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

ON DIRECT GRADIENT ENHANCED SIMULATION METAMODELS

Huashuai Qu

Department of Mathematics
University of Maryland, College Park,
MD, 20742, U.S.A

Michael C. Fu

The Robert H. Smith School of Business &
Institute for Systems Research,
University of Maryland, College Park,
MD, 20742, U.S.A

ABSTRACT

Traditional metamodel-based optimization methods assume experiment data collected consist of performance measurements only. However, in many settings found in stochastic simulation, direct gradient estimates are available. We investigate techniques that augment existing regression and stochastic kriging models to incorporate additional gradient information. The augmented models are shown to be compelling compared to existing models, in the sense of improved accuracy or reducing simulation cost. Numerical results also indicate that the augmented models can capture trends that standard models miss.

1 INTRODUCTION

Commonly used in simulation optimization, a metamodel provides an auxiliary functional relationship between input and output of a simulation model. Other simulation optimization approaches include ranking and selection, stochastic approximation and sample path optimization; see Fu (2002), Fu et al. (2008) for a recent survey and tutorial with references. Conducting simulations to collect experimental data is necessary to build metamodels, where the simulated data collected are usually performance measurements for parameters of interest. However, direct derivative information may also be available in stochastic simulation settings, where the output responses include not only the performance measurement, but also values of the gradient of performance measurement with respect to the parameters. Perturbation analysis (PA) (Ho and Cao 1991; Glasserman 1991; Glasserman 2004) and likelihood ratio/score function methods (LR/SF) (Rubinstein 1986; Rubinstein and Shapiro 1993) are techniques that aim at estimating the gradient the performance measure. Applications of direct gradient estimates have been studied extensively, including queueing, inventory and finance applications (Asmussen and Glynn 2007; Fu 2008).

Metamodel-based methods decouple optimization from simulation, as metamodels approximate stochastic responses through an algebraic function and deterministic optimization procedures are applied to the metamodel. In general, there are two types of metamodel strategies: iterated local metamodels and global metamodels. An overview of local and global metamodel-based optimization is given in Barton and Meckesheimer (2006) and Barton (2009).

Iterated local metamodels, also known as sequential response surface methodology, rely on low-order polynomial regression. A first-order polynomial is usually used to fit local response surface in a small region to determine the search direction. Following a line search, new regions for the parameters of interest are exploited repeatedly until the region of most interest is determined. At the final step, a quadratic approximation is chosen and deterministic optimization methods are applied to locate the optimum. Regression techniques and experiment design are critical in this procedure; see Kleijnen (2008) for details.

In global metamodels, high-order polynomial regression or nonlinear regression techniques based on existing knowledge about the response surface are appropriate; see Yang et al. (2007) for an example. To capture global characteristics of a response surface, more flexibilities in the models are required. Therefore, kriging, splines, neural networks and radial basis functions are more adequate to fit global metamodels.

Among all these, kriging metamodels have received a lot of attention in the stochastic simulation community over the past decade (Cressie 1993; Stein 1999; Kleijnen et al. 2010). Recently, Ankenman et al. (2010) proposed stochastic kriging as an extension of kriging, which takes the uncertainties in simulation noise into consideration. Stochastic kriging is considered to be flexible and promising in fitting global response surfaces, especially in stochastic simulation settings.

Gradient estimates have been used in local search procedures such as stochastic approximation (Ho and Cao 1983; Fu 1994). For metamodel-based optimization, researchers have made attempts to incorporate gradient estimates into iterate local metamodels and global metamodels. In the sequential aspect, another approach called gradient surface method (GSM) was proposed by Ho et al. (1992) as a simulation optimization procedure that fits the gradient response surface directly using the gradient estimates only; the function estimates themselves are not used in the procedure. Liu (2003) developed approaches to approximate response surface based upon artificial neutral networks and kriging. Chen et al. (2011) introduced stochastic kriging with gradient estimators (SKG) approach to exploit gradient estimates in stochastic kriging, showing that the new approach provides better prediction in the sense of smaller mean squared error (MSE). This approach is similar to cokriging used in deterministic simulations. Therefore, differentiability of correlation functions are required in their approach, as derivatives of stochastic processes are used to formulate models for gradient estimates.

In this paper, we examine the potential improvements in fitting local and global metamodels when gradient information is available. In the regression setting, we investigate the Direct Gradient Augmented Regression (DiGAR) in Fu and Qu (2012), which is a modification of the standard linear regression model to incorporate gradient estimates; in the stochastic kriging setting, the gradient extrapolated stochastic kriging (GESK) method proposed in Qu and Fu (2012) is investigated, where we use all available simulation outputs (both performance measurements and gradient estimates) to extrapolate more data. More spatial correlations are introduced to the data from extrapolation, which can be employed by stochastic kriging.

Experiments are used to illustrate the effectiveness of the augmented models. Preliminary results show that the DiGAR approach has several attractive features: it is less sensitive to outliers; it corrects the shape of the fitted curve - the slope for a linear fit and the curvature for a quadratic fit; it provides estimators with smaller variance than the standard regression model. Based on the numerical experiments considered here, GESK is shown to perform better than stochastic kriging, and it is comparable to or better than SKG. Moreover, GESK captures fluctuations of the response surface which are usually missed by the other two approaches.

The rest of the paper is organized as follows. Section 2 and 3 present models to incorporate gradient estimates in regression and stochastic kriging. In Section 4, numerical results for both enhanced metamodels are provided. Section 5 presents conclusions and topics for future research.

2 AUGMENTING REGRESSION MODELS

In this section we review the DiGAR models in Fu and Qu (2012) that incorporate gradient estimates under various assumptions.

2.1 Independent DiGAR

Given a data set $\{\mathbf{x}_i, y_i\}_{i=1}^k$ and $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{id})' \in \mathbb{R}^d$, the classical linear regression model takes the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_d x_{id} + \varepsilon_i,$$

where the prime denotes transpose. The parameters β_i's are regarded as unknown and to be estimated.

Now consider the enhanced setting where the k data points are $(\mathbf{x}_1, y_1, \mathbf{g}_1), \ldots, (\mathbf{x}_k, y_k, \mathbf{g}_k)$, with $\mathbf{g}_i = (g_i^1, g_i^2, \cdots, g_i^d)' \in \mathbb{R}^d$ representing a direct estimate of the gradient of y_i at $\mathbf{x}_i$, The DiGAR model takes the

form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_d x_{id} + \varepsilon_i, \tag{1}$$

$$g_i^j = \beta^j + \delta_i^j, \tag{2}$$

where y_i and g_i^j, for $i = 1,2,\ldots,k$ and $j = 1,2,\cdots,d$, are the performance measures and gradient estimates with residuals $\{\varepsilon_i\}$ and $\{\delta_i^j\}$, respectively.

If we stack all the $(d+1)\times k$ equations in (1) and (2) together, we have the same vector form model as in classical linear regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \mathbf{g}_1 \\ \vdots \\ y_k \\ \mathbf{g}_k \end{pmatrix}, \quad \mathbf{X} = \left(\begin{array}{cccc} 1 & x_{11} & \cdots & x_{1d} \\ 0 & 1 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & 1 \\ \hline \vdots & & & \vdots \\ \hline 1 & x_{k1} & \cdots & x_{kd} \\ 0 & 1 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & 1 \end{array}\right), \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \boldsymbol{\delta}_1 \\ \vdots \\ \varepsilon_k \\ \boldsymbol{\delta}_k \end{pmatrix}, \quad \boldsymbol{\delta}_i = \begin{pmatrix} \delta_i^1 \\ \delta_i^2 \\ \vdots \\ \delta_i^d \end{pmatrix}. \tag{4}$$

For illustration purposes, we consider the one-dimensional problem, i.e., the given data points are (x_1,y_1,g_1),..., (x_k,y_k,g_k) and $\mathbf{X}\boldsymbol{\beta} = \beta_0 + x\beta_1$. Using the ordinary least-squares approach, the function to be minimized is the sum of the squared deviations in both y_i and g_i,

$$L = \sum_{i=1}^{k}(y_i - \beta_0 - \beta_1 x_i)^2 + \sum_{i=1}^{k}(g_i - \beta_1)^2. \tag{5}$$

Here equal weights for the sum of squared error are used, and the extension to convex combination is considered in Fu and Qu (2012). Denote $\hat{\beta}_i^D$ and $\hat{\beta}_i^L$, $i = 0,1$, as estimators from DiGAR and classical linear regression, respectively. The resulting estimators that minimize (5) are

$$\hat{\beta}_0^D = \bar{y} - \hat{\beta}_1^D \bar{x}, \tag{6}$$

$$\hat{\beta}_1^D = \frac{\sum_{i=1}^{k}(x_i - \bar{x})(y_i - \bar{y}) + k\bar{g}}{\sum_{i=1}^{k}(x_i - \bar{x})^2 + k}, \tag{7}$$

while the estimators in classical linear regression are

$$\hat{\beta}_0^L = \bar{y} - \hat{\beta}_1^L \bar{x}, \qquad \hat{\beta}_1^L = \frac{\sum_{i=1}^{k}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{k}(x_i - \bar{x})^2},$$

where $\bar{x}$, $\bar{y}$ and $\bar{g}$ are the corresponding sample means of x_i, y_i and g_i. Note that in the augmented model, the form of $\hat{\beta}_0^D$ in (6) remains unchanged, whereas $\hat{\beta}_1^D$ in (7) has the additional terms $k\bar{g}$ and k in the numerator and denominator, respectively, reflecting the added gradient information.

Assumption 1

i) $E(\delta_i) = E(\varepsilon_i) = 0, \forall i.$
ii) $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$, $\text{Cov}(\delta_i, \delta_j) = 0$, $\text{Cov}(\varepsilon_i, \delta_j) = 0, \ \forall \ i, j.$
iii) $\text{Var}(\varepsilon_i) = \sigma^2$ and $\text{Var}(\delta_i) = \sigma_g^2$.

We have the following lemma about the variances of the slope estimator.

Lemma 1 Under Assumption 1, the variance of DiGAR estimator $\text{Var}(\hat{\beta}_1^D) \leq \text{Var}(\hat{\beta}_1^L)$ if $\sigma_g^2 \leq C \cdot \sigma^2$, where

$$C = \frac{k + \sum_{i=1}^{k} x_i^2 - k\bar{x}^2}{\sum_{i=1}^{k} x_i^2 - k\bar{x}^2}.$$

The primary reason that we are interested in the slope estimator is that we expect DiGAR can provide better search direction than classical linear regression in a sequential optimization procedure. Since $C > 1$, it suggests that as long as the variance of the gradient estimations g_i is not too much larger than the variance of y_i, the DiGAR estimator will have smaller variance.

Assumption 2 The residuals are normally distributed, i.e., $\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \ \delta_i \sim \mathcal{N}(0, \sigma_g^2)$.

Under Assumptions 1 and 2, y_i and g_i are independent due to the residuals being uncorrelated, and the likelihood function is given by

$$L(\beta_0, \beta_1, \sigma^2, \sigma_g^2) = (2\pi)^{-k}(\sigma\sigma_g)^{-k} \exp\left\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{k}(y_i - \beta_0 - \beta_1 x_i)^2 - \frac{1}{2\sigma_g^2}\sum_{i=1}^{k}(g_i - \beta_1)^2 \right\},$$

which leads to the following respective maximum likelihood estimators (MLEs) for β_0 and β_1:

$$\hat{\beta}_0^D = \bar{y} - \hat{\beta}_1^D \bar{x}, \qquad \hat{\beta}_1^D = \frac{\frac{1}{\sigma^2}\sum_{i=1}^{k} x_i y_i + \frac{k}{\sigma_g^2}\bar{g} - \frac{k}{\sigma^2}\bar{x}\bar{y}}{\frac{1}{\sigma^2}\sum_{i=1}^{k} x_i^2 + \frac{k}{\sigma_g^2} - \frac{k}{\sigma^2}\bar{x}^2}. \tag{8}$$

Lemma 2 Under Assumptions 1 and 2, the MLE $\hat{\beta}_1^D$ in (8) has smaller variance than $\hat{\beta}_1^L$.

2.2 Correlated DiGAR

In simulation metamodeling, the response outputs y_i often have heterogeneous variances and correlation induced by common random numbers. Moreover, the response outputs y_i and gradient estimates g_i are usually correlated. Thus, Assumption 1 is violated.

Similar to classical linear regression, generalized least squares (GLS) can be used to handle heteroscedasticity and correlations. Using the model in (3), we assume the residuals have zero mean, i.e., $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, so $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, and $\text{Cov}(\boldsymbol{\varepsilon}) = \mathbf{V}$, where the covariance matrix $\mathbf{V}$ is non-diagonal due to the correlations between y_i and g_i. The generalized least squares estimator is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y},$$

and the covariance matrix for $\hat{\boldsymbol{\beta}}$ is

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}.$$

If the residuals are assumed to be normally distributed, the MLE of $\boldsymbol{\beta}$ is the same as the GLS estimator.

We will analyze the GLS estimator, especially the slope estimator in the following. To make the analysis tractable, we consider a special case.

Assumption 3 $\mathbf{V}$ is a positive definite matrix such that y_i is correlated with g_i only when $i = j$ with correlation ρ.

Under Assumptions 1 (i), 1 (iii) and the variance of $\hat{\beta}_1^D$ is given by

$$\mathrm{Var}(\hat{\beta}_1^D) = \frac{\sigma^2}{\frac{1}{1-\rho^2}\left(\sum_{i=1}^{k} x_i^2 - k\bar{x}^2\right) + k\frac{\sigma^2}{\sigma_g^2}}, \tag{9}$$

If $0 < \sigma^2 < \infty$, $0 < \sigma_g^2 < \infty$ and $-1 < \rho < 1$, then we can show that $\mathrm{Var}(\hat{\beta}_1^D)$ in (9) is smaller than $\mathrm{Var}(\hat{\beta}_1^L)$.

Generally, ρ is unknown and must be estimated based on the data. The theoretical analysis indicates potential for variance reduction from estimating the correlation. However, the extra computational budget spent on estimating ρ must be traded off with any potential performance gains.

3 AUGMENTING STOCHASTIC KRIGING

Given an experiment design $(\mathbf{x}_i, n_i)$, $i = 1, 2, \cdots, k$, stochastic kriging introduced by Ankenman et al. (2010) models the simulation output $y_j(\mathbf{x}_i)$ from jth replication at design point $\mathbf{x}_i$ as:

$$y_j(\mathbf{x}_i) = \mathbf{f}(\mathbf{x}_i)'\boldsymbol{\beta} + \mathsf{M}(\mathbf{x}_i) + \varepsilon_j(\mathbf{x}_i), \tag{10}$$

where $\mathbf{f}(\mathbf{x}_i) \in \mathbb{R}^p$ with known functions of $\mathbf{x}_i$, $\boldsymbol{\beta} \in \mathbb{R}^p$ with unknown parameters to be estimated, M is a realization of a zero-mean random field. The trend term $\mathbf{f}(\mathbf{x}_i)'\boldsymbol{\beta}$ represents the overall surface mean and the measurement error is denoted as $\varepsilon_j(\mathbf{x}_i)$. The uncertainties in M and ε_j are referred as extrinsic and intrinsic uncertainties, respectively. Denote the sample mean of response output and the average simulation noise at $\mathbf{x}_i$ as

$$\bar{y}(\mathbf{x}_i) = \frac{1}{n_i}\sum_{j=1}^{n_i} y_j(\mathbf{x}_i), \quad \bar{\varepsilon}(\mathbf{x}_i) = \frac{1}{n_i}\sum_{j=1}^{n_i} \varepsilon_j(\mathbf{x}_i),$$

with $\bar{\mathbf{y}} = (\bar{y}(\mathbf{x}_1), \bar{y}(\mathbf{x}_2), \cdots, \bar{y}(\mathbf{x}_k))'$.

Suppose we want to predict the response $y(\mathbf{x}_0)$ at $\mathbf{x}_0$. Let $\boldsymbol{\Sigma}_{\mathsf{M}}$ be the $k \times k$ covariance matrix implied by the random field M and $\boldsymbol{\Sigma}_\varepsilon$ be the $k \times k$ covariance matrix implied by the simulation noise across all design point $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k$. Let $\boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot) = (\mathrm{Cov}(y(\mathbf{x}_0), y(\mathbf{x}_1), \cdots, \mathrm{Cov}(y(\mathbf{x}_0), y(\mathbf{x}_k))'$ denote the covariances between $y(\mathbf{x}_0)$ and the responses from all design points. Also, let $\mathbf{F} = (\mathbf{f}(\mathbf{x}_1), \mathbf{f}(\mathbf{x}_2), \cdots, \mathbf{f}(\mathbf{x}_k))$ be the design matrix. The MSE-optimal predictor is of the form

$$\hat{y}(\mathbf{x}_0) = \mathbf{f}(\mathbf{x}_0)'\hat{\boldsymbol{\beta}} + \boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot)'(\boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}_\varepsilon)^{-1}(\bar{\mathbf{y}} - \mathbf{F}'\hat{\boldsymbol{\beta}}), \tag{11}$$

and the optimal MSE is

$$\mathrm{MSE}(\hat{y}(\mathbf{x}_0)) = \boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \mathbf{x}_0) - \boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot)'[\boldsymbol{\Sigma}_{\mathsf{M}} + \boldsymbol{\Sigma}_\varepsilon]^{-1}\boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot). \tag{12}$$

In an enhanced data setting, we observe the responses $y_j(\mathbf{x}_i)$ and the gradient estimates $\mathbf{g}_j(\mathbf{x}_i)$ for the jth simulation replication at design points $\mathbf{x}_i$. Instead of modeling the gradient estimates by partial derivative of the random field M as in Chen et al. (2011), we model $\mathbf{g}_j(\mathbf{x}_i)$ as a noise measurement of the true gradient $\mathbf{g}(\mathbf{x}_i)$, i.e., $\mathbf{g}_j(\mathbf{x}_i) = \mathbf{g}(\mathbf{x}_i) + \boldsymbol{\delta}_j(\mathbf{x}_i)$. Denote the sample mean of gradient estimates and the average simulation noise at $\mathbf{x}_i$ as

$$\bar{\mathbf{g}}(\mathbf{x}_i) = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{g}_j(\mathbf{x}_i), \quad \bar{\boldsymbol{\delta}}(\mathbf{x}_i) = \frac{1}{n_i}\sum_{j=1}^{n_i} \boldsymbol{\delta}_j(\mathbf{x}_i).$$

Notice that the response and the gradient estimates are noisy and usually correlated, and we assume that $\bar{\boldsymbol{\delta}}(\mathbf{x}_i)$ is independent of the random field M.

3.1 Augmenting Dataset with Gradient Estimates Via Extrapolation

To incorporate gradient estimates into stochastic kriging, we extrapolate in the neighborhood of the original design points $\{\mathbf{x}_i\}$, $i = 1,2,\cdots,k$, i.e., additional response data is generated via linear extrapolations using the gradient estimates as follows:

$$\mathbf{x}_i^+ = \mathbf{x}_i + \Delta\mathbf{x}_i, \qquad \bar{y}^+(\mathbf{x}_i^+) = \bar{y}(\mathbf{x}_i) + \bar{\mathbf{g}}(\mathbf{x}_i)\cdot\Delta\mathbf{x}_i. \tag{13}$$

Different extrapolation techniques can be applied in (13), and we can also add multiple points to the neighborhood of $\mathbf{x}_i$. In this preliminary study we assume that the same step size is used for all design points, i.e., $\Delta\mathbf{x}_i = \Delta x$, $i = 1,2,\cdots,k$. We also assume that only one additional point is added in the neighborhood of $\mathbf{x}_i$. Let $\bar{y}_i = \bar{y}(\mathbf{x}_i)$ and $\bar{y}_i^+ = \bar{y}(\mathbf{x}_i^+)$ for simplicity and $\bar{\mathbf{y}}^*$ be the $2k\times 1$ vector containing all the original response outputs and the additional response outputs in (13):

$$\bar{\mathbf{y}}^* = (\bar{y}_1, \bar{y}_2, \cdots, \bar{y}_k; \bar{y}_1^+, \bar{y}_2^+, \cdots, \bar{y}_k^+).$$

Similarly, $\mathbf{x}^+$ is defined as

$$\mathbf{x}^* = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k; \mathbf{x}_1^+, \mathbf{x}_2^+, \cdots, \mathbf{x}_k^+).$$

To fit this augmented dataset into the stochastic kriging approach, we model the additional points similar to the original response output, i.e.,

$$\bar{y}^+(\mathbf{x}_i^+) = \mathbf{f}(\mathbf{x}_i^+)'\boldsymbol{\beta} + \mathsf{M}(\mathbf{x}_i^+) + \varepsilon^+(\mathbf{x}_i^+),$$

and the variance of the noise $\varepsilon^+(\mathbf{x}_i^+)$ and the covariance between $\varepsilon^+(\mathbf{x}_i^+)$ and $\varepsilon(\mathbf{x}_i)$ are approximated by

$$\begin{aligned}
\mathrm{Var}\left(\varepsilon^+(\mathbf{x}_i^+)\right) &= \mathrm{Var}\left(\varepsilon(\mathbf{x}_i)\right) + (\Delta x)^2\,\mathrm{tr}\left[\mathrm{Cov}(\bar{\boldsymbol{\delta}}(\mathbf{x}_i))\right] + 2(\Delta x)\mathbf{1}'\mathrm{Cov}\left(\bar{\varepsilon}(\mathbf{x}_i), \bar{\boldsymbol{\delta}}(\mathbf{x}_i)\right),\\
\mathrm{Cov}\left(\varepsilon^+(\mathbf{x}_i^+), \varepsilon(\mathbf{x}_i)\right) &= \mathrm{Var}\left(\varepsilon(\mathbf{x}_i)\right) + \Delta x\mathbf{1}'\mathrm{Cov}\left(\varepsilon(\mathbf{x}_i), \bar{\boldsymbol{\delta}}(\mathbf{x}_i)\right).
\end{aligned}$$

Let $\boldsymbol{\Sigma}_{\mathsf{M}}^{\dagger} = \mathrm{Cov}[\mathsf{M}(\mathbf{x}_i), \mathsf{M}(\mathbf{x}_j^+)]$, $\boldsymbol{\Sigma}_{\mathsf{M}}^{+} = \mathrm{Cov}[\mathsf{M}(\mathbf{x}_i^+), \mathsf{M}(\mathbf{x}_j^+)]$, $i,j = 1,2,\cdots,k$, and $\boldsymbol{\Sigma}_{\mathsf{M}}^*$ be a $2k\times 2k$ covariance matrix across all the original design points and additional design points, which takes the form

$$\boldsymbol{\Sigma}_{\mathsf{M}}^* = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathsf{M}} & \boldsymbol{\Sigma}_{\mathsf{M}}^{\dagger} \\ \boldsymbol{\Sigma}_{\mathsf{M}}^{\dagger} & \boldsymbol{\Sigma}_{\mathsf{M}}^{+} \end{bmatrix}. \tag{14}$$

Similarly, let

$$\boldsymbol{\Sigma}_{\varepsilon}^* = \begin{bmatrix} \boldsymbol{\Sigma}_{\varepsilon} & \boldsymbol{\Sigma}_{\varepsilon}^{\dagger} \\ \boldsymbol{\Sigma}_{\varepsilon}^{\dagger} & \boldsymbol{\Sigma}_{\varepsilon}^{+} \end{bmatrix}, \tag{15}$$

where

$$\begin{aligned}
\boldsymbol{\Sigma}_{\varepsilon}^{\dagger} &= \mathrm{diag}\left\{\mathrm{Cov}\left(\varepsilon^+(\mathbf{x}_1^+), \varepsilon(\mathbf{x}_1)\right), \cdots, \mathrm{Cov}\left(\varepsilon^+(\mathbf{x}_k^+), \varepsilon(\mathbf{x}_k)\right)\right\},\\
\boldsymbol{\Sigma}_{\varepsilon}^{+} &= \mathrm{diag}\left\{\mathrm{Var}\left(\varepsilon^+(\mathbf{x}_1^+)\right), \cdots, \mathrm{Var}\left(\varepsilon^+(\mathbf{x}_k^+)\right)\right\}.
\end{aligned}$$

Let $\boldsymbol{\Sigma}_{\mathsf{M}}^*(\mathbf{x}_0,\cdot)$ be the covariance between $y(\mathbf{x}_0)$ and all $2k$ design points. Also, let $\mathbf{F}^* = (\mathbf{f}(\mathbf{x}_1), \mathbf{f}(\mathbf{x}_2), \cdots, \mathbf{f}(\mathbf{x}_k), \mathbf{f}(\mathbf{x}_1^+), \mathbf{f}(\mathbf{x}_2^+), \cdots, \mathbf{f}(\mathbf{x}_k^+))'$ be the design matrix. Under the enhanced data setting, we can easily find the MSE-optimal predictor and the corresponding MSE by substituting $\bar{\mathbf{y}}^*$, $\mathbf{F}^*$, $\boldsymbol{\Sigma}_{\mathsf{M}}^*(\mathbf{x}_0,\cdot)$, $\boldsymbol{\Sigma}_{\mathsf{M}}^*$ and $\boldsymbol{\Sigma}_{\varepsilon}^*$ for $\bar{\mathbf{y}}$, $\mathbf{F}$, $\boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0,\cdot)$, $\boldsymbol{\Sigma}_{\mathsf{M}}$ and $\boldsymbol{\Sigma}_{\varepsilon}$ in (11) and (12), respectively.

The random field M is assumed to be second-order stationary, i.e.,

$$\boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_i, \mathbf{x}_j) = \tau^2 R(\mathbf{x}_i - \mathbf{x}_j; \boldsymbol{\theta}),$$

where $\tau^2 = \mathrm{Var}[\mathsf{M}(\mathbf{x})]$ and $R(\mathbf{x}_i - \mathbf{x}_j; \boldsymbol{\theta})$ is a correlation function with parameter $\boldsymbol{\theta}$ depending on the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. The extended covariance matrix $\boldsymbol{\Sigma}_{\mathsf{M}}^*$ follows the same correlation structure, and the parameters $(\tau^2, \boldsymbol{\theta})$ and $\boldsymbol{\beta}$ can be estimated from maximum likelihood estimators (MLEs) provided that $\boldsymbol{\Sigma}_{\varepsilon}^*$ is known.

3.2 The Step Size

Incorporating the gradient estimates into stochastic kriging via extrapolation requires choice of step sizes. Different step sizes, even with the same data set, will lead to different stochastic kriging metamodels. The linear approximation is only appropriate in a small neighborhood of the design points, so the step size cannot be too large. If the size step is too small, the additional points obtained from linear approximation provide little information. Therefore, it is crucial to find rational choices for the step sizes. Depending on the chosen objectives, the optimal step sizes may vary.

The step size Δx can simply be treated as a new parameter. Therefore, we can estimate Δx with other parameters $(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta})$ simultaneously by MLE. Under Assumption 1 in Ankenman et al. (2010), the objective function can be written as

$$\text{Maximize} \quad \mathscr{L}(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta}, \Delta x) = -\ln\left[(2\pi)^k\right] - \frac{1}{2}\ln\left[|\boldsymbol{\Sigma}_{\mathsf{M}}^* + \boldsymbol{\Sigma}_{\varepsilon}^*|\right] - \frac{1}{2}(\bar{\mathbf{y}}^* - \mathbf{F}^{*\prime}\boldsymbol{\beta})'\left[\boldsymbol{\Sigma}_{\mathsf{M}}^* + \boldsymbol{\Sigma}_{\varepsilon}^*\right]^{-1}(\bar{\mathbf{y}}^* - \mathbf{F}^{*\prime}\boldsymbol{\beta}), \tag{16}$$

where $\bar{\mathbf{y}}^*$, $\boldsymbol{\Sigma}_{\mathsf{M}}^*$ and $\boldsymbol{\Sigma}_{\varepsilon}^*$ are functions of Δx.

Another reasonable objective is to choose the stepsize to minimize the integrated mean squared error (IMSE). Lower IMSE suggests smaller deviation associated with the approximation over the region of interest. The problem can be formulated as

$$\text{Minimize} \quad \text{IMSE} = \int_{\mathbf{x}_0 \in \Omega} \text{MSE}^*(\mathbf{x}_0; \Delta x) d\mathbf{x}_0, \tag{17}$$

where Ω is the region of interest and

$$\text{MSE}^*(\mathbf{x}_0; \Delta x) = \boldsymbol{\Sigma}_{\mathsf{M}}^*(\mathbf{x}_0, \mathbf{x}_0) - \boldsymbol{\Sigma}_{\mathsf{M}}^*(\mathbf{x}_0, \cdot)'\left[\boldsymbol{\Sigma}_{\mathsf{M}}^* + \boldsymbol{\Sigma}_{\varepsilon}^*\right]^{-1}\boldsymbol{\Sigma}_{\mathsf{M}}^*(\mathbf{x}_0, \cdot).$$

Both (16) and (17) are unconstrained optimization problems. However, in the optimization process, maintaining the invertibility of the matrix $\boldsymbol{\Sigma}_{\mathsf{M}}^* + \boldsymbol{\Sigma}_{\varepsilon}^*$ is crucial, as an ill-conditioned matrix causes numerical stability issues. Therefore, adding a constraint on the condition number of the matrix $\boldsymbol{\Sigma}_{\mathsf{M}}^* + \boldsymbol{\Sigma}_{\varepsilon}^*$ is necessary.

One key difference between the objectives in (16) and (17) is that the MLE approach estimates all the parameters $(\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta})$ with Δx simultaneously. However, the IMSE approach requires $\boldsymbol{\Sigma}_{\mathsf{M}}^*$ to be known in advance. In practice, a two-stage strategy can achieve this:

1. In Stage 1, use $(\mathbf{x}_i, \mathbf{y}_i)$ to obtain MLE $\hat{\boldsymbol{\beta}}$, $\hat{\tau}^2$ and $\hat{\boldsymbol{\theta}}$.
2. Calculate $\text{MSE}^*(\mathbf{x}_0; \Delta x)$ with $\hat{\boldsymbol{\beta}}$, $\hat{\tau}^2$ and $\hat{\boldsymbol{\theta}}$ as a function of Δx.
3. In Stage 2, maximize IMSE in (17) as function of Δx.

It is worth mentioning that the problem of finding step sizes is closely related to the experiment design of stochastic kriging. To develop an experiment design, we need to decide the locations of design points $\mathbf{x}_i$ and allocate number of replications n_i to $\mathbf{x}_i$. In stochastic kriging, the locations of design points control the way we exploit the region of interest, while the number of replications determines sizes of variances of intrinsic noise. In the problem of finding step sizes, the k newly added design points depend on step sizes, and the variances of intrinsic noise $\boldsymbol{\Sigma}_{\varepsilon}^*$ also depends on step sizes.

The IMSE criterion was used in Ankenman, Nelson, and Staum (2010) to develop experiment designs for stochastic kriging. They concluded that both intrinsic and extrinsic uncertainty matter in the experiment design and suggested to choose design points that are centrally located. However, there is no simple extension of their results to determine step sizes.

3.3 Enhanced Experiment Design in Stochastic Kriging

In this section we investigate the potential to improve the experiment design for stochastic kriging with gradient estimates. The convexity information suggested by gradient estimates can be used to improve

experiment design. Consider a one-dimensional problem. If the gradient estimates at two consecutive design points x_i and x_{i+1} satisfy $g(x_i)g(x_{i+1}) < 0$, then it suggests that with a high probability that there exist local extreme values in $[x_i, x_{i+1}]$. Although the benefit of adding another design point between x_i and x_{i+1} is not fully justified, we believe that it helps the stochastic kriging metamodel to capture fluctuations in the response surface and offer accurate locations of optimizers. However, this can be only applied to response surfaces that have local extreme values within the region of interest.

In practice, given a total replication budget N, we propose the following two-stage design policy based on these results:

1. In Stage 1, implement a space-filling design with m design points $x_1, x_2, \cdots, x_m$ and allocate n_1 number of replications to each design point.
2. Collect simulation outputs $\bar{y}(x_1), \bar{y}(x_2), \cdots, \bar{y}(x_m)$ and $\bar{g}(x_1), \bar{g}(x_2), \cdots, \bar{g}(x_m)$.
3. For $\bar{g}(x_i)\bar{g}(x_{i+1}) < 0$, add another design point in between x_i and x_{i+1}. The new design point can be determined by fitting a linear model to the gradient surface within $[x_i, x_{i+1}]$.
4. In Stage 2, suppose there are k design points, including all x_i and the extrapolated design points, after Step 3. Allocate the remaining $N - mn_1$ replications to the k design points using the IMSE based experiment design proposed in Ankenman et al. (2010).

4 NUMERICAL EXAMPLES

We conduct some numerical experiments to illustrate DiGAR and GESK, to empirically investigate their properties in practice, and to compare with standard regression models and stochastic kriging. In the first example, we consider the mean total time for a customer (not the steady-state time in system), e.g., mean total time for the 2nd customer, in a first-come, first-served, single-server queue, where only DiGAR and standard regression models are considered. In the second example, we consider an example from Santner et al. (2003) with artificial noise to compare stochastic kriging (SK), stochastic kriging with gradient estimators (SKG) in Chen et al. (2011) and gradient extrapolated stochastic kriging (GESK). The third example considers the steady-state waiting time in an M/M/1 queue, where all these techniques for metamodels are investigated. Software for stochastic kriging is downloaded from http://www.stochastickriging.net, and code for SKG and GESK are modified based on this.

4.1 Example 1

In an M/M/1 queue with arrival rate 1/5, we consider the expected system time in system for each customer as a function of the mean service time. We compute the true theoretical dependence of the expected system time on the mean service time, which is used to judge the quality of the fitted curve. We choose 10 equally spaced design points in $[3.6, 4.5]$ and run 10 replications on each design point. The simulated data, true model and fitted models are plotted in Figure 1, with $y^{(i)}$ indicating the expected system time for ith customer. All methods fit the model reasonably well for the $2^{\text{nd}}$and 3^{rd} customer, but there are dramatic differences in $y^{(4)}$ and $y^{(5)}$. The slope of the OLS model has the incorrect sign, whereas all the DiGAR models capture the correct orientation of the curve. Independent DiGAR and DiGAR with normality assumption (DiGARn) are closer to the true model, compared to correlated DiGAR (DiGAR*). Many more numerical results using this example are contained in (Fu and Qu 2012).

4.2 Example 2

We take an example from Santner et al. (2003) to compare SK, SKG and GESK. Suppose the function is $y(x) = \exp(-1.4x)\cos(7\pi x/2) + \varepsilon$ over $-2 \le x \le 0$, where the noise $\varepsilon \sim \mathcal{N}(0,1)$. Direct gradient estimate is not available in this case, and we use $g(x) = y'(x) + \delta$ for gradient estimates where $\delta \sim \mathcal{N}(0, 25)$. We use a Gaussian correlation function $R(x, x') = \exp\{\theta(x - x')^2\}$ and choose 4 different experimental designs. The number of design points and the number of replications are 6 and 20, 8 and 20, 8 and 200, 20 and 200.

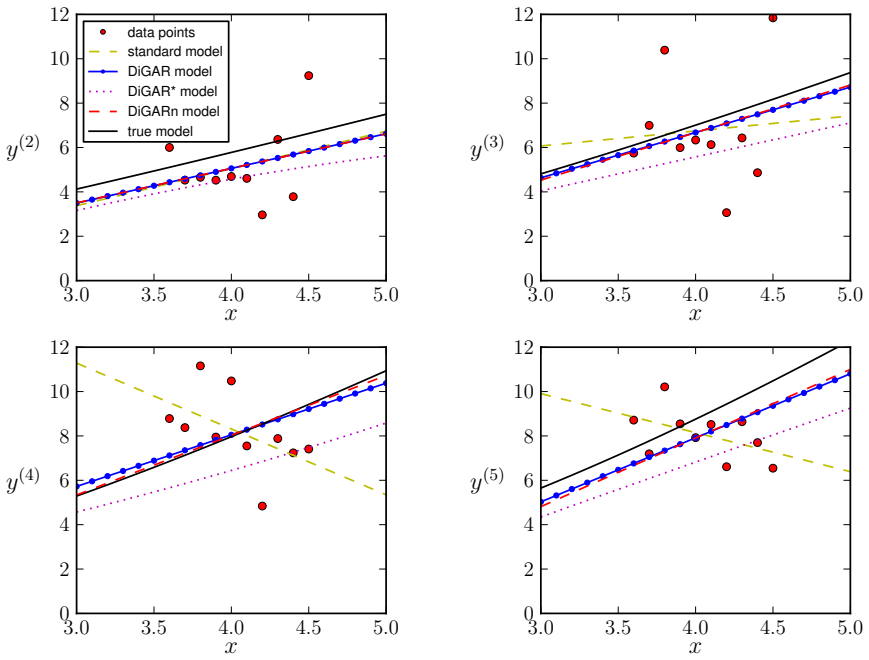

Figure 1: Expected time in system for each customer.

Predictions using 6 design points are shown in the leftmost plot in Figure 2. The number of design points is too small for SK and SKG to explore the design space, but GESK is able to explore the design space more via extrapolation. Although predictions from GESK are not that accurate, they still capture the fluctuations of the response surface. Results from experiment 2 are shown in the middle left plot. The fitted model using SK is close to the true model. Both SKG and GESK improve the fitting by incorporating gradient estimates, and improvement from SKG is bigger.

Another observation in this example is that predictions may become worse as the number of replications increases. These behaviors are illustrated in experiment 3 and 4. In experiment 3, the number of replications at each design point is increased to 200 (20 replications at each design point in experiment 2). The prediction using SK is even worse than experiment 2, but both SKG and GESK improve their predictions compared to experiment 2. In experiment 4, the number of design points is 20 and the number of replications at each design point is 200. GESK is the only method that make predictions close to the true values. This is an unexpected result. Generally, when we increase the number of replications, the effect of noises decreases and the prediction should be more accurate. A possible explanation is that as number of replications increases, the effect of intrinsic noise can be negligible. As mentioned in Staum (2009), stochastic kriging becomes like kriging when intrinsic variances are negligible. Kriging interpolates all response outputs, which is exactly what we observe for SK and SKG in experiment 4.

4.3 Example 3

We consider an M/M/1 queue with arrival rate 1. As a function of the service rate x, the expected steady-state waiting time is given by $y(x) = 1/(x(x-1))$. Metamodels fitted by different techniques are shown in Figure 3. The number of design points and the number of replications for each design point used in the left and middle plots are 5 and 1000, 20 and 250, each replication with 100 customers. The right plot uses 50 design points, 2000 replications for each design point and 1000 customers in each replication.

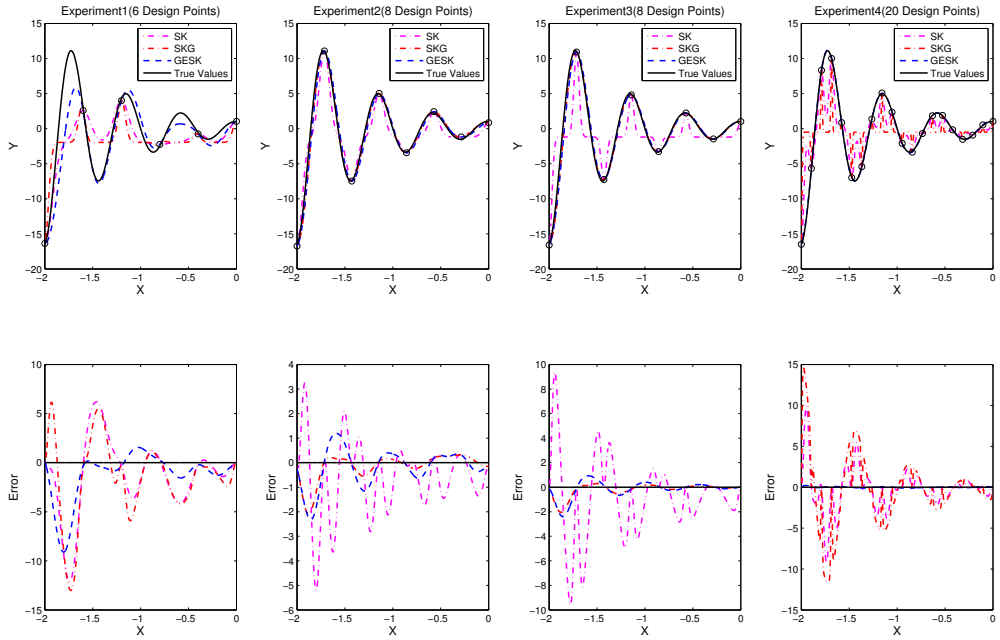

Figure 2: Test function: $y(x) = \exp(-1.4x)\cos(7\pi x/2) + \varepsilon$.

A quadratic model $y(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ is fitted via OLS regression and DiGAR, respectively. Both regression models do not fit the true model well. The reason for poor predictions is that regression models are misspecified, i.e., there is no quadratic function can precisely fit the expected waiting time.

We use constant trends and Gaussian correlation functions for SK, SKG and GESK. It is obvious that the fitted models are all better than regression models. The right plot shows that all three methods can fit the model really well when the computing budget is large enough. However, we can see performance differences when the number of design points is small. In the left plot, GESK and SKG are better than SK in relative errors; in the middle plot, GESK and SK are close and slightly better than SKG visually. Therefore, the improvement from incorporating gradient estimates is more significant when the computing budget is small, and GESK has good prediction consistently compared to the other two methods.

5 CONCLUSIONS AND FUTURE RESEARCH

In this paper we investigated both augmented regression models and augmented stochastic kriging models that exploit the availability of direct gradient estimates in stochastic simulation settings. Both augmented models are qualitatively able to capture trends that the standard models might miss. For stochastic kriging, preliminary numerical experiments indicate the following:

- GESK and SKG both improve predictions by incorporating gradient estimates.
- GESK doesn't require differentiability on the chosen correlation function, unlike SKG.
- GESK makes good predictions in different cases with good choice of step sizes.
- The performance of GESK highly depends on the choice of step sizes.

Our work points to several other directions for future research. The first direction is to apply DiGAR in simulation-based optimization using sequential RSM. The expected improvements in optimization efficiency from DiGAR models need to be characterized and quantified in theoretical work and numerical experimentations. Another direction is to focus on in-depth exploration of rational choice of step sizes used in GESK and develop multi-stage experiment design for stochastic kriging using gradient information.

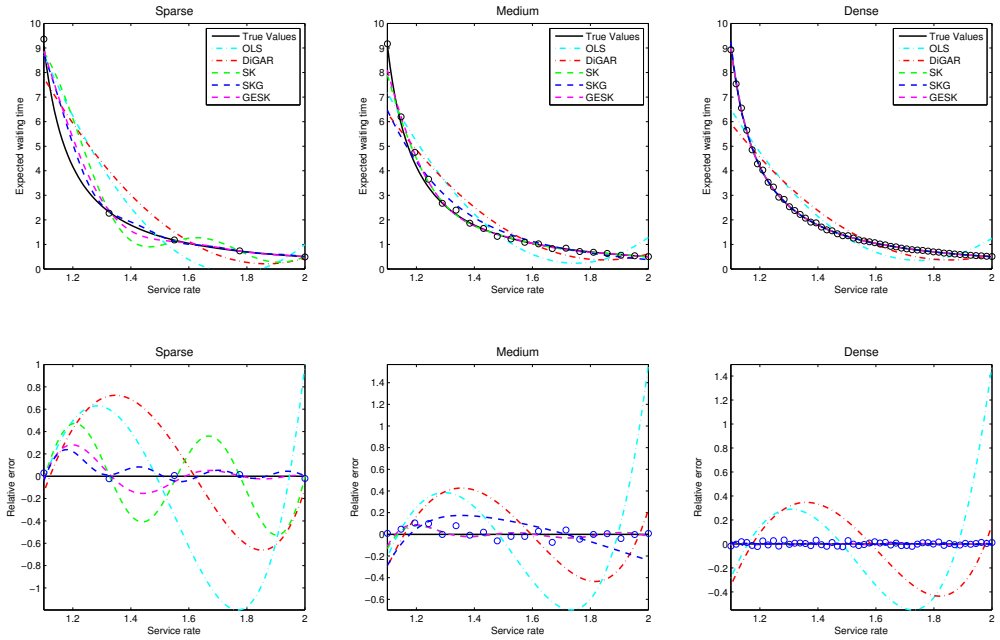

Figure 3: Expected waiting time in M/M/1 queue.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation (NSF) under Grants CMMI-0856256, EECS-0901543, and by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-10-10340.

REFERENCES

Ankenman, B., B. L. Nelson, and J. Staum. 2010, March. "Stochastic Kriging for Simulation Metamodeling". *Operations research* 58 (2): 371–382.

Asmussen, S., and P. Glynn. 2007. *Stochastic Simulation: Algorithms and Analysis*. New York: Springer.

Barton, R. R. 2009, December. "Simulation Optimization Using Metamodels". In *Proceedings of the 2009 Winter Simulation Conference*, edited by M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 230–238. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Barton, R. R., and M. Meckesheimer. 2006. "Metamodel-Based Simulation Optimization". In *Handbooks in Operations Research and Management Science: Simulation*, edited by S. G. Henderson and B. L. Nelson, Chapter 18, 535–574. Elsevier.

Chen, X., B. Ankenman, and B. L. Nelson. 2011. "Enhancing Stochastic Kriging Metamodels with Gradient Estimators". Working Paper.

Cressie, N. 1993. *Statistics for spatial data*. 2, revised ed. Wiley series in probability and mathematical statistics: Applied probability and statistics. New York: John Wiley and Sons.

Fu, M. C. 1994. "Optimization via Simulation: A Review". *Annals of Operations Research* 53:199–248.

Fu, M. C. 2002. "Optimization for Simulation: Theory vs. Practice (Feature Article)". *INFORMS Journal on Computing* 14 (3): 192–215.

Fu, M. C. 2008. "What you should know about simulation and derivatives". *Naval Research Logistics* 55 (8): 723–736.

Fu, M. C., C. H. Chen, and L. Shi. 2008, December. "Some Topics in Simulation Optimization". In *Proceedings of the 2008 Winter Simulation Conference*, edited by S. J. Mason, R. R. Hill, L. Moench, O. Rose, T. Jefferson, and J. W. Fowler, 27–38. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Fu, M. C., and H. Qu. 2012. "Augmented Regression with Direct Gradient Estimates". Working Paper.
Glasserman, P. 1991. *Gradient Estimation via Perturbation Analysis*. Kluwer Academic Publishers, Boston, Massachusetts.
Glasserman, P. 2004. *Monte Carlo Methods in Financial Engineering*. Springer, New York.
Ho, Y., L. Shi, L. Dai, and W. Gong. 1992, January. "Optimizing discrete event dynamic systems via the gradient surface method". *Discrete Event Dynamic Systems: Theory and Applications* 2:99–120.
Ho, Y. C., and X. Cao. 1983. "Perturbation analysis and optimization of queueing networks". *Journal of Optimization Theory and Applications* 40:559–582.
Ho, Y. C., and X. R. Cao. 1991. *Perturbation Analysis and Discrete Event Dynamic Systems*. Kluwer Academic.
Kleijnen, J. 2008. *Design and Analysis of Simulation Experiments*. New York: Springer.
Kleijnen, J. P. C., W. C. M. Van Beers, and I. v. Nieuwenhuyse. 2010, April. "Constrained optimization in expensive simulation: Novel approach". *European Journal of Operational Research* 202 (1): 164–174.
Liu, W. 2003. *Development of gradient-enhanced kriging approximations for multidisciplinary design optimization*. Ph. D. thesis, University of Notre Dame.
Qu, H., and M. C. Fu. 2012. "Augmented Stochastic Kriging with Direct Gradient Estimates". Working Paper.
Rubinstein, R. Y. 1986. *Monte Carlo Optimization, Simulation and Sensitivity of Queueing Networks*. John Wiley & Sons.
Rubinstein, R. Y., and A. Shapiro. 1993. *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*. John Wiley & Sons.
Santner, T. J., B. Williams, and W. Notz. 2003. *The Design and Analysis of Computer Experiments*. Springer-Verlag.
Staum, J. 2009, December. "Better simulation metamodeling: The why, what, and how of stochastic kriging". In *Proceedings of the 2009 Winter Simulation Conference*, edited by M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 119–133. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.
Stein, M. 1999. *Interpolation of spatial data: some theory for kriging*. Springer series in statistics. New York: Springer-Verlag.
Yang, F., B. Ankenman, and B. L. Nelson. 2007. "Efficient generation of cycle time-throughput curves through simulation and metamodeling". *Naval Research Logistics* 54 (1): 78–93.

AUTHOR BIOGRAPHIES

HUASHUAI QU is a Ph.D. student in Department of Mathematics at the University of Maryland. He received the B.S. degree in information and computational science in Beijing Jiaotong University in 2006. His research interest lies in the areas of simulation optimization and optimal learning. His email address is huashuai@math.umd.edu

MICHAEL C. FU is Ralph J. Tyser Professor of Management Science in the Robert H. Smith School of Business, with a joint appointment in the Institute for Systems Research and an affiliate appointment in the Department of Electrical and Computer Engineering, all at the University of Maryland. He received degrees in mathematics and EE/CS from MIT, and an M.S. and Ph.D. in applied mathematics from Harvard University. His research interests include simulation and applied probability modeling, particularly with applications towards manufacturing systems, supply chain management, and financial engineering. He has served as Stochastic Models and Simulation Department Editor of Management Science and as Simulation Area Editor of Operations Research. He is co-author (with J.Q. Hu) of the book, Conditional Monte Carlo: Gradient Estimation and Optimization Applications, which received the INFORMS Simulation Societys Outstanding Publication Award in 1998. He is a Fellow of INFORMS and IEEE. His email address is mfu@umd.edu.