Sequential Kriging Optimization Using Multiple Fidelity Evaluations

D. Huang, T. T. Allen, W. I. Notz, and R. A. Miller

Abstract

When cost-per-evaluation on a system of interest is high, surrogate systems can provide cheaper but lower-fidelity information. In the proposed extension of the Sequential Kriging Optimization method, surrogate systems are exploited to reduce the total evaluation cost. The method utilizes data on all systems to build a kriging meta-model that provides a global prediction of the objective function and a measure of prediction uncertainty. The location and fidelity level of the next evaluation are selected by maximizing an augmented expected improvement function, which is connected with the evaluation costs. The proposed method was applied to test functions from the literature and a metal-forming process design problem via Finite Element simulations. The method manifests sensible search patterns, robust performance, and appreciable reduction in total evaluation cost as compared to the original method.

Keywords: Multiple fidelity, Surrogate Systems, Kriging, Efficient Global Optimization, Computer Experiments

---

D. Huang, Scientific Forming Technologies Corporation, 5038 Reed Road, Columbus, Ohio 43220, U.S.A. Email: dhuang@deform.com.

T. T. Allen, Department of Industrial, Welding, and Systems Engineering, Ohio State University, 210 Baker Systems Building, 1971 Neil Avenue, Columbus, Ohio 43210, U.S.A. Email: allen.515@osu.edu.

W. I. Notz, Department of Statistics, Ohio State University, 404 Cockins Hall, 1958 Neil Avenue, Columbus, Ohio 43210, U.S.A. Email: win@stat.ohio-state.edu.

R. A. Miller, Department of Industrial, Welding, and Systems Engineering, Ohio State University, 210 Baker Systems Building, 1971 Neil Avenue, Columbus, Ohio 43210, U.S.A. Email: miller.6@osu.edu.

1. Introduction

The Sequential Kriging Optimization (SKO) method, also called the Efficient Global Optimization (EGO) method, has been developed in recent years for solving expensive noisy black-box problems (Jones et al. 1998, Sasena et al. 2002, Huang et al. 2005). These types of optimization problems arise in various areas, including large-scale circuit board design and manufacturing process improvement. For example, in a metal-forming shop, the manufacturer wants to adjust certain process parameters, such as the forming temperature or die speed, to maximize a system performance, such as the lifespan of the die. In such scenarios, the objective functions are often poorly-behaved, non-analytical, expensive to evaluate, and sometimes noisy.

The SKO method has its roots in the Bayesian Global Optimization method (Kushner 1964) and exploits the results from the Design and Analysis of Computer Experiments (DACE) area (Sacks et al. 1989a, 1989b). The basic concept of SKO is to approximate the objective with a kriging model and use it to indicate the most promising point for sequential sampling. Schonlau (1997) showed that SKO was able to efficiently locate the global optima of a number of classic test functions that were difficult for more traditional optimization methods to solve. Sasena et al. (2002) applied the SKO method for various practical problems in engineering design, and found that the method had good generality and efficiency on the whole.

Yet, sometimes the evaluation on the system of interest is so expensive that straightforward application of the SKO method might be too costly. In this case, one may consider drawing data with less cost from surrogate experimental systems. For instance, lab and pilot systems can be used to mimic production systems, and computer simulations can be used to approximate physical experiments. We call these systems "lower-fidelity systems", where the term "fidelity" relates to the extent to which a surrogate system can reproduce the input-output relationships of the system of interest. The system of interest, often a physical experiment or a high-resolution computer simulation, is called the "highest fidelity" system or the "real" system.

A variety of methods have been proposed for generating meta-models using variable fidelity data. Hutchinson et al. (1994) applied the so-called "correction response surface" model, where the difference or ratio between the low and high-fidelity systems are modeled with polynomials. To assess the difference or ratio, the high-fidelity points are usually a subset of the low-fidelity points. Similar concept was adopted by Watson et al. (1996), who used a neural

network to model differences between systems. Bandler et al. (1999) proposed an "input space-mapping" technique, where output of the low-fidelity system with the mapped input variables can accurately approximate that of the high-fidelity system. Leary et al. (2003) used the low fidelity data as prior knowledge to be incorporated in the training of neural networks and generation of kriging models. Kennedy et al. (2000) modeled the multi-fidelity outputs with auto-regressive Bayesian Gaussian stochastic processes. (Note that the Bayesian Gaussian stochastic process model, with proper prior assumptions, produces the same results as the kriging model.) Unlike the knowledge-based kriging by Leary et al. (2003), this model can conveniently handle more than two levels of fidelity. In addition, no particular design limitation is imposed, as the relationships between systems are inferred based on a measure of covariance.

The exploitation of multiple fidelity data for optimization has also received great interest, particularly in the area of multidisciplinary optimization (MDO). Kaufman et al. (1996) initiated the Variable Complexity Response Surface Model (VCRSM) method, using analyses of varying fidelity to reduce the design space to the region of interest and build response surface models of increasing accuracy. Alexandrov et al. (1998) and Rodriguez et al. (2001) coupled a low fidelity model with a numerical optimizer subject to a trust region constraint, where the lower fidelity model produces information that agrees with the higher fidelity model, within an acceptable error tolerance. Edy et al. (1998) used the solutions from genetic algorithm runs on the low-fidelity model to "seed" the population of a genetic algorithm run on the high-fidelity model. Keane (2003) proposed the "data fusion" technique, where a kriging model of the difference between high and low-fidelity systems was derived. The kriging model, combined with the low-fidelity data, provided a response surface of the high-fidelity system based on which the optimization is carried out.

In this paper, we will exploit the results by Kennedy et al. (2000) and propose an extension of the SKO method that utilizes multiple fidelity data to reduce total evaluation cost. This is probably the first optimization algorithm that uses an integrated criterion to determine both location and fidelity level of the subsequent search. We call this new method Multiple Fidelity Sequential Kriging Optimization (MFSKO). In Section 2, we describe the assumptions, algorithms, formulations, and some implementation issues. Section 3 uses a 1-dimention test problem to foster an intuitive understanding of the method. Additional numerical tests are documented in Section 4. In Section 5, the method is applied to a metal-forming process design

problem using Finite Element simulations. In Section 6, limitations and other relevant issues are addressed. Section 7 summarizes the conclusions and describes opportunities for future research.

2. Assumptions and Formulations

2.1. The Optimization Problem

Suppose there are a total of m systems to draw evaluations from, including the real and the surrogates. Denote the output functions of these systems in increasing order of fidelity by $f_1(\mathbf{x})$, $f_2(\mathbf{x})$, …, $f_m(\mathbf{x})$, where $\mathbf{x}$ is the input vector. Therefore, $f_1(\mathbf{x})$ has the lowest fidelity, and $f_m(\mathbf{x})$ has the highest fidelity. As mentioned previously, the highest-fidelity system is the system of interest, therefore the goal is to minimize $f_m(\mathbf{x})$ within the feasible region, χ, i.e.

$$\min_{\mathbf{x}\in\chi} f_m(\mathbf{x}) . \qquad (1)$$

We consider the systems as black boxes that provide no information other than measurements of the outputs. Denoting by d the dimension of the input space, we assume that the feasible region $\chi \subset \mathrm{R}^d$ is connected and compact.

Each system is associated with a cost-per-evaluation, which is denoted by C_1, C_2, …, C_m, respectively. In this research, the total cost of all evaluations measures the efficiency of the optimization scheme. Usually, a lower-fidelity evaluation is cheaper than a higher-fidelity evaluation, i.e. $C_1 < C_2 < \ldots < C_m$. Also, for now we assume that the cost of even the cheapest system is somewhat expensive, such that it is "worthwhile" to regenerate a kriging meta-model in order to determine the next search location. In Section 6, we will further discuss this issue and recommend some adaptations of our method for scenarios where very cheap evaluations are available.

In addition, the measurements of a system output may contain random error or noise. For each system, we assume that random errors from successive measurements are independent identically distributed (IID) normal deviates.

2.2. Overview of the Procedure

The outline for the proposed Multiple Fidelity Sequential Kriging Optimization (MFSKO) is as follows:

Step 1: Build the initial kriging meta-model for the system of interest. The kriging technique using multiple fidelity data is discussed in Section 2.3 and 2.4. The evaluation points can be allocated according to experimental designs provided in Section 2.5.

Step 2: Use a cross validation to diagnose whether the kriging prediction and the measure of uncertainty are satisfactory. If the test fails, appropriate transformations such as the logarithm or the inverse may be applied to the objective function until the test is passed. For details about the cross validation, please refer to the work by Jones et al. (1998).

Step 3: Find the location and fidelity level of the new evaluation that maximize the augmented Expected Improvement (EI) function, which is presented in Section 2.6. If the maximal EI is sufficiently small, terminate the optimization scheme.

Step 4: Conduct an evaluation where EI is maximized. Update the kriging meta-model with the new data point. Go to Step 3.

As conventions, step 1 is also referred to as the "initial fit" stage, while Steps 3 and 4 are called the "infill" or "update" stage. The sequentially added evaluations are also called the "infill" or "update" points. Note that the proposed method differs from its predecessors mainly in these two aspects: 1) the kriging meta-model is generated using multiple fidelity data. 2) The EI formulation takes into account not only the location but also the fidelity level of an infill point.

2.3. Kriging Meta-modeling for Multiple Fidelity Systems

To build kriging meta-models for multiple fidelity systems, we adopt a simplified version of the autoregressive assumption proposed by Kennedy and O'Hagan (2000). We assume that

$$f_l(\mathbf{x}) = f_{l-1}(\mathbf{x}) + \delta_l(\mathbf{x}) \qquad (l = 2, 3, \ldots, m) \tag{2}$$

where $\delta_l(\mathbf{x})$ is independent of $f_1(\mathbf{x})$, $f_2(\mathbf{x})$, …, $f_{l-1}(\mathbf{x})$. For the convenience in following notations, we also let

$$f_1(\mathbf{x}) = \delta_1(\mathbf{x}). \tag{3}$$

Note that for l = 2, 3, …, m, $\delta_l(\mathbf{x})$ can be understood as the "systematic error" of a lower-fidelity system, (l–1), as compared to the next higher-fidelity system, l. In these cases, $\delta_l(\mathbf{x})$ is usually small in scale as compared to $f_l(\mathbf{x})$, otherwise there will be no reason for the lower-fidelity system to exist.

In kriging meta-modeling, the response is assumed to be the sum of a linear model, a term representing the systematic departure (bias) from the linear model, and noise (Cressie 1993). We use kriging to model the lowest-fidelity system, $\delta_1(\mathbf{x})$, as well as the difference between systems, $\delta_l(\mathbf{x})$ (l = 2, 3, …, m). Therefore, we have

$$\delta_l(\mathbf{x}) = \mathbf{b}_l(\mathbf{x})^{\mathrm{T}}\,\boldsymbol{\beta}_l + Z_l(\mathbf{x}) + \varepsilon_l \qquad (l = 1, 2, \ldots, m) \tag{4}$$

where $\mathbf{b}_l$ and $\boldsymbol{\beta}_l$ are the basis functions and coefficients, respectively, of the linear model. Z_l is the systematic departure and ε_l is the random error. The basis functions of the linear model are often polynomials. In this paper we follow Jones et al. (1998) and use only one term, i.e. the constant term, for the linear model.

The kriging meta-model derives from an estimation process in which the systematic departure from the linear model, Z_l, is modeled as a zero-mean stationary Gaussian stochastic process. We use the following formula to describe the covariance between two points $\mathbf{x} = (x_1, \ldots x_d)$ and $\mathbf{x}' = (x'_1, \ldots x'_d)$:

$$\operatorname{cov}\left[\delta_l(\mathbf{x}), \delta_l(\mathbf{x}')\right] = \sigma^2_{Z,l} \exp\left[-\sum_{j=1}^{d} \theta_{l,j}(x_j - x'_j)^2\right] \tag{5}$$

where $\sigma^2_{Z,l}$ is the variance of the stochastic process, and $\theta_{l,j}$ is a "roughness" parameter associated with the dimension j. A larger $\theta_{l,j}$ implies a higher "activity", or lower spatial correlation, within the dimension j. (5) is sometimes referred to as a Gaussian covariance function.

From equation (2), it is trivial to derive that the output from system l at $\mathbf{x}$ and the output form system l' at $\mathbf{x}'$ have the following covariance

$$\operatorname{cov}\left[f_l(\mathbf{x}), f_{l'}(\mathbf{x}')\right] = \sum_{i=1}^{\min(l,l')} \operatorname{cov}\left[\delta_i(\mathbf{x}), \delta_i(\mathbf{x}')\right] \tag{6}$$

Intuitively, (6) implies that the covariance between two systems is due to the δ functions they share, which are determined by the lower fidelity level of the two.

We denote by Y_1, Y_2… Y_n the data drawn from an n-point design with point locations $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n\}$ and system indexes $\{l_1, l_2, \ldots, l_n\}$, respectively. Note that $1 \le l_1, l_2, \ldots, l_n \le m$. As mentioned previously, the data may contain random errors, which are assumed to be independent and identically distributed (IID). We denote by $\sigma^2_{\varepsilon,l}$ the variance of the random error associated with system l. To describe the kriging model predictor, we introduce the following notation:

$$\hat{\boldsymbol{\beta}} = \left[\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, ..., \hat{\boldsymbol{\beta}}_m\right]^T$$

$$\mathbf{h}_l(\mathbf{x}) = \left[\mathbf{b}_1(\mathbf{x})^T, \mathbf{b}_2(\mathbf{x})^T, ..., \mathbf{b}_l(\mathbf{x})^T, 0, ..., 0\right]^T \quad (l = 1, 2, \ldots, m)$$

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}_{l_1}(\mathbf{x})^T \\ \mathbf{h}_{l_2}(\mathbf{x})^T \\ \cdots \\ \mathbf{h}_{l_n}(\mathbf{x})^T \end{bmatrix}$$

$$\mathbf{V} = [\mathrm{cov}(Y_i, Y_j)]_{1 \le i, j \le n} = [\mathrm{cov}(f_{l_i}(\mathbf{x}_i), f_{l_j}(\mathbf{x}_j))]_{1 \le i, j \le n} + [\sigma^2_{\varepsilon,l_i} \eta_{ij}]_{1 \le i, j \le n}$$

$$\mathbf{t}_l(\mathbf{x}) = [\mathrm{cov}(f_{l_1}(\mathbf{x}_1), f_l(\mathbf{x})), ... \mathrm{cov}(f_{l_n}(\mathbf{x}_n), f_l(\mathbf{x})]^T$$

$$\mathbf{y}^T = [Y_1, ... Y_n]$$

where ^{T} denotes the transpose, $\eta_{ij} = 1$ for $i = j$, and $\eta_{ij} = 0$ for $i \ne j$. The best linear predictor (BLP) of $f_m(\mathbf{x})$ is:

$$\hat{f}_m(\mathbf{x}) = \mathbf{h}_m(\mathbf{x})^T \hat{\boldsymbol{\beta}} + \mathbf{t}_m(\mathbf{x})^T \mathbf{V}^{-1}(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}}) \qquad (7)$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{H}^T \mathbf{V}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{V}^{-1} \mathbf{y}$. Note that these formulations are essentially in agreement with Kennedy et al. (2000).

2.4. Maximum Likelihood Estimation (MLE) of the Hyper-Parameters

To compute (7), there are a number of hyper-parameters that need to be estimated. They include: $\sigma_{Z,l}^2$, $\sigma_{\varepsilon,l}^2$, and $\theta_{l,i}$, for l = 1, 2, …, m, and i = 1, 2, …, d. We estimate the hyper-parameters by maximizing the likelihood function of the samples

$$p(\mathbf{y}) = \frac{1}{(2\pi)^{n/2} |\mathbf{V}|^{1/2}} \exp\left[-\frac{(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}})}{2} \right] \tag{8}$$

(Note that, unlike in Sacks et al. (1989a), here we were not able to obtain a so-called "concentrated likelihood" by partially solving the maximization problem.)

To maximize (8), we use a multi-stage strategy proposed and justified by Kennedy and O'Hagan (2000). We assume that the data from a higher-fidelity system do not provide information about the hyper-parameters of a lower-fidelity system. In other words, to estimate the hyper-parameters of system l,

1) the hyper-parameters for systems 1 to (l – 1) are treated as fixed, and

2) the data from systems (l + 1) and above are ignored.

We believe that very little is lost in this simplification, due to the facts that the amount of lower-fidelity data is usually much larger than the amount of higher-fidelity data, and higher-fidelity data also depend on other hyper-parameters.

Therefore, the hyper-parameters for each system can be estimated separately, starting from the lowest-fidelity system (system 1) and ending with the highest-fidelity system (system m). This multi-stage strategy can greatly reduce the cost of solving the likelihood maximization problem. The original optimization problem has (2 + d) × m input variables. With this strategy, the problem is divided into m sub-problems, each of which has (2 + d) input variables. The latter is much easier to solve than the former as the input space is significantly smaller.

After the problem is broken into m stages, the likelihood maximization for each stage is still not trivial, as the likelihood function is often multimodal. In this research, we used a multiple-starting point Nelder-Mead simplex search method. We recommend the number of starting points to be at least three and increase with the number of dimensions.

2.5. The Design for Initial Fit

In Section 2.2, Step 1 involves an experimental design for the initial kriging fit. The designs for single-fidelity kriging models have been investigated extensively in the area of the Design and Analysis of Computer Experiments (DACE), where two main strategies, space-filling and criterion-based, have been proposed. For summaries on this area, see Santner et al. (2003) and Koehler, et al. (1996). The designs involving multiple-fidelity experiment data have been studied by a small number of groups. Kennedy et al. (2000 and 2002) suggested that the designs should give good coverage of the region of interest, and high-fidelity points should be close to the low-fidelity points in order to learn about the relationship between systems. Keane (2003) and Leary et al. (2003) used similar designs where the number of low-fidelity data is much larger than that of high-fidelity data. In this research, we follow the principles proposed in the prior work and adopt below guidelines for generating the initial-fit designs:

1) For the lowest-fidelity system, we use a Latin Hypercube (LHC) design (Stein 1987) with maximal minimum distance between points. LHC designs have good space-filling properties, and were also used by Jones et al. (1998).

2) For systems other than the lowest-fidelity system, points on a higher-fidelity system are a subset of the points on a lower-fidelity system. If subsets that are also LHC designs exist, choose the one that maximizes the minimum distance between points. If no subset can satisfy the LHC requirements, choose the subset with maximal minimum distance between points. We believe that this "subset" strategy can help generate accurate auto-regressive kriging models, as the differences between systems are observed.

3) If system output contains random errors, replicates are added where the best responses are found. These replicates allow direct estimations of the random errors. For details, refer to Huang, et al. (2005).

It is important to note that the "subset" design strategy recommended above is not required for generating auto-regressive kriging models. In fact, for the case study in Section 3, we use an initial-fit design where high-fidelity points that are not a subset of the low-fidelity points. Also, in the later infill stage, high-fidelity points are added without low-fidelity replicates.

The number of points used in the initial-fit designs may depend on the user's prior information about the objective function. If it is believed that the objective contains many important fine features (very "bumpy"), more points are preferred. Otherwise, fewer points may be adequate. In this study, we follow Jones et al. (1998) and use $10\times d$ as the default number of points on the lowest-fidelity system, where d is the dimension number of the input space. Note that current designs are based mostly on the intuitions and experience. We consider multi-fidelity experiment design an area that deserves much further research.

2.6. The Augmented Expected Improvement Function

As described in 2.2, Step 3 and 4, we will use the Expected Improvement (EI) as an integrated search criterion that determines both location and fidelity level of the subsequent evaluation. Construction of the EI function was one of the most challenging tasks in the research. Various forms had been considered before we settled with the current formations which are described below.

For intuitive presentation of the EI function, we will adopt a Bayesian point of view on kriging meta-modeling in this section. The Bayesian formulation is to quantify the uncertainty about the unknown function with data, using diffuse priors placed on $\boldsymbol{\beta}$, $\sigma^2_{z,l}$, $\sigma^2_{\varepsilon,j}$, and the $\theta_{l,j}$. For details, please refer to the work on Bayesian approaches for modeling computer experiment (Currin et. al., 1991, O'Hagan, 1989).

We denote by $f_l{}^p(\mathbf{x})$ the posterior distribution for the output of system l as a function of input $\mathbf{x}$. The posterior mean of $f_l{}^p(\mathbf{x})$ is equal to the BLP predictor, $\hat{f}_l(\mathbf{x})$, in (7), and the posterior covariance is

$$\operatorname{cov}\left[f_l{}^p(\mathbf{x}),\, f_{l'}{}^p(\mathbf{x}')\right] = \operatorname{cov}\left[f_l(\mathbf{x}),\, f_{l'}(\mathbf{x}')\right] - [\mathbf{h}_l(\mathbf{x})^T, \mathbf{t}_l(\mathbf{x})^T] \begin{bmatrix} \mathbf{0} & \mathbf{H}^T \\ \mathbf{H} & \mathbf{V} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{h}_{l'}(\mathbf{x}') \\ \mathbf{t}_{l'}(\mathbf{x}') \end{bmatrix} \qquad (9)$$

Note that when $l = l'$ and $\mathbf{x} = \mathbf{x}'$, (9) can be used as a measure for uncertainty of the prediction, which is referred to as the Mean Squared Error (MSE) of the prediction by Sacks et al. (1989b).

For Multiple Fidelity Sequential Kriging Optimization (MFSKO), we propose the following augmented Expected Improvement function:

$$EI(\mathbf{x},l) \equiv E\left[\max\left(\hat{f}_m(\mathbf{x}^*) - f_m^{\,p}(\mathbf{x}),0\right)\right]\cdot \alpha_1(\mathbf{x},l)\cdot \alpha_2(\mathbf{x},l)\cdot \alpha_3(l) \qquad (10)$$

where $\alpha_1(\mathbf{x},l) = \mathrm{corr}\left[f_l^{\,p}(\mathbf{x}),\, f_m^{\,p}(\mathbf{x})\right]$ (corr stands for correlation),

$$\alpha_2(\mathbf{x},l) = \left(1 - \frac{\sigma_{\varepsilon.l}}{\sqrt{s_l^2(\mathbf{x}) + \sigma_{\varepsilon.l}^2}}\right), \text{ where } s_l^2(\mathbf{x}) = \mathrm{cov}\left[f_l^{\,p}(\mathbf{x}),\, f_l^{\,p}(\mathbf{x})\right],$$

and $\alpha_3(l) = \dfrac{C_m}{C_l}$.

In (10), **x*** stands for the current "effective best solution" defined by

$$\mathbf{x}^* = \underset{\mathbf{x}\in\{\mathbf{x}_1,\mathbf{x}_2,\ldots\mathbf{x}_n\}}{\arg\max}\ [u(\mathbf{x})] \qquad (11)$$

where $u(\mathbf{x}) = -\hat{f}_m(\mathbf{x}) - cs_m(\mathbf{x})$.

The function $u(\mathbf{x})$ in (11) is introduced as a utility function to account for the uncertainty associated with the prediction of the objective function, and c is a constant that reflects the degree of risk aversion. We select c = 1.0 as our default, which implies a willingness to trade 1 unit of the predicted objective function for 1 unit of the standard deviation of prediction uncertainty.

In (10), the expectation is conditional given the past data and given estimates of the correlation parameters. Therefore, the expectation is computed by integrating over the distribution of $f_m^{\,p}(\mathbf{x})$, with $\hat{f}_m(\mathbf{x}^*)$ a fixed value. Based on results in Jones et al. (1998), the expectation can be calculated analytically as follows:

$$E\left[\max\left(\hat{f}_m(\mathbf{x}^*) - f_m^{\,p}(\mathbf{x}),0\right)\right] = \left(\hat{f}_m(\mathbf{x}^*) - \hat{f}_m(\mathbf{x})\right)\Phi\left(\frac{\hat{f}_m(\mathbf{x}^*) - \hat{f}_m(\mathbf{x})}{s_m(\mathbf{x})}\right) + s_m(\mathbf{x})\phi\left(\frac{\hat{f}_m(\mathbf{x}^*) - \hat{f}_m(\mathbf{x})}{s_m(\mathbf{x})}\right) \qquad (12)$$

where Φ and ϕ are the standard normal probability density and cumulative distribution functions, respectively.

Note that the expectation term in (10) relates only to the highest-fidelity system (real system), i.e., it is the expected gain if a highest-fidelity evaluation is added. A lower-fidelity evaluation at the same location would make less contribution. The term $\alpha_1(\mathbf{x},l)$ is designed to

account for the reduction in reward when a lower-fidelity evaluation is used. Clearly, when l = m, term $\alpha_1(\mathbf{x},l)$ should equal one. Also, $\alpha_1(\mathbf{x},l)$ should be zero when a sample at ($\mathbf{x}$, l) exists already and there is not random error, as such a replicate brings no additional benefit. In equation (10), we choose $\alpha_1(\mathbf{x},l)$ to be the correlation between the posterior estimate of system l and the posterior estimate of system m at location $\mathbf{x}$, because it satisfies abovementioned requirements and can be interpreted as the fraction of uncertainty on system m that can be eliminated once system l is known. Note that prior to the current form, we had experimented with a few other forms, for example, the ratio between prediction variances at different fidelity levels. They were not adopted because they either did not have desirable properties or were not as elegant.

The purpose of term $\alpha_2(\mathbf{x},l)$ is to adjust EI when outputs of system l contain random errors. It accounts for the diminishing return of additional replicates as the prediction becomes more accurate. As indicated in Huang et al. (2005), this factor is equal to the relative reduction in the posterior standard deviation after a new replicate is added. This factor equals one when the variance of the random errors is zero.

In (10), $\alpha_3(l)$ is the ratio between the cost-per-evaluation on the real system and that on system l. It represents an adjustment to the sampling strategy based on the evaluation costs. With the expected gains equal, a cheaper data point is preferred to a more expensive one. Again, prior to the current formulation, other forms have been considered. For example, instead of multiplying EI with a ratio, we could subtract EI by the cost of the evaluation (after scaling), and the search would stop when it is "indifferent" between the potential gain and the cost of additional evaluation. For this approach, however, users have to specify a ratio to convert between EI and the evaluation cost, and this ratio will affect behaviors of the search algorithm. We did not adopt this approach as we deemed these characteristics not desirable.

To summarize, each modifier term in (10) has important benefits. If $\alpha_1(\mathbf{x},l)$ is excluded, we will always choose on the lowest-fidelity system as it is cheaper. If $\alpha_2(\mathbf{x},l)$ is excluded, we will see many unnecessary replicates when random errors exist. And if $\alpha_3(l)$ is excluded, the cost-per-evaluation will be out of the picture, so the highest fidelity will always be chosen. As a whole, we believe the search criterion can resemble how an intelligent human conducts the

search. In addition, when only single fidelity data are used and random error is excluded, it reduces to the original form proposed by Jones et al. (1998). This criterion is further justified by the performance of the algorithm in the numerical test examples as shown in later sections. Nevertheless, rigorous studies on the properties and converge rate of the search criterion is a very important area of future research.

From Section 2.2, Step 3, the location and fidelity level of the next evaluation is selected by maximizing EI, i. e.:

$$(\mathbf{x}_{n+1}, l_{n+1}) = \arg\max_{\mathbf{x},l} EI(\mathbf{x}, l) \tag{13}$$

where $\mathbf{x}_{n+1}$ and l_{n+1} are the location and system index of the next evaluation. The EI maximization problem is solved by enumerating all possible l; and for each given l, the Nelder-Mead simplex approach is used to find the $\mathbf{x}$ that maximizes $EI(\mathbf{x},l)$. Note that the EI functions are usually multimodal and spiky at maxima. In addition, large areas of the search space have near-zero values, which can stall the search if the initial simplex has all zero vertices. Therefore, the EI functions are usually difficult to maximize. In this research, we use multiple-starting points, where the number of starting points is five per dimension. In addition, the starting points were chosen according to randomized space-filling designs, such as the uniform or Monte Carlo designs.

2.7. The Stopping Criterion

As mentioned in section 2.2, the optimization scheme stops when

$$\max_{\mathbf{x},l} EI(\mathbf{x}, l) < \Delta_s . \tag{14}$$

where Δ_s is the stopping criterion. A concern associated with implementing this criterion is that, with inadequate data, the kriging meta-model can sometimes underestimate the Expected Improvement and hence cause premature stopping. To reduce the probability of this, we generally require (13) to be satisfied a number of times consecutively before the final stopping. Specifically, in this paper, we use $(d + 1)$ for this number.

In addition, we adopt following principle in determining the stopping criterion

$$\Delta_s = r \times [\max(Y_1, Y_1, ..., Y_n) - \min(Y_1, Y_1, ..., Y_n)] \quad (15)$$

so that r is the ratio between the stopping criterion and the "active span" of the responses, in other words, the so-called "relative stopping criterion". In the rest of the paper, except in Section 4, r = 0.1% is used as the default, if not indicated otherwise.

3. An Illustrative Example

In this section, the proposed Multiple Fidelity Sequential Kriging Optimization (MFSKO) method is illustrated with a 1-dimensional test case. In this case, a two-system scenario is considered, which includes a real system and a surrogate system. We assume that the responses of the real system derives from a test function created by Sasena (2002)

$$f_2(x) = -\sin(x) - \exp(x/100) + 10 \qquad (0 < x < 10) \quad (16)$$

and we arbitrarily assume that the output of the surrogate system is

$$f_1(x) = -\sin(x) - \exp(x/100) + 10.3 + 0.03 \times (x-3)^2 \qquad (0 < x < 10). \quad (17)$$

These two functions are graphically displayed in Figure 1. Note that the two are somewhat "alike", with each function having two local optima. However, the global optimum of the surrogate system is different from that of the real system. In other words, data from the surrogate system can be "misleading" for finding the real global optimum. Therefore, it is important for the search algorithm be able to "escape" from a local optimum.

We further assume that the cost-per-evaluations of the real and surrogate systems are 4 and 1, respectively. We adopt a design for the initial fit that includes eight points, as indicated in Figure 1. In this design, six points, x = {0, 2, 4, 6, 8, 10}, are for the surrogate system and the remaining two points, x = {3.5, 6.5}, are for the real system.

The search pattern of the infill points is also displayed in Figure 1. Notice that the search initially concentrates on the left local optimum, until potential gain at the left local optimum is outweighed by uncertainty on the right local optimum. Then the algorithm explores near the right local optimum until the stopping criterion is met. The total number of evaluations is 15, 8 of which are on the surrogate system and 7 are on the real system. The total evaluation cost is 36. In a comparison study, the original Sequential Kriging Optimization (SKO) is run on the real

system. 11 evaluations are needed to reach the same stopping criterion, requiring a total evaluation cost of 44. In this case, by utilizing the surrogate system, the saving is 18.2%.

As mentioned previously, the Expected Improvement (EI) function is the criterion that determines the location and fidelity level of the subsequent search. To help understand properties of the EI function and benefit of each term in equation (10), Figure 2 displays the breakdown of EI just before the ninth point was added. Note that the expectation term in (10) incorporates prediction and the prediction uncertainty to provide a balanced search, and it is independent of the fidelity level. The term $\alpha_1(\mathbf{x},l)$ is a correlation that measures the reduction in reward when a lower-fidelity evaluation is used. As there is not random error in the outputs, $\alpha_2(\mathbf{x},1) = \alpha_2(\mathbf{x},2) = 1.0$. Also, based on the costs per evaluation, we have $\alpha_3(\mathbf{x},1) = 4.0$, and $\alpha_3(\mathbf{x},2) = 1.0$.

A kriging meta-model of the real system is generated at the end of the optimization scheme. Figure 3 shows the prediction error of the meta-model as compared to the true function. Note that the errors are smaller in areas near the local optima, where more evaluations have been allocated.

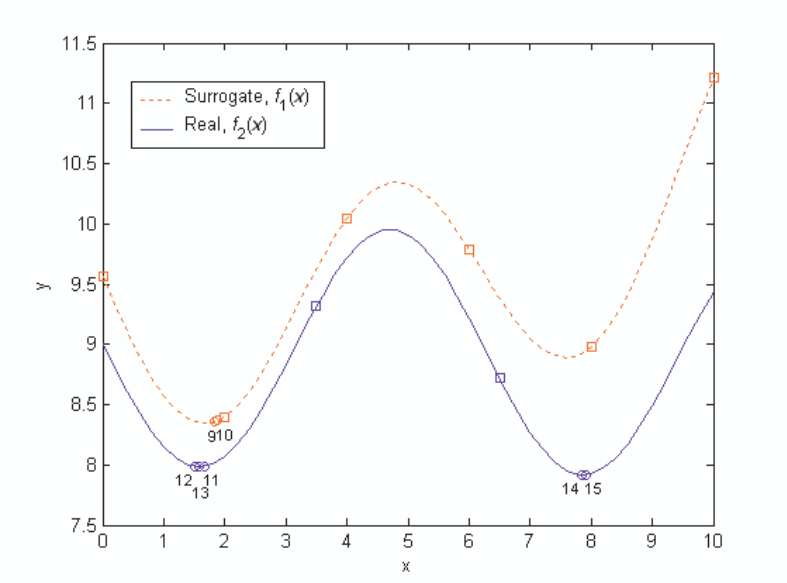
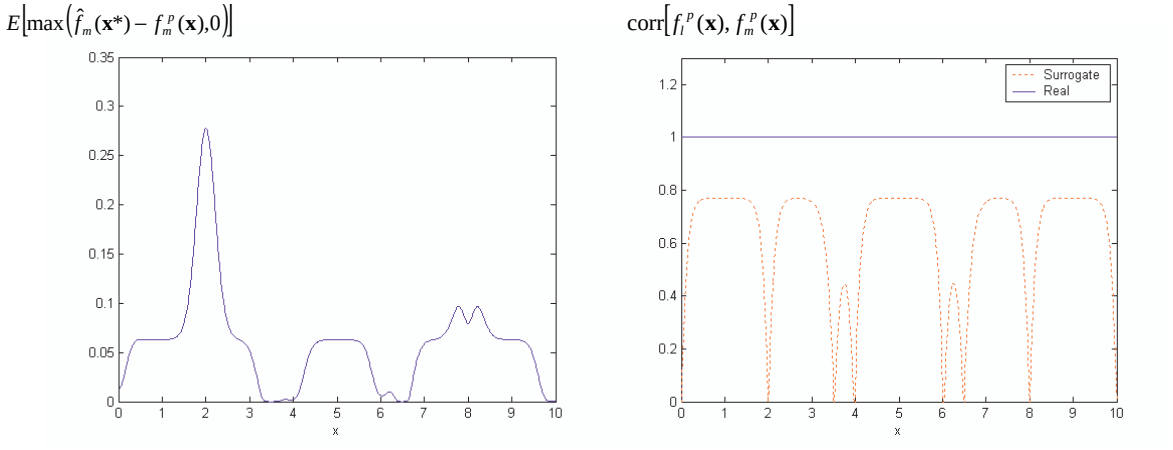
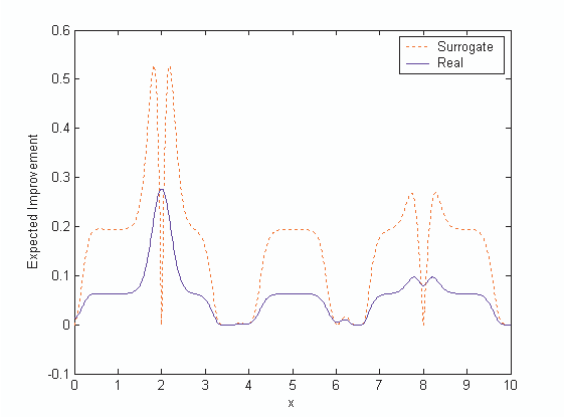


Figure 1. Search pattern for the Sasena (2002) test function
(□: initial-fit design; ○: infill points. Sequence of infill points is indicated.)



a) b)



c)

Figure 2. Breakdown of the EI function just before the 9th point was added. a) The expectation term in (10); b) The correlation term, $\alpha_1(\mathbf{x}, l)$, in (10); c) The expected improvement. (Note that $\alpha_2(\mathbf{x},1) = \alpha_2(\mathbf{x},2) = 1.0$, $\alpha_3(\mathbf{x},1) = 4.0$, and $\alpha_3(\mathbf{x},2) = 1.0$.)

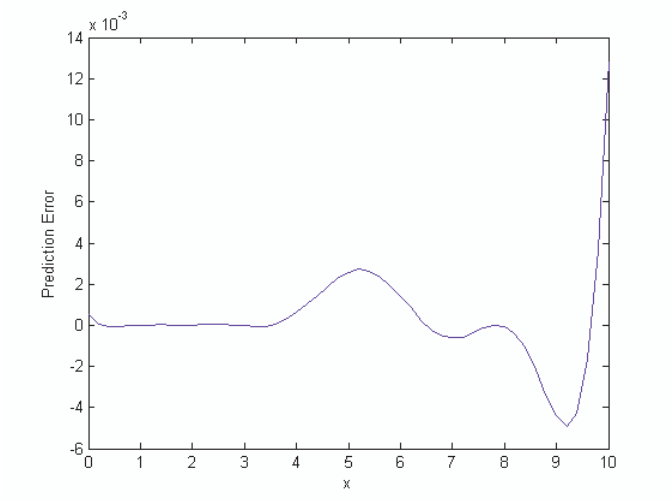


Figure 3. Prediction error of the final kriging meta-model for the real system

4. Some Numerical Tests

In this session, we further investigate the properties of the proposed MFSKO method using test functions from the literature. For the real systems, we use the three-dimension "Hartman 3" function (Hartman 1973) and the five-dimension "Ackley 5" function (Ackley 1987), as listed in Table 1. The surrogate systems are created by adding the systematic errors to the real system. For the systematic errors, we utilized the "polynomial test bed" method published by McDaniel et al. (2000) and randomly generated two polynomials, "MA 3" and "MA 5", which are also listed in Table 1. In this method, polynomials are created by fitting them to a number of random "responses" in the region of interest. Here, we used [0, 1] for the "range of responses" and default settings for other parameters such as the "effect heredity" and the "bumpiness". Note that due to "smoothing", the generated polynomials are in fact mostly between [0.25, 0.75]. These polynomials are then multiplied by a fraction of the objective function's "active span", i.e. the difference between minimal and maximal responses, to be the systematic errors. Here we only document the two-system scenarios, but the algorithm can be applied to more systems.

Five tests were conducted as listed in Table 2. Test #1 uses the "Hartman 3" for the real system and "MA 3" multiplied by 0.38 for the systematic error of the surrogate. The value of

0.38 was chosen such that the “active span” of the systematic error is approximately 5% of that of the objective function. Test #2 is similar to test #1 with the only difference being that the cost-per-evaluation of the surrogate system is larger. Test #3 also uses “Hartman 3” for real system, but the surrogate system's systematic error is three times that of test #1. In addition, to address the concern when the surrogate turns out to be a "bad" representation of the real system, in test #4 we make the systematic error about as large as the objective function itself. Finally, in test #5, we examine the performance of the MFSKO method in a five-dimension case, where “Ackley 5” is used for the real system and “MA 5” multiplied by 0.74 is used for the systematic error. In all of these tests, the initial-fit design contains $10\times d$ low-fidelity point and $3\times d$ high-fidelity points, where d is the number of dimensions.

Table 1 Test functions

Name and source	Function descriptions
Hartman 3 (Hartman 1973)	$d = 3$ $f(x) = -\sum_{i=1}^{4} c_i \exp\left[-\sum_{j=1}^{3} \alpha_{ij}(x_j - p_{ij})^2\right]$ where $\alpha_{ij} = \begin{bmatrix} 3 & 10 & 30 \\ 0.1 & 10 & 35 \\ 3 & 10 & 30 \\ 0.1 & 10 & 35 \end{bmatrix}$ $c_i = \begin{bmatrix} 1 \\ 1.2 \\ 3 \\ 3.2 \end{bmatrix}$ $p_{ij} = \begin{bmatrix} 0.3689 & 0.1170 & 0.2673 \\ 0.4699 & 0.4387 & 0.7470 \\ 0.1091 & 0.8732 & 0.5547 \\ 0.03815 & 0.5743 & 0.8828 \end{bmatrix}$ $0 \le x_i \le 15$, for $i = 1, 2, 3$, and $j = 1, \ldots, 4$ $N_{\text{local}} > 1$, $N_{\text{global}} = 1$ $\mathbf{x}^* = (0.114, 0.556, 0.852)$, $f^* = -3.8627$
Ackley 5 (Ackley 1987)	$d = 5$ $f(x) = -a\exp\left[-b\sqrt{\frac{1}{n}\sum_{i=1}^{d} x_i^2}\right] - \exp\left[\frac{1}{n}\sum_{i=1}^{d}\cos(cx_i)\right] + a + \exp(1)$, $a = 20$; $b = 0.2$; $c = 2\pi$ $-2.0 \le x_i \le 2.0$, for $i = 1, \ldots, d$ $N_{\text{local}} > 1$, $N_{\text{global}} = 1$ $\mathbf{x}^* = (0., 0., 0.)$, $f^* = 0.0$
MA 3 (McDaniel & Ankenman 2000)	$f(x) = 0.585 - 0.324x_1 - 0.379x_2 - 0.431x_3$ $-0.208x_1x_2 + 0.326x_1x_3 + 0.193x_2x_3 + 0.225x_1^2 + 0.263x_2^2 + 0.274x_3^2$
MA 5 (McDaniel & Ankenman 2000)	$f(x) = 0.588 - 0.00127x_1 - 0.00113x_2 - 0.00663x_3 - 0.0129x_4 - 0.00611x_5$ $+ 0.00526x_1x_4 + 0.0106x_1x_5 - 0.000626x_2x_4 - 0.00310x_2x_5 - 0.00724x_4x_5$ $-0.00096x_3^2 - 0.0124x_4^2 - 0.0101x_5^2$

Table 2. Numerical test results

#	Real system function	Surrogate System systematic error †	Cost-per-evaluation		Number of evaluations		Total cost	Relative gap to true optima ‡	Cost Reduction w.r.t. Orig. SKO
			Real	Surr.	Real	Surr.			
1	Hartmen 3 (d = 3)	MA 3 × 0.38 (5%)	1.0	0.25	10	37	19.25	0.01%	52%
2	Hartmen 3 (d = 3)	MA 3 × 0.38 (5%)	1.0	0.5	14	35	31.5	0.00%	21%
3	Hartmen 3 (d = 3)	MA 3 × 1.04 (15%)	1.0	0.25	12	38	21.5	0.03%	46%
3	Hartmen 3 (d = 3)	MA 3 × 7.6 (100%)	1.0	0.5	25	32	41	0.00%	- 2%
4	Ackley 5 (d = 5)	MA 5 × 0.74 (5%)	1.0	0.2	25	73	39.6	0.12%	56%

(†: Given in parentheses is the approximate ratio between the "active span" of the systematic error and that of the objective function.

‡: "Relative gap to true optima" is the difference between the solution and the true optima divided by the objective function's "active span".)

On the test results, as shown in Table 2, we have the following observations:

1) Global optima or near global optima are successfully found in all cases.

2) In cases where the surrogate is reasonably "good" (i.e. all except test #4), MFSKO more or less reduces total evaluation cost as compared to the original SKO method. The saving is particularly large in test #1 and #5, where the surrogate has small systematic errors and low evaluation costs.

3) Comparing test #2 to test #1, we see that as cost-per-evaluation of the surrogate system becomes higher, more points are allocated on the real system and fewer points are allocated on the surrogate system. As a result, the total evaluation cost increases.

4) Comparing test #3 to test #1, we find that as the systematic errors of the surrogate system become bigger, more evaluations are needed on both systems, while the final solution appears to worsen. Also, comparing test #3 to test #1, the increase in the number of evaluations on the real system is greater than the increase on the surrogate system, which may suggest that, relatively,

the surrogate system is utilized to a lesser degree. In addition, as we expected, total evaluation cost increases as the "accuracy" of the surrogate become poorer.

5) In test #4, where the surrogate is a very "bad" representation of the real, we found that nearly no additional samples are allocated on the surrogate system after the initial fit. More importantly, the total evaluation cost is more than without using the surrogate at all. This is understandable because the part of resource spend on assessing the surrogate ends up to be non-productive. One may suggests that if the systematic error is large enough, the usage of surrogate system should be completely discouraged even for the "initial-fit" stage. However, this may not be practical as the accuracy of the lower-fidelity model is often not known beforehand.

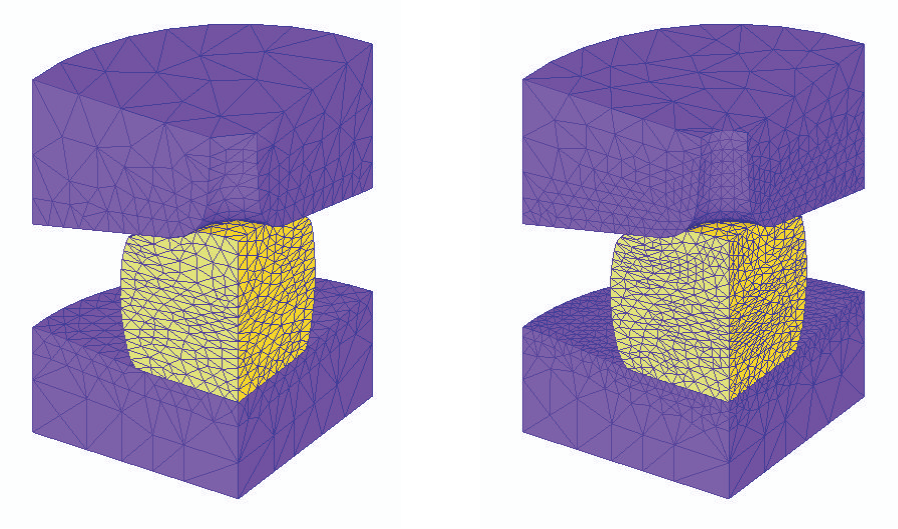
6) In general, the trends displayed in these tests seem to be consistent with our intuitions, and to some degree, justify the MFSKO algorithm, particularly the use of equation (10) for Expected Improvement (EI). Nevertheless, a wider variety of test functions should be included in future studies in order to draw more general conclusions on the behaviors of the algorithm.

5. Application Example: Die Wear Minimization

We now present an application of the proposed Multiple Fidelity Sequential Kriging Optimization (MFSKO) method for metal-forming process design improvement. In the forging industry, die wear is often the main factor that determines the lifespan of a die set, which in turn affects the cost of production. In this spike forging example, the manufacturers want to adjust the billet temperature (in the range 1650 to 2050 F) and die speed (in the range 0.01 to 10 inch/sec) to achieve the minimum wear rate. This is not a trivial task, because billet temperature and die speed both have dual effects on the die wear rate. On one hand, a higher billet temperature or a lower speed can reduce the material flow stress, leading to a lower wear rate. On the other hand, a higher billet temperature or a lower speed may increase the die temperature, causing a higher wear rate. Therefore, an optimization procedure is necessary.

In this application, evaluations are collected from computer simulations by the DEFORM® Finite Element Analysis (FEA) software. The simulation can be run at different levels of fidelity dictated by the fineness of the meshes. Here we use two levels: 10,000-element

and 27,000-element, which are referred to as system 1 and 2, respectively. Figure 4 displays these meshes graphically. We consider system 2 as the system of interest and system 1 a cheap surrogate. The computing time per simulation for system 1 and 2 is about 7.0 minutes and 41.6 minutes respectively, on a Pentium IV processor.



a) b)

Figure 4. Finite Element Models for Die Wear Simulations
(a: 10,000-element mesh, b: 27,000-element mesh)

We use a 26-point initial-fit design as shown in Figure 5. 20 of them are on system 1, and 6 of them are on system 2. As mentioned in section 2.5, the points on system 1 form a Latin Hypercube design; and the points on system 2 are a subset of the points on system 1, and in this case form another Latin Hypercube design.

Figure 5 also displays the search patterns of the optimization method. In this application, a relative stopping criterion r = 1% is applied, due to concerns on the computation cost. The optimization is terminated in 42 runs, which includes 31 cheap and 11 expensive runs. Table 3 lists the history of the infill points in detail. From this table, we see that the final best solution is in fact found by point 33, however, as a global optimization scheme, the algorithm continued to explore other potential areas until the expected improvement was sufficiently small.

As a comparison, we also use the original Sequential Kriging Optimization (SKO) method, with the same starting design and stopping criterion, to solve this problem entirely using

the expensive system. It took the method 35 runs to finish. Considering that the cost-per-evaluation on the surrogate system is only about one-sixth of that on the real system, MFSKO in this case is 54% more cost-effective than its original counterpart.

Note that in this case the optima of the high and low-fidelity models happen to be in the same place. Some may argue that the high-fidelity evaluations are wasted, as we could have reached the same optimal solution had we used the low-fidelity systems only. However, before we run the optimization, we usually do not know how well the high and low-fidelity systems are correlated or whether their optima are close in the input space. Without evaluating high-fidelity points, we could not find the best solution with good confidence.

In addition, solution difference between different fidelity levels may become greater when variable fidelity constraints are involved. For example, in structural designs using FEA simulations, a coarse-mesh model can underestimate maximal stresses when compared to a fine-mesh model. When maximal stress is used as a constraint, the feasible region of the high and low-fidelity may be different, which may lead to different optima, if the optima occur on the constraint boundary. In subsection 6.3, we will briefly discuss about potential approaches for constrained optimization, which is an important area of future research.

An additional advantage of MFSKO (and other SKO-type approaches) is that kriging global meta-models are created as by-products of the optimization. Figure 6 shows such meta-models for both the real and surrogate systems. This visualization can help engineers grasp an intuitive understanding of the phenomenon.

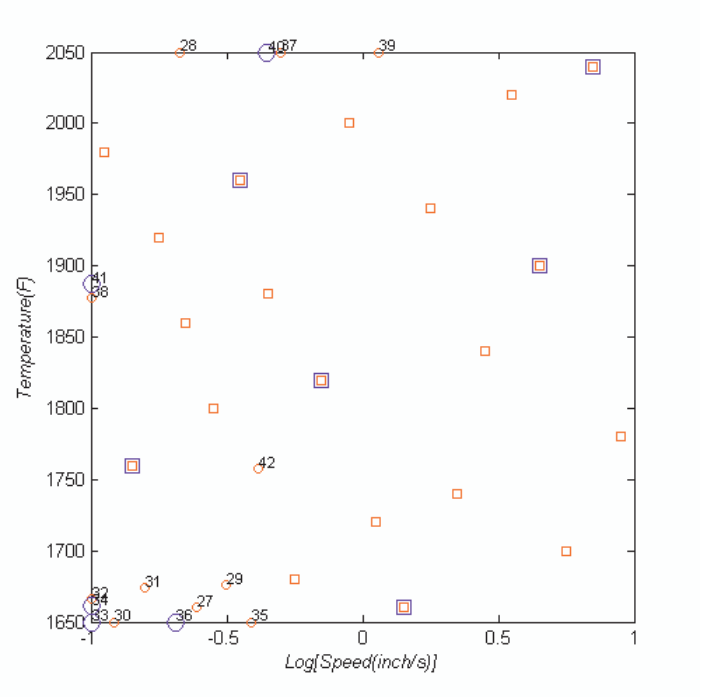


Figure 5. The search patterns of the die wear minimization problem
(▫: system 1 initial-fit design points; □: system 2 initial-fit design points;
○: system 1 infill points; ○: system 2 infill points.
The numbers indicate the sequence of the infill points.)

Table 3. Detailed history of the infill points

#	Effective best solution X*	Predicted best response	Maximal EI	Infill Point Location	Evaluated System	Response
27	(-0.45, 1960)	0.1622	0.15908	(-0.611, 1660)	1	0.10424
28	(-0.611, 1660)	0.13484	0.031961	(-0.674, 2050)	1	0.13243
29	(-0.611, 1660)	0.13484	0.012704	(-0.503, 1676.4)	1	0.10916
30	(-0.611, 1660)	0.13518	0.055904	(-0.916, 1650)	1	0.093685
31	(-0.916, 1650)	0.12628	0.0088	(-0.800, 1673.8)	1	0.10224
32	(-0.916, 1650)	0.12628	0.009954	(-1.000, 1666.1)	1	0.091498
33	(-1.000, 1666.1)	0.12478	0.006156	(-1.000, 1650)	2	0.13564
34	(-1.000, 1650)	0.13564	0.008841	(-1.000, 1661.6)	2	0.13831
35	(-1.000, 1650)	0.13564	0.002603	(-0.408, 1650)	1	0.11079
36	(-1.000, 1650)	0.13564	0.003573	(-0.687, 1650)	2	0.14579
37	(-1.000, 1650)	0.13564	0.00598	(-0.301, 2050)	1	0.11875
38	(-1.000, 1650)	0.13564	0.00467	(-1.000, 1877.6)	1	0.11493
39	(-1.000, 1650)	0.13564	0.003119	(0.057, 2050)	1	0.14091
40	(-1.000, 1650)	0.13564	0.0009	(-0.351, 2049.3)	2	0.15644
41	(-1.000, 1650)	0.13564	0.000395	(-1.000, 1887.7)	2	0.17386
42	(-1.000, 1650)	0.13564	0.000192	(-0.387, 1757.3)	1	0.11809

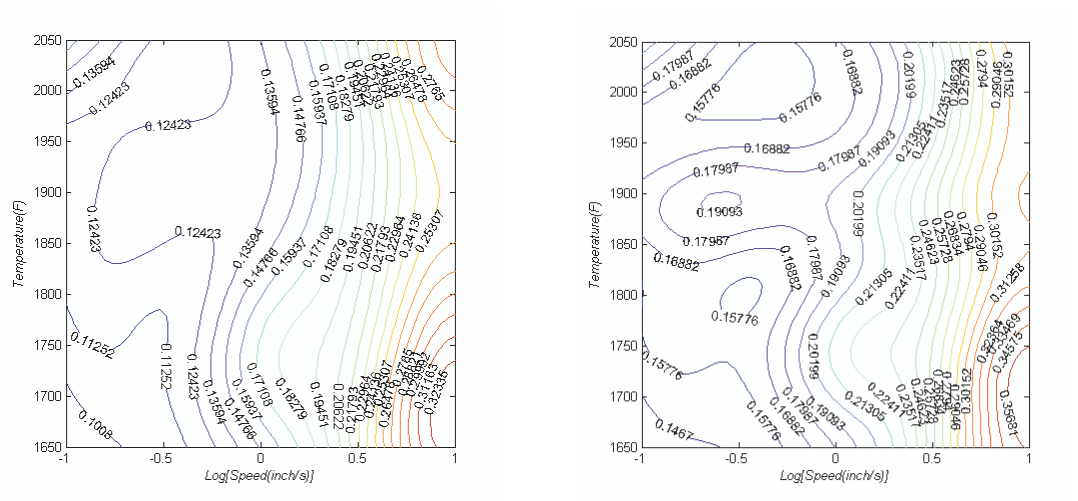


Figure 6. Kriging meta-models of the die wear rate (in 10^{-6} mm per forge)
(a: system 1, 27,000-element FEA model; b: system 2, 10,000-element FEA model)

6. Other Relevant Issues

6.1 Overhead cost

An important drawback of the SKO method (and its variations such as MFSKO) is the high overhead cost as compared to most standard optimization methods. This is not a surprise, because the SKO method processes all previous data points in order to determine the point of next search. For example, when the number of samples is 150, the computing time per iteration reaches about 60 seconds on a Pentium III 1.2G processor. Therefore, SKO is worthwhile only when the objectives are "expensive" enough, i.e. the cost-per-evaluation should be at least higher than the overhead cost per iteration.

Specifically, the overhead cost mainly includes fitting the kriging meta-model and maximizing the Expected Improvement functions. The cost of fitting the kriging increases with the number of data, as the matrix $\mathbf{V}^{-1}$ becomes larger and more expensive to invert. In addition, when the input space dimensionality becomes higher, a larger number of samples are needed to

generate useful predictions and more starting points are needed to search for the maximal EI. Therefore, the algorithm may be prohibitively expensive when the dimensionality is too high. In our studies, the maximum number of dimensions tried was 10.

6.2 When the cheapest system is very cheap

As mentioned in Section 2.1, in this paper we assume that even the cheapest system is somewhat expensive, such that it is worthwhile to use the MFSKO method to determine search points on it. For the scenarios where this assumption is not true, we proposed some adaptations of the MFSKO method as discussed below.

Let us consider Scenario I, where the cost per evaluation of the cheapest system is less than that of evaluating a given kriging model. (Note that an evaluation on a kriging model, as indicated in equation (7), is very cheap, because with hyper-parameters fixed, $\mathbf{V}^{-1}$ can be pre-computed, the cost is only of multiplying a few matrixes.) In this case, the cheapest system can be treated as a baseline, and the distances to this baseline are treated as the outputs of all other systems. Therefore, the total number of fidelity is reduced by one, and the baseline is evaluated every time any other system is evaluated. Note that similar concept has been adopted in the "data fusion" technique by Keane (2003). For the MFSKO method, to accommodate the baseline, the Expected Improvement function needs some modification, i.e., the expectation term in (10) should be replaced by:

$$E\left[\max\left(\hat{f}_m(\mathbf{x}^*) + B(\mathbf{x}^*) - f_m^p(\mathbf{x}) - B(\mathbf{x}),\quad 0\right)\right],$$

where $B(\mathbf{x})$ is the baseline function.

Let us consider another scenario, Scenario II, where the cost of the cheapest evaluation is higher than an evaluation on a kriging model but lower than refitting a kriging meta-model, which requires regenerating the hyper-parameters. In this case, we can generate a kriging meta-model of the cheapest system and replace the cheapest system with it. After the replacement, the cheapest system is a kriging model, thus the problem becomes a Scenario I problem which we can solve accordingly. Note that for the kriging meta-model to be a good global predictor, the number of points in the design needs to be large.

Scenario II may also be resolved by using a special initial-fit design in the MFSKO method. In the initial-fit design, allocate a large number of points on the cheapest system. According to (10), it is unlikely that additional evaluation will be allocated on the cheapest system, as the uncertainty on it is little. Without additional data, the hyper-parameters associated with the cheapest system need not be updated, so the overhead of refitting kriging is reduced significantly. One may wonder whether this approach is essentially the same as the approach described in the last paragraph. The answer is no, because this approach takes into account the uncertainty with the kriging prediction of the cheapest system, but the previous approach does not. However, when number of points is large, the uncertainty of the kriging prediction is usually small, so these two methods may lead to similar behaviors.

6.3 Constrained optimization

In this paper, we focus on problems with simple bound constraints. We intend to leave more complex constrained problems for future research, but want to provide some thoughts here. A number of approaches have been proposed to handle constraints in SKO depending on the cost of constraint evaluations. When constraints are inexpensive, we can simply maximize the infill sampling criteria as a constrained optimization problem, as proposed by Sasena et al. (2002). However, when constraints are expensive, we may need to utilize meta-models, thus the problem is complicated due to the prediction uncertainty on the constraints. Schonlau (1997) suggested multiplying the value of the expected improvement by the probability that the point is feasible. Björkman et al. (2000) applied a penalty method whereby a large constant is added to the search criterion in the infeasible region in order to discourage adding samples there. Audet et al. (2000) calculate the so-called “expected violation” of the constraints, before the expected improvement criterion is considered. Williams et al. (2004) adopted a Bayesian improvement function for a scenario where mean of one computer code is optimized under constraints defined by a second computer code.

Like the objective, constraints can also be evaluated at different levels of fidelity. Note that sometimes the constraints are evaluated in conjunction with the objective, in which cases the fidelity levels of the constraints will be the same as those of the objective. In other scenarios, the

constraints are independent, thus their fidelity levels may be different from those of the objective. Two approaches may be considered to deal with variable fidelity constraints:

1) Combine the constraints into the objective function using a penalty method and apply MFSKO to the combined objective function. This approach may be natural when the constraints and the objective share the levels of fidelity. For the approach to be effective, careful construction of the penalty function is important. The combined objective function should not be too "kinky", which may cause difficulties in generating good kriging fit. Also, the penalty needs to be large enough the make difference, but not too much larger in magnitude than the original objective function, otherwise the optimization tends to stop prematurely.

2) Create separate meta-models for the objectives and the constraints, and then maximize the expected improvement of the objective subject to the approximate constraints. Note that the meta-models for constraints may also be auto-regressive multi-fidelity kriging models as discussed in Section 2.2. Proposed constrained optimization methods, such as the one by Sasena et al. (2002), can be adopted accordingly. Note that in this case the infill points for the objective and the constraints may be determined separately.

7. Conclusions and Future work

We propose an extension of the Sequential Kriging Optimization (SKO) method that exploits cheaper data from surrogate systems for finding the global optima of the real system. This new method, named Multiple Fidelity Sequential Kriging Optimization (MFSKO), uses an integrated criterion to determine both location and fidelity level of the subsequent search. Applications to several test functions from the literature and a metal-forming process design problem show sensible search patterns, robust performance and appreciable reduction in total evaluation cost as compared to the original method. Test results also suggest that when a surrogate system's cost-per-evaluation becomes higher or its systematic error becomes larger, relatively fewer evaluations will be allocated to that surrogate system and the total evaluation cost will increase. Finally, an additional advantage of MFSKO is that global meta-models, which can be used for visualization, are created for every fidelity level as by-products of the optimization.

Future work should include optimal designs for the initial fit, which uses up an important portion of the total evaluation cost. As multiple levels of fidelity are involved, such designs represent new challenges to the Design and Analysis of Computer Experiments (DACE) community. Also, how is the size of the initial-fit design related to how long the method takes to reach the stopping criterion? Too many initial-fit points may be wasteful, but too few may cause us to miss important information. Another area of future study may relate to the selection of surrogate systems, especially when the fidelity level is a matter of choice, such as in Finite Element simulations. As indicated in the study, cheap and accurate surrogates are preferred, and very "bad" surrogates may be counter-productive. Therefore, it will be useful to have guidelines to determine whether using certain surrogate is beneficial, and in some cases, what fidelity level can bring the best results. At last but not least, variable fidelity constrained optimization is an essential direction of future research which promises great applications in areas such as structural optimization.

Acknowledgments

This research was partially funded by Scientific Forming Technologies Corporation. We would like to thank Wei-Tsu Wu and Thomas Santner for ideas, references, and encouragement. In addition, we thank Matthias Schonlau, Michael Sasena, and Ofelia Marin for providing source code, documentation, and discussions. Finally, we thank the anonymous reviewers for valuable suggestions that greatly helped us improve this paper.

Reference:

Ackley, D.H. (1987), A Connectionist Machine for Genetic Hill-climbing, Kluwer Academic Publishers, Boston.

Alexzndrov, N. M., Dennis, J. E., Jr., Lewis, R. M., Torczon, V. (1998), A Trust Region Framework for Managing the Use of Approximation Models in Optimization. Structural Optimization 15(1), 16-23.

Audet, C., Dennis, J.E., Jr., Moore, D.W., Booker, A., and Frank, P.D. (2000), A Surrogate-Model-based Method for Constrained Optimization. In Proceedings of the 8th AIAA/NASA/USAF/ISSMO Symposium on Multidisciplinary Analysis and Optimization, AIAA-2000-4891.

Bandler, J. W., Ismail, M. A., Rayas-Sanchez, J. E. and Zhang., Q. (1999). Neuromodelling of microwave circuits exploiting space-mapping technology, IEEE Transactions on Microwave Theory and Techniques, 47, 2417–2427.

Björkman, M., and Holström, K. (2000), Global Optimization of Costly Nonconvex Functions Using Radial Basis Functions. Optimization and Engineering, 1, 373–397.

Cressie, N.A.C. (1993), Statistics for Spatial Data (Revised edition). Wiley, New York.

Currin, C., Mitchell, M. Morris, M., and Ylvisaker D. (1991), Bayesian Prediction of Deterministic functions, with Applications to the Design and Analysis of Computer Experiments. Journal of American Statistics Association 86, 953-963.

Edy, D., Averill, R. C., Punch, W. F. III, and Goodman, E. D. (1998), Evaluation of Injection Island GA Performance on Flywheel Design Optimization. In: I.C. Parmee (ed.), Adaptive Computing in Design and manufacture, Springer-Verlag.

Hartman, J.K. (1973), Some Experiments in Global Optimization. Naval Research Logistics Quarterly 20, 569-576.

Huang, D., Allen, T. T., Notz, W. I., and Zheng, N. (2005), Global Optimization of Stochastic Black-Box Systems via Sequential Kriging Meta-Models, to appear on the Journal of Global Optimization.

Hutchinson, M. G., Unger, E. R., Mason, W. H., Grossman, B. and Haftka, R. T. (1994). Variable-complexity aerodynamic optimization of a high speed civil transport wing, Journal of Aircraft, 31, 110–116.

Jones, D., Schonlau, M. and Welch, W. (1998), Efficient Global Optimization of Expensive Black-Box Functions. Journal of Global Optimization 13, 455-492.

Koehler, J.R., and Owen, A.B. (1996), Computer Experiments. In: S. Ghosh and C. R. Rao (eds), Handbook of Statistics, Vol. 13 , Elsevier Science B. V.

Kaufman, M., Balabanov, V., Burgee, S. L., Giunta, A. A., Grossman, B., Mason, W. H. Watson, L. T. (1996), Variable-Complexity Response Surface Approximations for Wing Structural Weight in HSCT Design, 34th Aerospace Sciences Meeting and Exhibit, Reno, NV, AIAA-96-0089.

Keane, A.J. (2003) Wing Optimization Using Design of Experiment, Response Surface, and Data Fusion Methods. Journal of Aircraft, 40(4), 741-750.

Kennedy, M. C. and O'Hagan, A. (2000) Predicting the output of a complex computer code when fast approximation are available. Biometrika, 87, 1, 1-13.

Kushner, H.J. (1964), A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise. Journal of Basic Engineering 86, 97-106.

Leary, S.J., Bhaskar, A. and Keane, A.J., (2003) A Knowledge-Based Approach To Response Surface Modelling in Multifidelity Optimization, Journal of Global Optimization, 26(3), 297-319.

McDaniel, W. R. and Ankenman, B. E. (2000), A Response Surface Test Bed. Quality and Reliability Engineering International, 16, 363-372.

O'Hagan, A. (1989), Comment: Design and Analysis of Computer Experiments. Statistic Science 4, 430-432.

Rodriguez, J. F., Perez, V. M., Padmanabhan, D. and Renaud, J. E. (2001), Sequential Approximate Optimization Using Multiple Fidelity Response Surface Approximation. Structural and Multidisciplinary Optimization 22(1), 23-34.

Sacks, J., Welch W.J., Mitchell, T.J. and Wynn, H.P. (1989a), Design and Analysis of Computer Experiments (with discussion). Statistical Science 4, 409-430.

Sacks, J., Schiller, S.B., and Welch, W. (1989b), Design for Computer Experiments. Technometrics 31, 41-47.

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

Sasena, M.J. (2002), Flexibility and Efficiency Enhancements for Constrained Global Design Optimization with Kriging Approximations, Ph. D. dissertation, University of Michigan.

Sasena, M.J., Papalambros, P.Y. and Goovaerts, P. (2002), Exploration of Metamodeling Sampling Criteria for Constrained Global Optimization. Engineering Optimization 34, 263–278.

Schonlau, M. (1997), Computer Experiments and Global Optimization, Ph.D. Dissertation, University of Waterloo.

Stein, M. (1987), Large Sample Properties of Simulation Using Latin Hypercube Sampling. Technometrics 29, 143-151.

Williams, B.J., Lehman, J. S., Santner, T.J., and Notz, W.I. (2004), Sequential Design of Computer Experiments for Constrained Optimization of Integrated Response Functions. In revision for Statistica Sinica .