscr{N}(\mathbf{x}) = \{\mathbf{y} : \mathbf{y} \in \Theta \text{ and } \|\mathbf{x}-\mathbf{y}\| = 1\}$ be the local neighborhood of $\mathbf{x} \in \Theta$, where $\|\mathbf{x}-\mathbf{y}\|$ denotes the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$. Then $\mathbf{x}$ is a *local minimum* if $\mathbf{x} \in \Theta$ and either $\mathscr{N}(\mathbf{x}) = \emptyset$ or $g(\mathbf{x}) \le g(\mathbf{y})$ for all $\mathbf{y} \in \mathscr{N}(\mathbf{x})$. Let $\mathscr{M}$ denote the set of local minimizers of the function g in Θ.

AHA is an iterative adaptive random search algorithm and has two essential components: a sampling scheme and an estimation scheme. At iteration k, AHA randomly samples M_k feasible solutions (duplicates possible) from a hyperbox-shaped Most Promising Area (MPA) $\mathscr{H}_k \subseteq \Theta$ according to a sampling distribution F_k defined on $\mathscr{H}_k$. Xu, Hong, and Nelson (2011) used the uniform distribution defined on $\mathscr{H}_k$ as F_k. Denote the set of unique sampled solutions as $\mathscr{S}_k$ and the set of all sampled solutions through iteration k as $\mathscr{S}(k)$. The estimation scheme chooses a subset of solutions $\mathscr{E}_k \subseteq \mathscr{S}(k)$, and allocates $a_k(\mathbf{x})$ additional simulation observations to all $\mathbf{x} \in \mathscr{E}_k$. Let $a_k(\mathbf{x}) = 0$ for all $\mathbf{x} \notin \mathscr{E}_k$. Then the total number of simulation observations $\mathbf{x}$ has received up to iteration k is $N_k(\mathbf{x}) = \sum_{i=0}^{k} a_i(\mathbf{x})$. The cumulative sample mean of solution $\mathbf{x}$ is $\bar{G}_k(\mathbf{x}) = \sum_{j=1}^{N_k(\mathbf{x})} G_j(\mathbf{x})/N_k(\mathbf{x})$ if $N_k(\mathbf{x}) > 0$, where $G_j(\mathbf{x})$ is the jth observation of $G(\mathbf{x})$.

Figure 1 illustrates how AHA works in a two-dimensional example. The initial MPA is the feasible solution space (the bold rectangle) and the current best solution $\mathbf{x}_0$ is the user provided initial feasible solution. In the first iteration, AHA sampled $\mathbf{x}_{11}$ and $\mathbf{x}_{12}$. AHA simulates $\mathbf{x}_0$, $\mathbf{x}_{11}$ and $\mathbf{x}_{12}$ and determines $\mathbf{x}_0$ is the current sample best solution. The new MPA is then the largest rectangle that contains the current sample best solution $\mathbf{x}_0$ in its interior, but not any of the previously sampled solutions. Figure 1 plots two more iterations. Formally, AHA is described below

Algorithm 1 Adaptive Hyperbox

Step 0 Let $\mathbf{x}_0$ be the starting solution provided by the user. Set the iteration counter $k = 0$. Let $\mathscr{S}_0 = \mathscr{S}(0) = \{\mathbf{x}_0\}$ and $\widehat{\mathbf{x}}_0^* = \mathbf{x}_0$. Set $\mathscr{E}_0 = \{\mathbf{x}_0\}$. Determine $a_0(\mathbf{x}_0)$. Take $a_0(\mathbf{x}_0)$ observations from $\mathbf{x}_0$, set $N_0(\mathbf{x}_0) = a_0(\mathbf{x}_0)$, and calculate $\bar{G}_0(\mathbf{x}_0)$.

Step 1 Let $k = k+1$. Identify $\mathscr{H}_k$ (for k = 1, $\mathscr{H}_k = \Theta$). Sample $\mathbf{x}_{k1}, \mathbf{x}_{k2}, \ldots, \mathbf{x}_{km}$ independently from $\mathscr{H}_k$ using a sampling distribution F_k defined on $\mathscr{H}_k$. Remove any duplicates from $\mathbf{x}_{k1}, \mathbf{x}_{k2}, \ldots, \mathbf{x}_{km}$ and let $\mathscr{S}_k$ be the remaining set. Let $\mathscr{S}(k) = \mathscr{S}(k-1) \cup \mathscr{S}_k$.

Step 2 Let $\mathscr{E}_k = \mathscr{S}_k \bigcup \{\widehat{\mathbf{x}}_{k-1}^*\}$. For all $\mathbf{x} \in \mathscr{E}_k$, take $a_k(\mathbf{x})$ simulation observations and update $N_k(\mathbf{x})$ and $\bar{G}_k(\mathbf{x})$. For all $\mathbf{x} \notin \mathscr{E}_k$, let $N_k(\mathbf{x}) = N_{k-1}(\mathbf{x})$ and $\bar{G}_k(\mathbf{x}) = \bar{G}_{k-1}(\mathbf{x})$.

Step 3 Let $\widehat{\mathbf{x}}_k^* = \arg\min_{\mathbf{x}\in\mathscr{E}_k} \bar{G}_k(\mathbf{x})$. Go to Step 1.

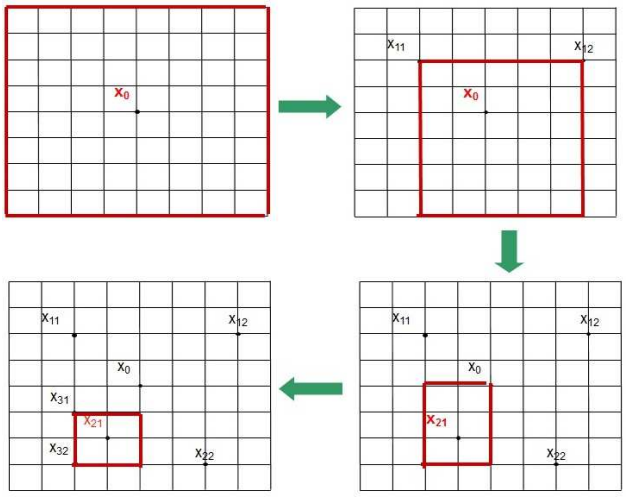

Figure 1: A two-dimensional example for AHA.

Under Assumption 1, Xu, Hong, and Nelson (2011) proved that the infinite sequence $\{\widehat{\mathbf{x}}_0^*, \widehat{\mathbf{x}}_1^*, \ldots\}$ generated by AHA converges with probability 1 to the set $\mathscr{M}$ in the sense that $\Pr\{\widehat{\mathbf{x}}_k^* \notin \mathscr{M} \text{ infinitely often}\} = 0$ when AHA satisfies the following conditions on the sampling and estimation schemes:

Condition 1 : The sampling scheme satisfies the following requirement:

The sampling distribution F_k guarantees that $\Pr\{\mathbf{x} \in \mathscr{S}_k\} \geq \varepsilon$ for all $\mathbf{x} \in \mathscr{N}(\widehat{\mathbf{x}}_{k-1}^*)$ for some $\varepsilon > 0$ that is independent of k.

Condition 2 : The estimation scheme satisfies the following requirements:

1. $\mathscr{E}_k$ is a subset of $\mathscr{S}(k)$;
2. $\mathscr{E}_k$ contains $\widehat{\mathbf{x}}_{k-1}^*$ and $\mathscr{S}_k$;
3. $a_k(\mathbf{x})$ is allocated such that $\min_{\mathbf{x}\in\mathscr{E}_k} N_k(\mathbf{x}) \geq 1$ for all $k = 1, 2, \ldots$ and $\min_{\mathbf{x}\in\mathscr{E}_k} N_k(\mathbf{x}) \to \infty$ w.p. 1 as $k \to \infty$.

2.2 Stochastic Kriging

Stochastic kriging (Ankenman, Nelson, and Staum 2010) is a simulation metamodeling technique based on the well-known kriging technique widely used in spatial statistics. Stochastic kriging explicitly models the intrinsic noise in stochastic simulation output. As a result, unlike kriging, the interpolation surface of a stochastic kriging model almost always does not pass through experiment points and is not an "exact" interpolation. Compared to polynomial based simulation metamodeling techniques (Barton and Meckesheimer 2006), stochastic kriging is able to capture uncertainty globally.

Stochastic kriging models $g(\mathbf{x})$ as

$$g(\mathbf{x}) = \beta_0 + \mathrm{M}(\mathbf{x}) \tag{2}$$

where β_0 is a constant representing the overall surface mean, M is a stationary Gaussian random field with mean 0. A more general trend term can be used to model the surface mean. But in kriging literature, it has been shown that a constant overall surface mean works very well for a large variety of applications. Using a Gaussian random field M in the model captures the uncertainty before running simulations on $\mathbf{x}$ and is referred to as *extrinsic uncertainty* in Ankenman, Nelson, and Staum (2010). The correlation function reflects the "similarity" between solutions that are spatially close to each other. We adopt a commonly

used model with a a secondary-order stationary Gaussian correlation function

$$\mathrm{Cov}\left[\mathrm{M}(\mathbf{x}),\mathrm{M}(\mathbf{x}')\right]=\tau^2\exp\left(-\sum_{d=1}^{D}\theta_d(x_d-x_d')^2\right).$$

The variance of $\mathrm{M}(\mathbf{x})$ for all $\mathbf{x}'$ is given by τ^2 and the amount of correlation depends only on the distance between $\mathbf{x}$ and $\mathbf{x}'$, with different weights $\theta=[\theta_1,\ldots,\theta_D]^T$. The jth simulation observation of $G(\mathbf{x})$ is then modeled as

$$G_j(\mathbf{x})=\beta_0+\mathrm{M}(\mathbf{x})+\varepsilon_j(\mathbf{x}), \tag{3}$$

where the simulation noise $\varepsilon_j(\mathbf{x})$, referred to as "intrinsic uncertainty" in Ankenman, Nelson, and Staum (2010), is I.I.D. normal distributed with a mean 0 and variance $\mathrm{V}(\mathbf{x})$, independent of $\mathrm{M}(\mathbf{x})$. After $N(\mathbf{x})$ simulations, the point estimator of $g(\mathbf{x})$ is the sample mean $\bar{G}(\mathbf{x})=\sum_{j=1}^{N(\mathbf{x})}G_j(\mathbf{x})/N(\mathbf{x})$.

Let $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_L$ be the design points. Let $g_L=[g(\mathbf{x}_1),g(\mathbf{x}_2),\ldots,g(\mathbf{x}_L)]^T$, $\bar{G}_L=[\bar{G}(\mathbf{x}_1),\bar{G}(\mathbf{x}_2),\ldots,\bar{G}(\mathbf{x}_L)]^T$, and Σ^{LL} be the covariance matrix of g_L. Also let Σ_ε be the covariance matrix for the simulation noises for all design points. In this paper, we assume that all simulations use independent random number streams since it has been shown that Common Random Numbers (CRN) may actually hurt the performance of stochastic kriging (Chen, Ankenman, and Nelson 2012). So Σ_ε is diagonal and $\Sigma_\varepsilon(i,i)=\mathrm{V}(\mathbf{x}_i)/N_k(\mathbf{x}_i)$. We denote $\Sigma_\varepsilon=VN^{-1}$ where V and N are diagonal matrices and the (i,i)-th elements are $\mathrm{V}(\mathbf{x}_i)$ and $N_k(\mathbf{x}_i)$, respectively. Denote the covariance matrix of $\bar{G}_L$ as $\Sigma=\Sigma^{LL}+\Sigma_\varepsilon$. Let the prediction points be $\mathbf{x}^1,\mathbf{x}^2,\ldots,\mathbf{x}^S$. Let $g^S=[g(\mathbf{x}^1),g(\mathbf{x}^2),\ldots,g(\mathbf{x}^S)]^T$. Denote the $L\times S$ covariance matrix between g_L and g^S as Σ^{LS}, and its transpose as G^{SL}. Further denote the covariance matrix of g^S as Σ^{SS}. Ankenman, Nelson, and Staum (2010) show that the joint distribution of $\bar{G}_L$ and g^S is multivariate normal

$$\begin{pmatrix}g^S\\ \bar{G}_L\end{pmatrix}\sim\mathrm{MVN}\left[\beta_0\mathbf{1}_{L+S},\begin{pmatrix}\Sigma^{SS}&\Sigma^{SL}\\ \Sigma^{LS}&\Sigma\end{pmatrix}\right], \tag{4}$$

where $\mathbf{1}_{L+S}$ is a $L+S$ column vector of ones. Given $\bar{G}_L=\bar{G}$, the conditional distribution of g^S is multivariate normal and its mean is the MSE-optimal stochastic kriging predictor

$$\widehat{G}^S=\beta_0\mathbf{1}_S+\Sigma^{SL}\Sigma^{-1}\left(\bar{G}-\beta_0\mathbf{1}_k\right), \tag{5}$$

and the covariance matrix is

$$\widehat{\Sigma}^{SS}=\Sigma^{SS}-\Sigma^{SL}\Sigma^{-1}\Sigma^{LS}. \tag{6}$$

The unknown parameters $\beta_0,\tau^2,\theta_1,\cdots,\theta_D$ are estimated via maximal likelihood estimation. Sample variances are used to estimate $\mathrm{V}(\mathbf{x}_1),\cdots,\mathrm{V}(\mathbf{x}_L)$ and Ankenman, Nelson, and Staum (2010) showed that doing so does not introduce any prediction bias.

3 SKOPE PROCEDURE

The SKOPE procedure provides a specific implementation of the sampling scheme of AHA. On iteration k, instead of uniformly randomly sampling solutions from the MPA $\mathscr{H}_k$, SKOPE builds a stochastic kriging metamodel to help predict solution quality inside $\mathscr{H}_k$ and select the "most promising" solutions to be included in the sampling set $\mathscr{S}_k$.

3.1 Experiment Design

It is common to use a space-filling design such as the Latin Hypercube Design (LHD) when one uses stochastic kriging or ordinary kriging to construct a metamodel (Santner, Williams, and Notz 2003). On iteration k, when AHA constructs the hyperbox MPA $\mathscr{H}_k$, we apply the Matlab function lhsdesign() to

create a Latin hypercube design that maximizes the minimal distance between any pair of design points inside $\mathscr{H}_k$. The more design points we have, the more accurate the metamodel will be. However, more design points also means there will be less simulation budget remaining for future AHA iterations. How to balance this trade-off and choose the number of design points (denoted as L_k) is a very challenging research question. A simple rule of thumb (Jones, Schonlau, and Welch 1998) suggests that there should be 10 design points per dimension. We will follow this rule up to $D = 10$. We also require that $L_k \leq 0.05V_k$, where V_k is the number of solutions in $\mathscr{H}_k$. This is to avoid using unnecessarily many design points when AHA has made $\mathscr{H}_k$ small enough. In addition, it also helps avoid numerical problems when stochastic kriging calculates the inverse of the correlation matrix for design points.

Experiment design for SKOPE has its unique challenges. As AHA searches the solution space, we will have simulation observations for all previously sampled solutions $\mathbf{x} \in \mathscr{S}(k)$, in addition to the L_k design points inside $\mathscr{H}_k$. As simulations are expensive, we want to fully utilize information available and also include $\mathscr{S}(k)$ or a subset of it as design points. However, as $\mathbf{x} \in \mathscr{S}(k)$ are not spread evenly and they tend to concentrate around $\mathscr{H}_k$, using all $\mathbf{x} \in \mathscr{S}(k)$ as design points will likely make the correlation matrix almost singular. It will be an interesting research issue to study how to select a subset of $\mathscr{S}(k)$ to maximally increase the predictive power of the metamodel without causing numerical problems. In this paper, we will adopt the simple strategy of using all $\mathbf{x} \in \mathscr{S}(k)$ and the L_k LHD points inside $\mathscr{H}_k$. If there is a numerical problem with computing the inverse of the correlation matrix, SKOPE will only use the L_k LHD points.

The maxmin LHD only supports hyperbox-shaped design space, which may not be the case if the original DOvS problem constraints involve more than boundary constraints for decision variables. It is possible to approximately use an LHD in a convex design space as explained in Liu and Staum (2010). But this is not essential to the discussion of SKOPE and we will test SKOPE on problems with only boundary constraints.

3.2 Selecting the Sampling Set

The goal of SKOPE is to use the stochastic kriging metamodel to help AHA select a sampling set that has average solution quality higher than what a simpler sampling scheme like the uniform random distribution is able to achieve. There are multiple ways to utilize a kriging metamodel to identify a good solution to simulate. Huang et al. (2006) used ordinary kriging for optimizing stochastic black-box systems. For a minimization problem, they introduced a utility function $u(\mathbf{x})$ for a solution $\mathbf{x}$ not sampled yet

$$u(\mathbf{x}) = -\widehat{G}_o(\mathbf{x}) - c\sigma_o(\mathbf{x}), \tag{7}$$

where $\widehat{G}_o(\mathbf{x})$ is the ordinary kriging predictor for $G(\mathbf{x})$, $\sigma_o(\mathbf{x})$ is its standard deviation, and c is a user chosen constant. In their study, they chose $c = 1$, i.e., they are willing to trade one unit of the predicted objective value for one unit of prediction uncertainty as measured by the standard deviation $\sigma_o(\mathbf{x})$. Then they determine the "effective best" solution $\mathbf{x}^{**}$ among all of the n previously sampled solutions $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ as

$$\mathbf{x}^{**} = \arg\max_{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n} u(\mathbf{x}). \tag{8}$$

They then computed the Expected Improvement (EI) function for each $\mathbf{x}$ defined as

$$\mathrm{E}[I(\mathbf{x})] \equiv \mathrm{E}\left[\max\left(\widehat{G}_o(\mathbf{x}^{**}) - G(\mathbf{x}), 0\right)\right] \cdot \left(1 - \frac{\sqrt{\mathrm{V}(\mathbf{x})}}{\sqrt{\mathrm{V}(\mathbf{x}) + \sigma_o(\mathbf{x})^2}}\right), \tag{9}$$

where the expectation is taken with respect to the conditional distribution of $G(\mathbf{x}) \sim N(\widehat{G}(\mathbf{x}), \sigma_o(\mathbf{x})^2)$. In the right hand side of (9), the first term gives the expected improvement in objective value from the current effective best solution. The second term attempts to balance the intrinsic uncertainty $\mathrm{V}(\mathbf{x})$ and extrinsic

uncertainty $\sigma_o(\mathbf{x})^2$ in a heuristic manner. They assumed that $\mathrm{V}(\mathbf{x})$ is the same for all $\mathbf{x}$, which is an unrealistic assumption in most simulation applications. They then maximized EI across the solution space using the Nelder-Mead simplex algorithm.

Although we could adopt their approach and use stochastic kriging to replace the ordinary kriging model and modify the utility function $u(\mathbf{x})$ and $\mathrm{E}[I(\mathbf{x})]$ accordingly, we choose not do to so for the following reasons. First, $u(\mathbf{x})$ is only a heuristic way to measure the overall quality of $\mathbf{x}$. There is no justifiable way to choose the parameter c, which has a significant impact on the performance of the procedure. Second, the $\mathrm{E}[I(\mathbf{x})]$ maximization problem introduces another nonlinear optimization problem. Although it is deterministic, it still adds considerably to the complexity of the procedure as the formula of $\mathrm{E}[I(\mathbf{x})]$ would involve the stochastic kriging predictor and the cumulative distribution function of the normal distribution. Third, AHA is a population-based random search algorithm. If we need to select, say, 25 solutions in the sampling set, it would be extremely difficult to find solutions with the top 25 EI values. Finally, the EI approach only uses marginal distributions for $G(\mathbf{x})$ and fails to fully leverage the information available in the multivariate normal distribution of G^S.

We propose to address this problem using a combination of Ordinal Optimization (OO) (Ho, Zhao, and Jia 2007) and Monte Carlo simulation. We first present some notations. Recall that V_k is the number of solutions inside the MPA $\mathscr{H}_k$ on iteration k. Let T_k be the set of good enough solutions among all V_k solutions (typically the top solutions), S_k be the set of prediction points for the stochastic kriging model, and s_k be the size of S_k.

We first blindly pick enough prediction points via uniform random sampling inside $\mathscr{H}_k$ to ensure that with a large probability γ_1, a good enough solution $\mathbf{x} \in T_k$ is in S_k, i.e., $\Pr(|T_k \bigcap S_k| \geq 1) \geq \gamma_1$. When we uniformly randomly pick S_k from $\mathscr{H}_k$, the probability is given by

$$\Pr(|T_k \bigcap S_k| \geq 1) = 1 - \binom{V_k - T_k}{s_k} \Big/ \binom{V_k}{s_k}. \tag{10}$$

We can then search for the smallest s_k such that (10) is at least γ_1. We set $\gamma_1 = 0.95$ in our numerical experiments. When V_k is large, we set $T_k = V_k/10000$ such that s_k will not be too big. We also require that $s_k \leq \min(0.1V_k, 5000)$ to avoid memory usage issue and numerical problem of inverting the correlation matrix when S_k fills $\mathscr{H}_k$ too densely.

We then generate Monte Carlo samples from the conditional distributions of the unknown objective values of S_k to rank these prediction points and select a subset of S_k that with a large probability γ_2 includes the best solution in S_k in the sampling set $\mathscr{S}_k$. We let $p(\mathbf{x})$ be the probability that $\mathbf{x}$ is the best solution in S_k. We then rank $\mathbf{x} \in S_k$ in descending order of $p(\mathbf{x})$ and select the first $S \equiv \min\{S : \Sigma_{s=1}^{S} p(\mathbf{x}_s) > \gamma_2\}$ solutions to be included in $\mathscr{S}_k$. We let $\gamma_2 = 0.9$ in our numerical experiments. We also impose an upper bound $S = 25$. We admit this is an arbitrary choice. However, we need to balance effort spent in the current iteration and future iterations and it is extremely challenging to analyze this tradeoff. Indeed, if including 25 solutions is not enough to ensure that with probability γ_2, the best solution in S_k is included in $\mathscr{S}_k$, it probably means the metamodel either cannot predict the objective values of the prediction points with reasonable certainty or the objective values of these prediction points are close. In either case, AHA can most likely do just as well by selecting the best $S = 25$ as by selecting a lot more of the prediction points to be included in $\mathscr{S}_k$.

In the Monte Carlo simulation procedure, we generate J samples from the multivariate normal distribution with mean and covariance matrix given in (5) and (6). For the jth sample of g^S, denoted as G_j^S, we let $I_s(j)$ be the indicator function for prediction point $\mathbf{x}_s, s = 1, 2, \ldots, s_k$. When $\mathbf{x}_s$ is the top solution (including a tie) in this sample G_j^S, $I_s(j) = 1$; otherwise $I_s(j) = 0$. The probability that $\mathbf{x}_s$ is the best solution in the prediction set S_k is then estimated as $\widehat{p}_s = \sum_{j=1}^{J} I_s(j)/J$. We want J to be large enough to have good estimates. Since this procedure is typically much faster than the actual simulation, we use $J = 1000$ in our numerical study.

Another practical issue is the uniform random sampling procedure may generate duplicate solutions among the s_k sampled solutions. We may keep sampling until we have s_k unique prediction points. However, during the early stage of DOvS, the design space is typically quite large and the number of duplicate solutions is negligible. Also the computation overhead of doing so probably outweighs its benefit. We thus will skip this issue and just uniformly sample the MPA $\mathscr{H}_k$ to generate s_k solutions and keep the unique ones as the prediction set S_k.

3.3 The AHA-SKOPE Procedure

We are now ready to describe the SKOPE procedure implemented in the AHA framework to improve the sampling scheme of AHA. We assume that the feasible solution space is a hyperbox so that LHD can be directly applied.

Algorithm 2 AHA-SKOPE

Step 0 Same as Algorithm 1.

Step 1.1 Let $k = k+1$. Identify $\mathscr{H}_k$ (for k = 1, $\mathscr{H}_k = \Theta$). Use the max-min LHD to fill $\mathscr{H}_k$ with a design point set L_k. For each $\mathbf{x} \in L_k$, take $a_k(\mathbf{x})$ simulation observations and update $N_k(\mathbf{x})$ and $\bar{G}_k(\mathbf{x})$.

Step 1.2 Determine s_k according to (10). Using a uniform random distribution defined on $\mathscr{H}_k$, sample s_k solutions (excluding the design set L_k), remove any duplicates and obtain the prediction set S_k.

Step 1.3 Use the Monte Carlo simulation procedure in Section 3.2 to select at most S solutions from S_k. Let S'_k be the remaining set. Let $\mathscr{S}_k = L_k \bigcup S'_k$ and let $\mathscr{S}(k) = \mathscr{S}(k-1) \cup \mathscr{S}_k$.

Step 2 Same as Algorithm 1.

Step 3 Same as Algorithm 1.

We have the following proposition to show that AHA-SKOPE maintains the local convergence property of AHA under Assumption 1.

Proposition 1 Let $\widehat{\mathbf{x}}^*_k, k = 0,1,2,\ldots$ be a sequence of solutions generated by Algorithm 2 when applied to Problem (1). Suppose that Assumption 1 is satisfied. Then $\Pr\{\widehat{\mathbf{x}}^*_k \notin \mathscr{M} \text{ i.o.}\} = 0$.

Proof: Because AHA-SKOPE only modifies the sampling scheme, we only need to verify that AHA-SKOPE satisfies Condition 1. To do so, we need to compute $\Pr\{\mathbf{x} \in \mathscr{S}_k\}$ for all $\mathbf{x} \in \mathscr{N}(\widehat{\mathbf{x}}^*_{k-1})$. We first notice that $\mathscr{N}(\widehat{\mathbf{x}}^*_{k-1}) \subseteq \mathscr{H}_{k-1}$ by construction (Xu, Hong, and Nelson 2011). Denote the s_k prediction points independently and uniformly sampled within $\mathscr{H}_{k-1}$ as $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{s_k}$. For all $\mathbf{x} \in \mathscr{N}(\widehat{\mathbf{x}}^*_{k-1})$, we have $\Pr\{\mathbf{x}_1 = \mathbf{x}_2 = \cdots = \mathbf{x}_{s_k} = \mathbf{x}\} = \Pr\{\mathbf{x}_1 = \mathbf{x}\}^{s_k} = (1/V_{k-1})^{s_k} \geq (1/|\Theta|)^{|\Theta|}$. So there is a nonzero probability that S_k only contains $\mathbf{x}$ and thus $\mathbf{x} \in S'_k$. By construction, $\mathbf{x} \in \mathscr{S}_k$. So Condition 1 is satisfied. □

4 Numerical Experiments

Xu, Hong, and Nelson (2011) compared AHA to another locally convergent DOvS algorithm COMPASS, which has been tested against a leading commercial OvS solver (Xu, Hong, and Nelson 2010) via extensive numerical experiments and shown to be highly competitive. AHA's performance has been shown to be much better than that of COMPASS when dimension is high (e.g., ≥ 10) and slightly worse than that of COMPASS for lower dimensional problems. Because these studies have established the competitiveness of AHA, we will focus on comparing AHA-SKOPE with AHA.

Because the current implementation of AHA-SKOPE can only deal with boundary constraints, we will only use two of the test problems in Xu, Hong, and Nelson (2010) with hyperbox constraints only. We provide brief descriptions of these test problems and details can be found in Xu, Hong, and Nelson (2010).

The first test problem is the *singular function* of Hong and Nelson (2006):

$$g_1(x_1,x_2,x_3,x_4) = (x_1+10x_2)^2 + 5(x_3-x_4)^2 + (x_2-2x_3)^4 + 10(x_1-x_4)^4 + 1. \tag{11}$$

When only integer solutions are considered, this function has three local minima: $(0,0,0,0)$ with $g_1(0,0,0,0) = 1$; $(1,0,0,1)$ with $g_1(1,0,0,1) = 7$; and $(-1,0,0,-1)$ with $g_1(-1,0,0,-1) = 7$. We add normally distributed noise with zero mean and standard deviation $\max(\sqrt{g_1(x_1,x_2,x_3,x_4)},30)$ to make it a stochastic optimization problem. The feasible solution space is $-50 \le x_i \le 50, x_i \in \mathscr{Z}^+, i = 1,2,3,4$ with 104,060,401 feasible solutions.

The second test problem is the *high-dimensional* test problem of Xu, Hong, and Nelson (2010):

$$g_2(x_1,x_2,\ldots,x_D) = -\beta \exp\left\{-\gamma \sum_{d=1}^{D} (x_d - \xi^*)^2\right\}, \tag{12}$$

where we set $\gamma = 0.001, \beta = 10000$ and $\xi^* = 0$. The response surface has the shape of an inverted multivariate normal density function with a single globally optimal solution at $\mathbf{x} = (\xi^*, \xi^*, \ldots, \xi^*)$. We experiment with $D = 10$ and the feasible region is $-15 \le x_d \le 15, d = 1,\ldots,D$, with a total of 8,196,282,869,808,010 feasible solutions. Normally distributed noise with standard deviation $0.3 \times |g_2(\mathbf{x})|$ is added.

We run AHA and AHA-SKOPE on these test problems for 5 trails each. For the singular function, the initial solution is $\mathbf{x}_0 = (-30,-30,-30,-30)$; for the high-dimensional function, $\mathbf{x}_0 = (12, 12, 12, 12, 12, 12, 12, 12, 12, 12)$. For AHA, we set the number of solutions sampled per iteration close to what AHA-SKOPE sampled each iteration. For the singular function, we set it to 50; for the high-dimensional test function, it is 100. We use the same simulation budget allocation rule as in Xu, Hong, and Nelson (2011) to determine $a_k(\mathbf{x})$. AHA converges asymptotically to a local optimal solution. But in numerical experiments, we have to stop AHA and AHA-SKOPE in finite time. We apply a local optimality test (Xu, Hong, and Nelson 2011) when $\mathscr{H}_k$ contains only one interior solution $\widehat{\mathbf{x}}_k^*$, and stop AHA/AHA-SKOPE when the test finds $\widehat{\mathbf{x}}_k^*$ to be the local optimal solution with a statistical guarantee.

Figure 2 and 3 are the performance plots for the singular function and the high-dimensional test function with $D = 10$, respectively. Since we know the true objective value of a solution, we measure the progress of the algorithm using the true objective value of the current sample best solution. The curves are the average of 5 independent AHA/AHA-SKOPE runs. As can be seen clearly from the plot, AHA-SKOPE achieves much faster progress than AHA alone early on in the optimization progress, demonstrating the value of SKOPE when the DOvS application has a very limited simulation budget and the optimization process has to be stopped well before the local optimal solution can be found.

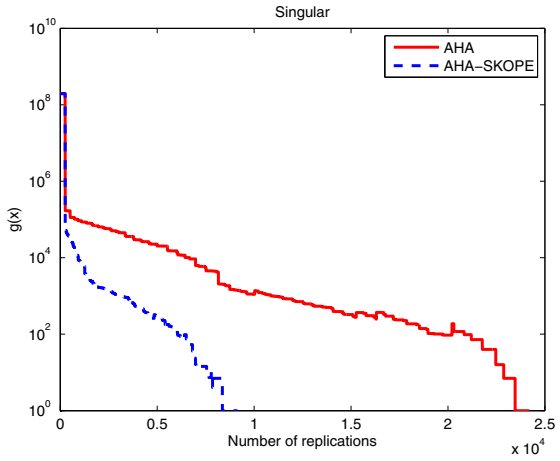

Figure 2: Performance plot for the singular function.

We have found in numerical experiments that there are two major factors hindering the effectiveness of SKOPE. As the MPA $\mathscr{H}_k$ becomes smaller, which can happen very quickly due to the efficiency of AHA in cutting down the search space (Xu, Hong, and Nelson 2011), we have to limit the number of design

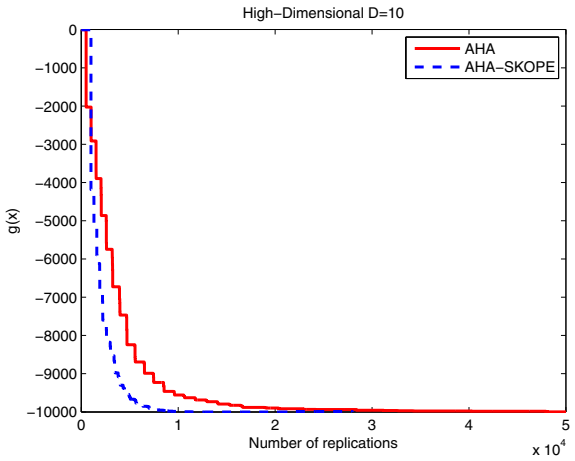

Figure 3: Performance plot for the high-dimensional test function $D = 10$.

points and prediction points to avoid numerical problems when computing the inverse of correlation matrix in the stochastic kriging step. Our approach now is quite crude and impose a somewhat arbitrary bound on the number of design points and prediction points. We also only use LHD points inside $\mathscr{H}_k$ when using all $\mathbf{x} \in \mathscr{S}(k)$ as design points makes the design correlation matrix near singular. Undoubtedly, our approach is not making full use of available information to construct a metamodel with the best predictive power possible. This is an area that needs further research.

The response surface shape also has a big impact on the effectiveness of SKOPE. If the response surface over the MPA $\mathscr{H}_k$ is almost flat everywhere but has a narrow valley around the optimal solution(s), it will be very difficult to obtain a metamodel with the necessary predictive power to guide AHA. This is especially relevant in the early stage of optimization when $\mathscr{H}_k$ is huge. In addition, if the magnitude of differences in objective values are too big, the predictive capability of stochastic kriging is also impaired considerably. In all cases, a poor metamodel can mislead AHA and performs worse than a uniform random sampling distribution. It will also make AHA's performance highly volatile from trial to trial. A better implementation of stochastic kriging may help alleviate some of these problems. However, we feel that it can also be effectively addressed by a pre-processing step of DOvS. Through a combination of experience, expert input, first-order analytical model, and preliminary simulation experiments, the decision maker should and could remove a significant part of feasible solution space that are orders of magnitude inferior to good solutions.

5 CONCLUSIONS AND DISCUSSIONS

We propose a stochastic-kriging based SKOPE procedure as a sampling scheme for a class of locally convergent adaptive random search DOvS algorithm to improve their finite-time performance, especially in the early stage of optimization. We feel this is very important for many practical DOvS applications where simulation is very time-consuming and a very limited number of simulations can be expended for an optimization task. We integrate the SKOPE procedure with a locally convergent DOvS algorithm AHA and prove that AHA-SKOPE maintains the local convergence property of AHA. We show in numerical experiments that SKOPE leads to significant performance improvement, especially in the early stage of optimization.

To fully utilize the benefit of SKOPE, there remain several further theoretical and computational issues that warrant further research. The first question is how to choose design points. Since AHA and most other DOvS algorithms are an iterative process, there are many previously sampled solutions that can be used as design points. Although more design points in principle can reduce the prediction uncertainty

of the stochastic kriging metamodel, in the later stage of optimization, we tend to have many sampled solutions closely clustered together, which makes the design point correlation matrix very close to singular. Therefore, it is very important to select a subset of visited solutions that will not make the correlation matrix near singular while reducing the prediction uncertainty of the metamodel maximally. We have restricted our attention to improving the sampling scheme of AHA using SKOPE and used the same simulation budget allocation rule as before. Ankenman, Nelson, and Staum (2010) discussed how to allocate simulation budget for a set of fixed design points to minimize the Integrated MSE (IMSE) of the metamodel. In the DOvS setting, IMSE probably is not the most useful measure of the utility of the metamodel because the metamodel should reduce prediction uncertainty for promising solutions while pay less attention to clearly inferior ones. When applied to higher dimensional problems, the current stochastic kriging implementation (www.stochastickriging.net) can have numerical difficulties. A more scalable implementation of stochastic kriging will also make SKOPE more applicable in high-dimensional DOvS applications.

ACKNOWLEDGMENTS

This article is based upon work supported by the National Science Foundation under Grant No. CMMI-1233376.

REFERENCES

Alrefaei, M. H., and S. Andradóttir. 1999. "A simulated annealing algorithm with constant temperature for discrete stochastic optimization". *Management Science* 45:748–764.

Andradóttir, S. 1995. "A method for discrete stochastic optimization". *Management Science* 41:1946–1961.

Ankenman, B., B. L. Nelson, and J. Staum. 2010. "Stochastic Kriging for Simulation Metamodeling". *Operations Research* 58:371–382.

Chen, C.-H., J. Lin, E. Ycesan, and S. E. Chick. 2000. "Simulation budget allocation for further enhancing the efficiency of ordinal optimization". *Discrete Event Dynamic Systems* 10:251–270.

Chen, X., B. Ankenman, and B. L. Nelson. 2012. "The Effects of Common Random Numbers on Stochastic Kriging Metamodels". *ACM Transactions on Modeling and Computer Simulation* 22:7:1–7:29.

Chick, S. E., and K. Inoue. 2001. "New two-stage and sequential procedures for selecting the best simulated system". *Operations Research* 49:732–743.

Frazier, P. 2012. "Indifference-Zone Ranking and Selection for More Than 15,000 Alternatives". Technical report, School of Operations Research and Information Engineering, Cornell University, Ithaca, New York.

Fu, M. C. 2002. "Optimization for simulation: Theory vs. practice". *INFORMS Journal on Computing* 14:192–215.

Fu, M. C., F. W. Glover, and J. April. 2005, December. "Simulation optimization: A review, new developments, and applications". In *Proceedings of the 2005 Winter Simulation Conference*, edited by M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, 83–95. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Ho, Y.-C., Q.-C. Zhao, and Q.-S. Jia. 2007. *Ordinal Optimization*. New York: Springer.

Hong, L. J., and B. L. Nelson. 2006. "Discrete optimization via simulation using COMPASS". *Operations Research* 54:115–129.

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* 34:441–466.

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

Kim, S., and B. L. Nelson. 2001. "A Fully Sequential Procedure for Indifference-Zone Selection in Simulation". *ACM Transactions on Modeling and Computer Simulation* 11:251–273.

Liu, M., and J. Staum. 2010. "Stochastic Kriging for Efficient Nested Simulation of Expected Shortfall". *Journal of Risk* 12:3–27.

Nelson, B. L., J. Swann, D. Goldsman, and W. Song. 2001. "Simple procedures for selecting the best simulated system when the number of alternatives is large". *Operations Research* 49:950–963.

Pichitlamken, J., and B. L. Nelson. 2003. "A combined procedure for optimization via simulation". *ACM Transactions on Modeling and Computer Simulation* 13:155–179.

Santner, T. J., B. J. Williams, and W. I. Notz. 2003. *The Design and Analysis of Computer Experiments*. New York: Springer.

Shi, L., and S. Ólafsson. 2000. "Nested partitions method for stochastic optimization". *Methodology and Computing in Applied Probability* 2:271–291.

Sun, L., L. J. Hong, and Z. Hu. 2011, December. "Optimization via Simulation Using Gaussian Process-Based Search". In *Proceedings of the 2011 Winter Simulation Conference*, edited by S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, 4134–4145. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Xu, J., L. J. Hong, and B. L. Nelson. 2010. "Industrial Strength COMPASS: A Comprehensive Algorithm and Software for Optimization via Simulation". *ACM Transactions on Modeling and Computer Simulation* 20:3:1–3:29.

Xu, J., L. J. Hong, and B. L. Nelson. 2011. "An Adaptive Hyperbox Algorithm for High-Dimensional Discrete Optimization via Simulation Problems". *INFORMS Journal on Computing* published online before print.

Yan, D., and H. Mukai. 1992. "Stochastic discrete optimization". *SIAM Journal on Control and Optimization* 30:594–612.

AUTHOR BIOGRAPHIES

JIE XU is an assistant professor in the Department of Systems Engineering and Operations Research at George Mason University. His research interests include Monte Carlo simulation, stochastic optimization, computational intelligence, and applications in risk management and aviation. He is a member of INFORMS, IIE, and IEEE. His email address is jxu13@gmu.edu.