OPERATIONS RESEARCH
Vol. 61, No. 2, March–April 2013, pp. 512–528
ISSN 0030-364X (print) | ISSN 1526-5463 (online)

http://dx.doi.org/10.1287/opre.1120.1143

Enhancing Stochastic Kriging Metamodels with Gradient Estimators

Xi Chen
Department of Statistical Sciences and Operations Research, Virginia Commonwealth University, Richmond, Virginia 23284, xchen4@vcu.edu

Bruce E. Ankenman, Barry L. Nelson
Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60208
{ankenman@northwestern.edu, nelsonb@northwestern.edu}

Stochastic kriging is a new metamodeling technique for effectively representing the mean response surface implied by a stochastic simulation; it takes into account both stochastic simulation noise and uncertainty about the underlying response surface of interest. We show theoretically, through some simplified models, that incorporating gradient estimators into stochastic kriging tends to significantly improve surface prediction. To address the issue of which type of gradient estimator to use, when there is a choice, we briefly review stochastic gradient estimation techniques; we then focus on the properties of infinitesimal perturbation analysis and likelihood ratio/score function gradient estimators and make recommendations. To conclude, we use simulation experiments with no simplifying assumptions to demonstrate that the use of stochastic kriging with gradient estimators provides more reliable prediction results than stochastic kriging alone.

Subject classifications: stochastic simulation; metamodeling; gradient estimation.
Area of review: Simulation.
History: Received November 2011; revisions received June 2012, September 2012; accepted October 2012.

1. Introduction

Simulation models, both deterministic and stochastic, can provide high-fidelity predictions of the behavior of complex or complicated systems at different settings of their controllable factors or decision variables. However, simulation runs may be (and frequently are) time consuming to execute, potentially limiting the usefulness of simulation in some settings, including support for real-time decision making and system optimization. To mitigate this deficiency, carefully designed simulation experiments can be employed to fit metamodels—equation-based approximations of some aspect of system performance, such as the mean—and the metamodels may be exercised in real time, or searched efficiently using nonlinear optimization methods. Of course, to be useful, the metamodels need to be accurate, and the experiment design to fit them should not be too computationally expensive to execute.

In the stochastic simulation community there is a long history of research on experiment design for fitting regression metamodels to simulation output. Unfortunately, standard linear regression models tend not to provide good global representations of the response surface over the feasible space of controllable factors, and we are particularly interested in applications where a global model is valuable, such as using the metamodel for real-time decision support. In the design and analysis of computer experiments (DACE) community—where the simulation models are typically deterministic and very expensive to run—the use of semiparametric "kriging" metamodels has been remarkably effective for global metamodeling (see for instance, Santner et al. 2003). More recently Ankenman et al. (2010) introduced *stochastic kriging* as a tool for representing stochastic simulation response surfaces. The benefit of kriging and stochastic kriging as metamodels, however, comes with a price: although the predicted surface may have low error, it is difficult to enforce known properties like monotonicity, or to avoid bumpiness of the predicted surface because of mean reversion (e.g., Siem 2008). In DACE it has been known for some time that incorporating known gradient information in the construction of the metamodel can mitigate such deficiencies (which is quite different from using kriging models to estimate the gradients themselves). We investigate whether similar benefits can be obtained in the stochastic simulation setting.

To the best of our knowledge this idea is new and has yet to be exploited in stochastic simulation metamodeling. One reason that it is promising is that gradient estimation in stochastic simulation has been studied extensively, and there exist gradient estimators whose properties are well understood and that are ready to use for large classes of problems (e.g., queueing networks; see Glasserman 1991). Further, unlike the typical deterministic gradient calculation in DACE, the additional computational burden to obtain a gradient estimator in stochastic simulation is often negligible compared to the effort required to obtain the response itself. Therefore, the research question is whether including

stochastic gradient estimators leads to better prediction, rather than whether the improvement is worth the extra computation. Assuming stochastic gradient estimators are helpful, we are also interested in which of the well-known types of gradient estimators it is better to use. As discussed more fully in §4, if the additional burden is substantial, as it is for finite-difference gradient estimators, then the effort would be better spent on additional experiment design points.

The remainder of this paper is organized as follows. In §2 we build the framework for stochastic kriging with gradient estimators. In §3 we analyze the effects on prediction of incorporating gradient estimators through simplified two-design-point and k-design-point models and try to infer desirable properties of the potential gradient estimators. We review stochastic gradient estimators in §4, with a focus on infinitesimal perturbation analysis (IPA) and likelihood ratio/score function (LR/SF) approaches, and we analyze their respective properties in stochastic kriging metamodels. We conclude the paper with two experiments that apply stochastic kriging with gradient estimators in §§5 and 6.

Here is a summary of what we will show: For simple two-design-point and k-design-point models with known parameters, we will prove that incorporating gradient estimators provides better prediction of the response surface in terms of the mean squared error of prediction. Further, when we have a choice of unbiased gradient estimators, we show that one that has lower variance and stronger correlation with the response estimate is preferred, which (as we also show) tends to favor IPA. This theoretical analysis is backed up by simulation experiments in which all parameters are estimated, as would be done in practice. Our experiments include predicting the value of a call option as a function of its underlying stock price, where we find that incorporating either IPA and LR/SF estimators of the Black-Scholes delta leads to substantially better prediction; this is critical in option pricing applications. A second experiment examines a realistic problem of predicting the throughput of a closed-loop flexible assembly system. Again we find that incorporating IPA gradient estimators into stochastic kriging significantly improves the prediction performance over stochastic kriging without gradient estimators.

2. The Model Formulation

In this section we provide background on kriging metamodels that incorporate gradient information, and introduce our approach for exploiting gradient estimators in stochastic kriging.

2.1. Kriging

Although originating in geostatistics, kriging has become increasingly popular in engineering design following the work of Sacks et al. (1989) who applied the method to approximate the output of deterministic computer experiments. In a deterministic computer experiment, the response $\mathsf{Y}(\mathbf{x})$ is observed without noise and a metamodel is developed after observing $\mathsf{Y}(\mathbf{x})$ at some design points $\mathbf{x}_i$, with $\mathbf{x}_i \in \Re^d$. The unknown response surface is represented by

$$\mathsf{Y}(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top \boldsymbol{\beta} + \mathsf{M}(\mathbf{x}), \tag{1}$$

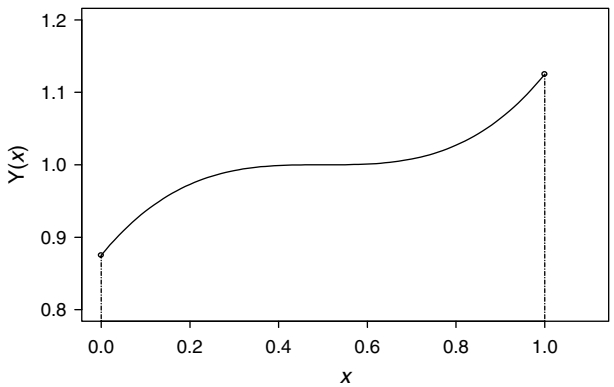
where $\mathbf{f}(\mathbf{x})$ and $\boldsymbol{\beta}$ are, respectively, a $p \times 1$ vector of known functions of $\mathbf{x}$, and a $p \times 1$ vector of unknown parameters. This form of the model is sometimes called "universal kriging." In kriging it is assumed that M is a realization of a mean zero stationary Gaussian random process (or random field) of second order (i.e., $\mathrm{E}[|\mathsf{M}(\mathbf{x})|^2] < \infty$ for any $\mathbf{x} \in \Re^d$), which can be thought of as being sampled from a space of functions mapping $\Re^d \to \Re$. The functions in this space are assumed to exhibit spatial correlation: the values $\mathsf{M}(\mathbf{x}_h)$ and $\mathsf{M}(\mathbf{x}_l)$ will tend to be similar if $\mathbf{x}_h$ and $\mathbf{x}_l$ are close to each other in $\Re^d$. Specifically, the covariance function between the responses at design points is usually described by τ^2, the spatial variance of the random process, and a correlation function $\mathcal{R}(\cdot, \cdot)$; that is, at $\mathbf{x}_h$ and $\mathbf{x}_l$,

$$\begin{aligned}\mathrm{Cov}[\mathsf{Y}(\mathbf{x}_h), \mathsf{Y}(\mathbf{x}_l)] &= \mathrm{Cov}[\mathsf{M}(\mathbf{x}_h), \mathsf{M}(\mathbf{x}_l)] \\ &= \tau^2 \mathcal{R}(\mathbf{x}_h, \mathbf{x}_l).\end{aligned} \tag{2}$$

We introduce a simple two-design-point model for illustration and continue building this example throughout §2. Suppose that $d = 1$ and we observe the responses $\mathsf{Y}(x_1), \mathsf{Y}(x_2)$ with no simulation noise at design points $x_i \in \Re$, $i = 1, 2$. Without loss of generality, let $x_1 = 0$ and $x_2 = 1$. Figure 1 illustrates the idea of kriging. The curve $\mathsf{Y}(x)$ is the unknown response surface of interest. Kriging uses the observed responses $\mathsf{Y}(x_i)$ at design points x_i, $i = 1, 2$ (denoted by ∘ on the curve) to predict the responses at other points.

Now suppose that a simulation experiment has been run at k design points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$, and we want to predict

Figure 1. A two-design-point model for kriging.



Note. Curve is the true response function $\mathsf{Y}(x)$, and the ∘'s are the observed values at x_1 and x_2.

the response at $\mathbf{x}_0$. For ease of exposition, we abuse the notation slightly by letting $\mathsf{Y} = (\mathsf{Y}(\mathbf{x}_1), \mathsf{Y}(\mathbf{x}_2), \ldots, \mathsf{Y}(\mathbf{x}_k))^\top$ be the vector of observed responses. Let $\boldsymbol{\Sigma}_\mathsf{M}$ be the $k \times k$ variance–covariance matrix that contains the spatial covariance between the responses at a pair of design points, i.e., $\Sigma_\mathsf{M}(\mathbf{x}_h, \mathbf{x}_l) = \tau^2 \mathcal{R}(\mathbf{x}_h, \mathbf{x}_l)$ because of (2). Take the $k \times p$ matrix of regression functions $\mathbf{F}$ as $\mathbf{F} = (\mathbf{f}(\mathbf{x}_1)^\top, \mathbf{f}(\mathbf{x}_2)^\top, \ldots, \mathbf{f}(\mathbf{x}_k)^\top)^\top$. Let $\boldsymbol{\Sigma}_\mathsf{M}(\mathbf{x}_0, \cdot)$ be the $k \times 1$ vector of spatial covariances between $\mathsf{Y}(\mathbf{x}_0)$ and the responses at $\mathbf{x}_i, i = 1, 2, \ldots, k$, i.e., $\boldsymbol{\Sigma}_\mathsf{M}(\mathbf{x}_0, \cdot) = \tau^2(\mathcal{R}(\mathbf{x}_0, \mathbf{x}_1), \mathcal{R}(\mathbf{x}_0, \mathbf{x}_2), \ldots, \mathcal{R}(\mathbf{x}_0, \mathbf{x}_k))^\top$.

Sections 1.2 and 1.5 in Stein (1999) give the following results. When the spatial parameters (hence $\boldsymbol{\Sigma}_\mathsf{M}$ and $\boldsymbol{\Sigma}_\mathsf{M}(\mathbf{x}_0, \cdot)$) and the vector $\boldsymbol{\beta}$ are known, then the MSE-optimal predictor (which is also the best linear predictor (BLP)), which minimizes the mean squared error among linear predictors of the response at $\mathbf{x}_0$, is

$$\hat{\mathsf{Y}}(\mathbf{x}_0) = \mathbf{f}(\mathbf{x}_0)^\top \boldsymbol{\beta} + \boldsymbol{\Sigma}_\mathsf{M}(\mathbf{x}_0, \cdot)^\top \boldsymbol{\Sigma}_\mathsf{M}^{-1}(\mathsf{Y} - \mathbf{F}\boldsymbol{\beta}), \tag{3}$$

with mean squared error of prediction (MSE)

$$\mathrm{MSE}(\hat{\mathsf{Y}}(\mathbf{x}_0)) = \Sigma_\mathsf{M}(\mathbf{x}_0, \mathbf{x}_0) - \boldsymbol{\Sigma}_\mathsf{M}(\mathbf{x}_0, \cdot)^\top \boldsymbol{\Sigma}_\mathsf{M}^{-1} \boldsymbol{\Sigma}_\mathsf{M}(\mathbf{x}_0, \cdot). \tag{4}$$

When $\boldsymbol{\Sigma}_\mathsf{M}$ and $\boldsymbol{\Sigma}_\mathsf{M}(\mathbf{x}_0, \cdot)$ are known, but $\boldsymbol{\beta}$ is unknown and needs to be estimated, the MSE-optimal predictor (which is also the best linear unbiased predictor (BLUP)) of the response at $\mathbf{x}_0$ is

$$\hat{\mathsf{Y}}(\mathbf{x}_0) = \mathbf{f}(\mathbf{x}_0)^\top \hat{\boldsymbol{\beta}} + \boldsymbol{\Sigma}_\mathsf{M}(\mathbf{x}_0, \cdot)^\top \boldsymbol{\Sigma}_\mathsf{M}^{-1}(\mathsf{Y} - \mathbf{F}\hat{\boldsymbol{\beta}}), \tag{5}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{F}^\top \boldsymbol{\Sigma}_\mathsf{M}^{-1} \mathbf{F})^{-1} \mathbf{F}^\top \boldsymbol{\Sigma}_\mathsf{M}^{-1} \mathsf{Y}$ is the generalized least squares estimator. The MSE of $\hat{\mathsf{Y}}(\mathbf{x}_0)$ equals

$$\begin{aligned}\mathrm{MSE}(\hat{\mathsf{Y}}(\mathbf{x}_0)) = {} & \Sigma_\mathsf{M}(\mathbf{x}_0, \mathbf{x}_0) - \boldsymbol{\Sigma}_\mathsf{M}(\mathbf{x}_0, \cdot)^\top \boldsymbol{\Sigma}_\mathsf{M}^{-1} \boldsymbol{\Sigma}_\mathsf{M}(\mathbf{x}_0, \cdot) \\ & + \eta^\top (\mathbf{F}^\top \boldsymbol{\Sigma}_\mathsf{M}^{-1} \mathbf{F})^{-1} \eta,\end{aligned} \tag{6}$$

where $\eta = f(\mathbf{x}_0) - \mathbf{F}^\top \boldsymbol{\Sigma}_\mathsf{M}^{-1} \boldsymbol{\Sigma}_\mathsf{M}(\mathbf{x}_0, \cdot)$. Equations (3) and (5) are the kriging metamodels. We next enhance them by incorporating gradients.

2.2. Kriging with Gradient Information

Research on metamodels with gradient information has been conducted in the context of deterministic computer experiments. See Morris et al. (1993), chapter 4 of Santner et al. (2003), Näther and Šimák (2003), Šimák (2002), and Stephenson (2010) for modeling and experiment design issues; for metamodel–based optimization, see for instance, Forrester and Keaney (2009) and Yamazaki et al. (2010). A key research issue in deterministic computer simulation is whether the (often substantial) computational effort required to generate gradients might be better spent on additional response estimates. This is in contrast to our research question as to whether there is value in incorporating gradient *estimates*, because the computational cost of obtaining them is often not significant.

To introduce the idea of kriging with gradient information, we start with the derivative of a stochastic process. Consider the stochastic process $\{\mathsf{Y}(\mathbf{x}), \mathbf{x} \in \Re^d\}$ defined by Equation (1) with mean function $\mathbf{f}(\mathbf{x})^\top \boldsymbol{\beta}$ and covariance given in Equation (2). If $\{\mathsf{M}(\mathbf{x}), \mathbf{x} \in \Re^d\}$ is a zero-mean covariance-stationary Gaussian stochastic process, then the first-order partial derivative processes $\mathsf{D}^r(\mathbf{x})$, $r = 1, 2, \ldots, d$ are defined as follows:

$$\mathsf{D}^r(\mathbf{x}) = \frac{\partial \mathsf{Y}(\mathbf{x})}{\partial x_r} = \left(\frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_r}\right)^\top \boldsymbol{\beta} + \frac{\partial \mathsf{M}(\mathbf{x})}{\partial x_r}, \tag{7}$$

$$\frac{\partial \mathsf{M}(\mathbf{x})}{\partial x_r} = \lim_{t \to 0} \frac{\mathsf{M}(\mathbf{x} + t\mathbf{e}_r) - \mathsf{M}(\mathbf{x})}{t}, \quad r = 1, 2, \ldots, d;$$

where $\mathbf{e}_r$ is the $d \times 1$ unit vector with the rth element being one while the others are zero. Notice that the limit is taken in the sense of convergence in mean square. Sufficient conditions for Equation (7) to hold are given in chapter 3 of Parzen (1962): the mean function $\mathbf{f}(\mathbf{x})^\top \boldsymbol{\beta}$ is differentiable and the second-order mixed derivative of $\mathcal{R}(\mathbf{x}_h, \mathbf{x}_l)$ exists and is continuous. Under these conditions the operations of differentiation and expectation can be interchanged. Therefore, the general first-order partial derivative process $\mathsf{D}^r(\mathbf{x})$, $r = 1, 2, \ldots, d$ of $\mathsf{Y}(\mathbf{x})$ is Gaussian, with mean function

$$\mathrm{E}[\mathsf{D}^r(\mathbf{x})] = \left(\frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_r}\right)^\top \boldsymbol{\beta}$$

and covariance of the first-order partial derivatives at a pair of points $\mathbf{x}_h, \mathbf{x}_l \in \Re^d$ being

$$\begin{aligned}\mathrm{Cov}[\mathsf{D}^r(\mathbf{x}_h), \mathsf{D}^s(\mathbf{x}_l)] &= \mathrm{Cov}\left[\frac{\partial}{\partial x_{hr}} \mathsf{Y}(\mathbf{x}_h), \frac{\partial}{\partial x_{ls}} \mathsf{Y}(\mathbf{x}_l)\right] \\ &= \frac{\partial}{\partial x_{hr}} \frac{\partial}{\partial x_{ls}} \mathrm{Cov}[\mathsf{Y}(\mathbf{x}_h), \mathsf{Y}(\mathbf{x}_l)] \\ &= \tau^2 \frac{\partial}{\partial x_{hr}} \frac{\partial}{\partial x_{ls}} \mathcal{R}(\mathbf{x}_h, \mathbf{x}_l).\end{aligned}$$

Further, the covariance between $\mathsf{Y}(\mathbf{x})$ and its first-order partial derivative $\mathsf{D}^r(\mathbf{x})$ is given by

$$\begin{aligned}\mathrm{Cov}[\mathsf{D}^r(\mathbf{x}_h), \mathsf{Y}(\mathbf{x}_l)] &= \mathrm{Cov}\left[\frac{\partial}{\partial x_{hr}} \mathsf{Y}(\mathbf{x}_h), \mathsf{Y}(\mathbf{x}_l)\right] \\ &= \tau^2 \frac{\partial}{\partial x_{hr}} \mathcal{R}(\mathbf{x}_h, \mathbf{x}_l).\end{aligned}$$

A common choice for the correlation function is the exponential form

$$\mathcal{R}(\mathbf{x}_h, \mathbf{x}_l) = \exp\left\{-\sum_{r=1}^{d} \theta_r |x_{hr} - x_{lr}|^{p_r}\right\}. \tag{8}$$

The parameter θ_r affects how quickly the correlation decreases as two points become farther apart in the direction of the rth coordinate: the greater θ_r is, the less correlation exists in that direction. The parameter p_r describes

how smooth the response surface is. When $p_r = 1$, the correlation function—and therefore the predicted response surface—is continuous but not differentiable; this is clearly not appropriate for modeling a response surface for which gradients exist. When $p_r = 2$, which gives what is known as the "Gaussian correlation function," the correlation function and surface are infinitely continuously differentiable. Although this may be smoother than necessary, there is a long history of successful use of the Gaussian correlation function in practice, so we adopt it here. To see other choices of correlation functions that are differentiable, refer to Stein (1999), Näther and Šimák (2003), Šimák (2002), and Stephenson (2010).

To connect with the discussion in §2.1, the observations made at k design points now include both the response and all the first-order partial derivatives at each design point. We organize them into the $k(d+1) \times 1$ vector as follows:

$$\mathsf{Y}_+ = (\mathsf{Y}^\top, (\mathsf{D}^1)^\top, \ldots, (\mathsf{D}^d)^\top)^\top, \tag{9}$$

where Y is the $k \times 1$ vector of observed responses; and for $r = 1, 2, \ldots, d$

$$\mathsf{D}^r = (\mathsf{D}^r(\mathbf{x}_1), \mathsf{D}^r(\mathbf{x}_2), \ldots, \mathsf{D}^r(\mathbf{x}_k))^\top$$

denotes the vector of observed partial derivatives of Y with respect to the rth design variable. Figure 2 illustrates kriging with gradient information through the one-dimensional two-design-point model. On the response surface, we observe the response $\mathsf{Y}(x_i)$ and its derivative $\mathsf{D}^1(x_i)$ at $x_i \in \Re$, $i = 1, 2$ with no simulation noise present. In this illustration, the observation vector is $\mathsf{Y}_+ = (\mathsf{Y}(x_1), \mathsf{Y}(x_2), \mathsf{D}^1(x_1), \mathsf{D}^1(x_2))^\top$.

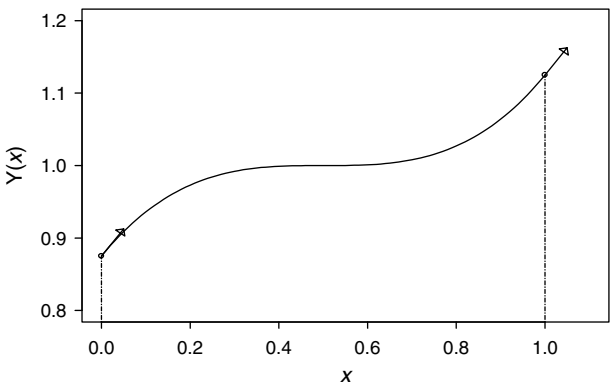
Throughout the paper, we assume that the mean function $\mathbf{f}(\mathbf{x})^\top\boldsymbol{\beta}$ is differentiable. The correlation function is chosen as the exponential form given in Equation (8) with $p_r = 2$ for $r = 1, 2, \ldots, d$. Notice that these are sufficient conditions for the existence of the derivative processes, and they insure that the various variance–covariance matrices we require are positive definite. Now kriging with gradient information can be constructed through augmenting the relevant objects of the metamodel given in §2.1. Let $\boldsymbol{\Sigma}_{\mathsf{M}_+}$ be the $k(d+1) \times k(d+1)$ variance–covariance matrix that includes three types of the spatial covariances at a pair of design points: covariances between the responses, covariances between the derivatives, and covariances between the response at one design point and the derivative at another:

$$\boldsymbol{\Sigma}_{\mathsf{M}_+} = \begin{pmatrix} C^M_{0,0}(1,1) & \cdots & C^M_{0,0}(1,k) & C^M_{0,1}(1,1) & \cdots \\ \vdots & & \vdots & \vdots & \\ C^M_{0,0}(k,1) & \cdots & C^M_{0,0}(k,k) & C^M_{0,1}(k,1) & \cdots \\ \vdots & & \vdots & \vdots & \\ C^M_{d,0}(1,1) & \cdots & C^M_{d,0}(1,k) & C^M_{d,1}(1,1) & \cdots \\ \vdots & & \vdots & \vdots & \\ C^M_{d,0}(k,1) & \cdots & C^M_{d,0}(k,k) & C^M_{d,1}(k,1) & \cdots \end{pmatrix}$$
$$\begin{pmatrix} C^M_{0,1}(1,k) & \cdots & C^M_{0,d}(1,1) & \cdots & C^M_{0,d}(1,k) \\ \vdots & & \vdots & & \\ C^M_{0,1}(k,k) & \cdots & C^M_{0,d}(k,1) & \cdots & C^M_{0,d}(k,k) \\ \vdots & & \vdots & & \\ C^M_{d,1}(1,k) & \cdots & C^M_{d,d}(1,1) & \cdots & C^M_{d,d}(1,k) \\ \vdots & & \vdots & & \\ C^M_{d,1}(k,k) & \cdots & C^M_{d,d}(k,1) & \cdots & C^M_{d,d}(k,k) \end{pmatrix}. \tag{10}$$

In the entries above, $i, h = 1, 2, \ldots, k$ represents the indices for design points and the subscripts $m, g = 1, 2, \ldots, d$ $(m \neq g)$ indicate the coordinate with respect to which we have taken the partial derivative. For notation simplicity, the subscript 0 means that there is no differentiation. To be specific,

$$\begin{aligned} C^M_{0,0}(i,h) &= \mathrm{Cov}[\mathsf{Y}(\mathbf{x}_i), \mathsf{Y}(\mathbf{x}_h)] \\ &= \tau^2 \exp\left\{-\sum_{r=1}^{d} \theta_r (x_{ir} - x_{hr})^2\right\}, \\ C^M_{0,m}(i,h) &= \mathrm{Cov}[\mathsf{Y}(\mathbf{x}_i), \mathsf{D}^m(\mathbf{x}_h)] \\ &= (2\theta_m)(x_{im} - x_{hm}) C^M_{0,0}(i,h), \\ C^M_{m,0}(i,h) &= \mathrm{Cov}[\mathsf{D}^m(\mathbf{x}_i), \mathsf{Y}(\mathbf{x}_h)] = -C^M_{0,m}(i,h), \\ C^M_{m,g}(i,h) &= \mathrm{Cov}[\mathsf{D}^m(\mathbf{x}_i), \mathsf{D}^g(\mathbf{x}_h)] \\ &= (-4\theta_m\theta_g)(x_{im} - x_{hm})(x_{ig} - x_{hg}) C^M_{0,0}(i,h), \\ C^M_{m,m}(i,h) &= \mathrm{Cov}[\mathsf{D}^m(\mathbf{x}_i), \mathsf{D}^m(\mathbf{x}_h)] \\ &= (2\theta_m)[1 - 2\theta_m(x_{im} - x_{hm})^2] C^M_{0,0}(i,h). \end{aligned} \tag{11}$$

Figure 2. A two-design-point model for kriging with gradient information.



Note. Arrows indicate the observed values of the gradients at x_1 and x_2.

Notice that $C^M_{0,m}(i,i) = C^M_{m,0}(i,i) = 0$ for $m = 1, 2, \ldots, d$ at all design points. In particular, for the two-design-point model, Equation (10) reduces to

$$\boldsymbol{\Sigma}_{\mathsf{M}_+} = \tau^2 \begin{pmatrix} 1 & C^M_{0,0}(1,2) & 0 & C^M_{0,1}(1,2) \\ C^M_{0,0}(2,1) & 1 & C^M_{0,1}(2,1) & 0 \\ 0 & C^M_{1,0}(1,2) & 2\theta & C^M_{1,1}(1,2) \\ C^M_{1,0}(2,1) & 0 & C^M_{1,1}(2,1) & 2\theta \end{pmatrix}. \quad (12)$$

The matrix $\mathbf{F}$ in §2.1 now becomes $\mathbf{F}_+$, which is the $k(d+1) \times p$ vector of functions

$$\mathbf{F}_+ = \big(\mathbf{f}(\mathbf{x}_1)^\top, \ldots, \mathbf{f}(\mathbf{x}_k)^\top, (\partial \mathbf{f}(\mathbf{x}_1)/\partial x_1)^\top, \ldots, (\partial \mathbf{f}(\mathbf{x}_k)/\partial x_1)^\top, \ldots, (\partial \mathbf{f}(\mathbf{x}_1)/\partial x_d)^\top, \ldots, (\partial \mathbf{f}(\mathbf{x}_k)/\partial x_d)^\top\big)^\top.$$

For instance, if $\mathbf{f}^\top(x)\boldsymbol{\beta} = \beta_0$ is used in the two-design-point model, then $\mathbf{F}_+ = (1, 1, 0, 0)^\top$.

To do prediction at $\mathbf{x}_0 \in \Re^d$, $\boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}, \cdot)$ in §2.1 has to be updated to $\boldsymbol{\Sigma}_{\mathsf{M}_+}(\mathbf{x}, \cdot)$, the $k(d+1) \times 1$ vector that consists of not only the spatial covariances between the response at $\mathbf{x}_0$ and those at $\mathbf{x}_i, i = 1, 2, \ldots, k$, but the covariances between the response at $\mathbf{x}_0$ and the partial derivatives at the k design points as well. Specifically,

$$\boldsymbol{\Sigma}_{\mathsf{M}_+}(\mathbf{x}_0, \cdot) = \big(C^M_{0,0}(0,1), \ldots, C^M_{0,0}(0,k), C^M_{0,1}(0,1), \ldots, C^M_{0,1}(0,k), \ldots, C^M_{0,d}(0,1), \ldots, C^M_{0,d}(0,k)\big)^\top.$$

References such as Santner et al. (2003) and Šimák (2002) give the following results on kriging with derivative information. When $\boldsymbol{\beta}$ is known, the MSE-optimal predictor and its corresponding MSE can be obtained by substituting $\mathsf{Y}_+, \mathbf{F}_+, \boldsymbol{\Sigma}_{\mathsf{M}_+}, \boldsymbol{\Sigma}_{\mathsf{M}_+}(\mathbf{x}_0, \cdot)$, respectively, for $\mathsf{Y}, \mathbf{F}, \boldsymbol{\Sigma}_{\mathsf{M}}, \boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot)$ into Equations (3) and (4). On the other hand, if $\boldsymbol{\beta}$ has to be estimated, Equations (5) and (6) are used instead. In the next section, we develop a metamodel with gradient estimators that is adapted to stochastic simulation output, which is a central contribution of this paper.

2.3. Stochastic Kriging with Gradient Estimators

To start, we review the idea of stochastic kriging without gradient estimators. Stochastic kriging models the simulation's output on replication j at design point $\mathbf{x}_i \in \Re^d$, $i = 1, 2, \ldots, k$ as

$$\mathscr{Y}_j(\mathbf{x}_i) = \mathsf{Y}(\mathbf{x}_i) + \varepsilon_j(\mathbf{x}_i) = \mathbf{f}(\mathbf{x})^\top\boldsymbol{\beta} + \mathsf{M}(\mathbf{x}_i) + \varepsilon_j(\mathbf{x}_i). \quad (13)$$

What is new in Equation (13) is accounting for stochastic simulation noise. At design point $\mathbf{x}_i$, $n_i \geqslant 2$ simulation replications are obtained; $\varepsilon_1(\mathbf{x}_i), \varepsilon_2(\mathbf{x}_i), \ldots, \varepsilon_{n_i}(\mathbf{x}_i)$ represent the independent and identically distributed mean-zero sampling noise observed for each replication taken at design point $\mathbf{x}_i$.

On the jth simulation replication, suppose that we not only observe the simulation output $\{\mathscr{Y}_j(\mathbf{x}_i)\}_{i=1}^k$ at all k design points, but we are also able to obtain unbiased gradient estimators $\{\mathscr{D}_j^r(\mathbf{x}_i),\ r = 1, 2, \ldots, d\}_{i=1}^k$, where $\mathscr{D}_j^r(\mathbf{x}_i)$ denotes the estimator of the rth gradient component in the jth simulation replication at design point $\mathbf{x}_i$. We propose to use Equation (14) to describe the gradient estimators, which will be explored more fully in §4:

$$\mathscr{D}_j^r(\mathbf{x}_i) = \frac{\partial \mathsf{Y}(\mathbf{x}_i)}{\partial x_r} + \zeta_j^r(\mathbf{x}_i) = \left(\frac{\partial \mathbf{f}(\mathbf{x}_i)}{\partial x_r}\right)^\top \boldsymbol{\beta} + \frac{\partial \mathsf{M}(\mathbf{x}_i)}{\partial x_r} + \zeta_j^r(\mathbf{x}_i), \quad (14)$$

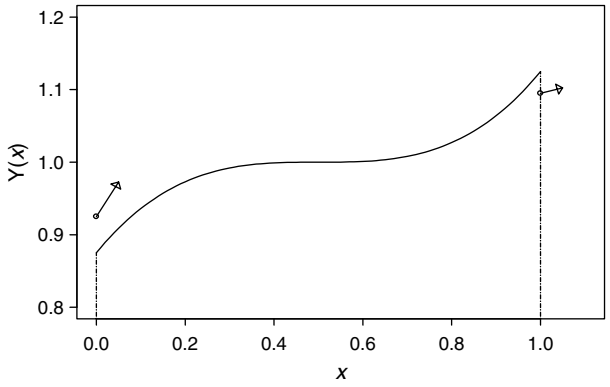
where $\zeta_1^r(\mathbf{x}_i), \zeta_2^r(\mathbf{x}_i), \ldots, \zeta_{n_i}^r(\mathbf{x}_i)$ represent the mean-zero, independent and identically distributed (not necessarily normal) sampling noise in the estimators of the rth gradient component in different simulation replications at design point $\mathbf{x}_i$. Equations (13) and (14) jointly provide the framework of stochastic kriging with gradient estimators.

Figure 3 illustrates the simulation responses and gradient estimators obtained in a particular stochastic simulation replication j for the two-design-point model. Comparing Figure 3 with Figure 2, we see that the observed simulation response $\mathscr{Y}_j(x_i)$ deviates from the true value at design point x_i on the response surface; meanwhile, the gradient estimator $\mathscr{D}_j^1(x_i)$ does not reflect the true trend of the response surface at x_i, $i = 1, 2$ either.

For the theoretical analysis in this paper, the following assumptions are made for stochastic kriging with gradient estimators, in addition to those for stochastic kriging without gradient estimators given in Ankenman et al. (2010):

1. Common random numbers (CRN) are not used across design points. Just as CRN causes the responses across design points to be positively correlated, it would cause the gradient estimators across design points to be positively correlated as well. As shown in Chen et al. (2010, 2012),

Figure 3. A two-design-point model for stochastic kriging with gradient estimators.



Note. Curve is the true response function $\mathsf{Y}(x)$; o's are the observed values at x_1 and x_2; and arrows indicate the observed values of the gradient estimators at x_1 and x_2 in a particular simulation replication.

which studied the effect of CRN on stochastic kriging, CRN inflates the MSE at prediction points; parallel reasoning makes us believe that CRN will have a similar impact on prediction when applied with the gradient-enhanced stochastic kriging. However, to verify this reasoning we conduct experiments with and without CRN in §§5 and 6, and they do confirm our conjecture. It is worth noting that Chen et al. (2012) showed that CRN may be beneficial when stochastic kriging itself is used to predict gradients.

2. The simulation noise ζ associated with the gradient estimators is independent of the random process M and its derivative processes.

Under these assumptions, correlation only exists for $(\varepsilon_j(\mathbf{x}_i), \zeta_j^1(\mathbf{x}_i), \ldots, \zeta_j^d(\mathbf{x}_i))^\top$ and not between components with different replication index j or design point i. At design point $\mathbf{x}_i$, let the variance of the simulation noise in the response be $\mathrm{Var}[\varepsilon_j(\mathbf{x}_i)] = \sigma_{i0}^2$ and the variance of the simulation noise in the estimator of the rth gradient component be $\mathrm{Var}[\zeta_j^r(\mathbf{x}_i)] = \sigma_{ir}^2$, $r = 1, 2, \ldots, d$. Define the correlation between the simulation noise in the response and in the estimator of the rth gradient component as $\varrho_i^{(0,r)} = \mathrm{Corr}[\varepsilon_j(\mathbf{x}_i), \zeta_j^r(\mathbf{x}_i)]$, $r = 1, 2, \ldots, d$. Let the correlation between the simulation noise in the estimators of a pair of distinct gradient components be $\varrho_i^{(r,s)} = \mathrm{Corr}[\zeta_j^r(\mathbf{x}_i), \zeta_j^s(\mathbf{x}_i)]$, $r, s = 1, 2, \ldots, d$ $(r \neq s)$. Notice that the $\varrho_i^{(0,r)}$'s and the $\varrho_i^{(r,s)}$'s at different design points are not necessarily equal. Let $\bar{\mathscr{Y}}_+$ be the $k(d+1) \times 1$ vector that contains all of the sample average simulation responses and gradient estimators:

$$\begin{aligned}\bar{\mathscr{Y}}_+ &= \big(\bar{\mathscr{Y}}(\mathbf{x}_1), \ldots, \bar{\mathscr{Y}}(\mathbf{x}_k), \bar{\mathscr{D}}^1(\mathbf{x}_1), \ldots, \bar{\mathscr{D}}^1(\mathbf{x}_k), \ldots, \\ &\qquad \bar{\mathscr{D}}^d(\mathbf{x}_1), \ldots, \bar{\mathscr{D}}^d(\mathbf{x}_k)\big)^\top \\ &= \mathsf{Y}_+ + \bar{\boldsymbol{\varepsilon}}_+, \end{aligned} \tag{15}$$

where Y_+ is defined in Equation (9) and the $k(d+1) \times 1$ vector of averaged simulation noise

$$\begin{aligned}\bar{\boldsymbol{\varepsilon}}_+ = \big(&\bar{\varepsilon}(\mathbf{x}_1), \ldots, \bar{\varepsilon}(\mathbf{x}_k), \bar{\zeta}^1(\mathbf{x}_1), \ldots, \bar{\zeta}^1(\mathbf{x}_k), \ldots, \\ &\bar{\zeta}^d(\mathbf{x}_1), \ldots, \bar{\zeta}^d(\mathbf{x}_k)\big)^\top\end{aligned}$$

has mean zero and $k(d+1) \times k(d+1)$ variance–covariance matrix $\boldsymbol{\Sigma}_{\varepsilon_+}$, with elements specified as follows: for $r, s = 0, 1, 2, \ldots, d$, $i = 1, 2, \ldots, k$, $\boldsymbol{\Sigma}_{\varepsilon_+}[rk+i, sk+i] = n_i^{-1}\varrho_i^{(r,s)}\sigma_{ir}\sigma_{is}$, and $\boldsymbol{\Sigma}_{\varepsilon_+}[rk+l, sk+h] = 0$ for all $l \neq h, l, h = 1, 2, \ldots, k$. Notice that $\varrho_i^{(r,r)} = 1$ for $r = 0, 1, 2, \ldots, d$. For the one-dimensional two-design-point illustration, using $\varrho_i = \varrho_i^{(0,1)}$ for short, we have

$$\boldsymbol{\Sigma}_{\varepsilon_+} = \begin{pmatrix} \sigma_{10}^2/n_1 & 0 & \varrho_1\sigma_{10}\sigma_{11}/n_1 & 0 \\ 0 & \sigma_{20}^2/n_2 & 0 & \varrho_2\sigma_{20}\sigma_{21}/n_2 \\ \varrho_1\sigma_{10}\sigma_{11}/n_1 & 0 & \sigma_{11}^2/n_1 & 0 \\ 0 & \varrho_2\sigma_{20}\sigma_{21}/n_2 & 0 & \sigma_{21}^2/n_2 \end{pmatrix}. \tag{16}$$

Combining our analysis on Y_+ in §2.2 and the analysis on $\bar{\boldsymbol{\varepsilon}}_+$, we see that $\bar{\mathscr{Y}}_+$ has mean $\mathbf{F}_+\boldsymbol{\beta}$ and variance–covariance matrix $\boldsymbol{\Sigma}_+$, where $\boldsymbol{\Sigma}_+ = \boldsymbol{\Sigma}_{\mathsf{M}_+} + \boldsymbol{\Sigma}_{\varepsilon_+}$; $\mathbf{F}_+$ and $\boldsymbol{\Sigma}_{\mathsf{M}_+}$ are as specified in §2.2. The model for the averaged simulation outputs follows

$$\begin{aligned}&\bar{\mathscr{Y}}(\mathbf{x}_i) = \mathsf{Y}(\mathbf{x}_i) + \bar{\varepsilon}(\mathbf{x}_i), \\ &\bar{\mathscr{D}}^r(\mathbf{x}_i) = \mathsf{D}^r(\mathbf{x}_i) + \bar{\zeta}^r(\mathbf{x}_i), \quad i = 1, 2, \ldots, k; \\ &\qquad\qquad\qquad\qquad r = 1, 2, \ldots, d. \end{aligned} \tag{17}$$

To do prediction at $\mathbf{x}_0 \in \Re^d$, $\boldsymbol{\Sigma}_{\mathsf{M}_+}(\mathbf{x}_0, \cdot)$ remains the same as in §2.2. By following similar derivations as given in Appendix EC.1. of Ankenman et al. (2010), we can obtain the following results: when $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_{\varepsilon_+}$ and the spatial parameters are known, the MSE-optimal predictor and its corresponding MSE by stochastic kriging with gradient estimators are given by simply substituting $\bar{\mathscr{Y}}_+$, $\mathbf{F}_+$, $\boldsymbol{\Sigma}_+$, $\boldsymbol{\Sigma}_{\mathsf{M}_+}(\mathbf{x}_0, \cdot)$, respectively, for Y, $\mathbf{F}$, $\boldsymbol{\Sigma}_{\mathsf{M}}$, $\boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot)$, into Equations (3) and (4). When $\boldsymbol{\beta}$ has to be estimated, results follow a similar proof as given in Appendix A.1 of Chen et al. (2012): the MSE-optimal predictor and its corresponding MSE can be obtained by substituting $\bar{\mathscr{Y}}_+$, $\mathbf{F}_+$, $\boldsymbol{\Sigma}_+$, $\boldsymbol{\Sigma}_{\mathsf{M}_+}(\mathbf{x}_0, \cdot)$, respectively, for $\mathsf{Y}, \mathbf{F}, \boldsymbol{\Sigma}_{\mathsf{M}}, \boldsymbol{\Sigma}_{\mathsf{M}}(\mathbf{x}_0, \cdot)$ into Equations (5) and (6).

In reality, we have to estimate $\boldsymbol{\Sigma}_{\varepsilon_+}$. If we are willing to further assume that on simulation replication j, $j = 1, 2, \ldots, n_i$, the simulation noise vectors $(\varepsilon_j(\mathbf{x}_i), \zeta_j^1(\mathbf{x}_i), \ldots, \zeta_j^d(\mathbf{x}_i))^\top$ are i.i.d. multivariate normally distributed with mean zero and $(d+1) \times (d+1)$ variance–covariance matrix

$$\begin{pmatrix} \sigma_{i0}^2 & \varrho_i^{(0,1)}\sigma_{i0}\sigma_{i1} & \cdots & \varrho_i^{(0,d)}\sigma_{i0}\sigma_{id} \\ \varrho_i^{(1,0)}\sigma_{i1}\sigma_{i0} & \sigma_{i1}^2 & \cdots & \varrho_i^{(1,d)}\sigma_{i1}\sigma_{id} \\ \vdots & \vdots & \ddots & \vdots \\ \varrho_i^{(d,0)}\sigma_{id}\sigma_{i0} & \varrho_i^{(d,1)}\sigma_{id}\sigma_{i1} & \cdots & \sigma_{id}^2 \end{pmatrix}, \quad i = 1, 2, \ldots, k, \tag{18}$$

then we can show, by proof entirely analogous to the proof of Theorem 1 in Ankenman et al. (2010), that the plug-in predictor $\hat{\mathsf{Y}}(\mathbf{x}_0)$ obtained by replacing $\boldsymbol{\Sigma}_{\varepsilon_+}$ with its sample counterpart $\hat{\boldsymbol{\Sigma}}_{\varepsilon_+}$ is unbiased for $\mathsf{Y}(\mathbf{x}_0)$ (additional details, including an expression for the MSE of this estimator, are in the following section). Furthermore, with this normality assumption we can first estimate the intrinsic variance–covariance $\boldsymbol{\Sigma}_{\varepsilon_+}$ and then use $\hat{\boldsymbol{\Sigma}}_{\varepsilon_+}$ in the likelihood expression to estimate $\boldsymbol{\beta}$, τ^2 and $\boldsymbol{\theta}$. The likelihood expression that we numerically maximize is given in EC.1 of the online companion (available as supplemental material at http://dx.doi.org/10.1287/opre.1120.1143), which is how we estimated the metamodel parameters in the experiments described in §§5 and 6.

In the next section, we analyze the effects on prediction of incorporating gradient estimators through simplified

two-design-point and k-design-point models and try to infer desirable properties of the potential gradient estimators.

3. Analysis of Stochastic Kriging Metamodels with Gradient Estimators

Is it beneficial to include gradient estimators into stochastic kriging metamodels when they are available cheaply? And if it is helpful, what properties of the gradient estimators lead to the most positive impact? In this section we work with simplified two-design-point and k-design-point models to gain some insights as to the answers to these questions. We make these models tractable by enforcing simplifying assumptions, including known metamodel parameters in some cases. In our empirical study we make none of these simplifications and we estimate all metamodel parameters.

3.1. A Two-Design-Point Model

To make the analysis tractable, we start with a one-dimensional ($d = 1$) two-design-point stochastic kriging metamodel with gradient estimators and further simplify as follows: The two design points are $x_1 = 0$ and $x_2 = 1$, and the prediction point x_0 is in the design space $[0, 1]$. A constant trend model is used as is common in practice, i.e., $\mathbf{f}(\mathbf{x})^\top\boldsymbol{\beta} = \beta_0$; and the parameters β_0, τ^2, θ are known. Finally only the gradient estimators $\{\bar{\mathcal{D}}_j^1(x_1)\}_{j=1}^{n_1}$ at x_1 are included in building the stochastic kriging metamodel. Recall that n_i is the number of simulation replications used at design point $x_i, i = 1, 2$.

Our goal in this section is to investigate the interplay among the variances of the response and gradient estimator and the correlation between them as it impacts the MSE of prediction at x_0.

Let $s_{11} = \sigma_{11}/(\sqrt{n_1}\tau)$ measure the sampling noise of the gradient estimator relative to the spatial variation of the unknown response surface. Similarly, let $s_{i0} = \sigma_{i0}/(\sqrt{n_i}\tau)$, $i = 1, 2$ measure the noise of the simulation responses at the two design points relative to the spatial variation of the unknown response surface. Let $\varrho = \mathrm{Corr}[\bar{\mathcal{Y}}(x_1), \bar{\mathcal{D}}^1(x_1)]$ be the correlation between the stochastic simulation response and the gradient estimator obtained at design point $x_1 = 0$. Based on the assumptions given in §2.3, it follows that $\varrho = \mathrm{Corr}[\varepsilon_j(x_1), \zeta_j^1(x_1)]$.

As discussed in §2.3, the observation vector $\bar{\mathcal{Y}}_+ = (\bar{\mathcal{Y}}(x_1), \bar{\mathcal{Y}}(x_2), \bar{\mathcal{D}}^1(x_1))^\top$ has mean $[\beta_0, \beta_0, 0]^\top$ and variance–covariance matrix $\boldsymbol{\Sigma}_{\mathcal{Y}_+} = \boldsymbol{\Sigma}_{\mathsf{M}_+} + \boldsymbol{\Sigma}_{\varepsilon_+}$, which takes the form

$$\boldsymbol{\Sigma}_{\mathcal{Y}_+} = \tau^2 \left(\begin{array}{cc|c} 1+s_{10}^2 & e^{-\theta} & \varrho s_{10}s_{11} \\ e^{-\theta} & 1+s_{20}^2 & 2\theta e^{-\theta} \\ \hline \varrho s_{10}s_{11} & 2\theta e^{-\theta} & 2\theta + s_{11}^2 \end{array}\right) = \begin{pmatrix} \boldsymbol{\Sigma}_2 & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix}, \tag{19}$$

where $\boldsymbol{\Sigma}_2$ is the variance–covariance matrix of $(\bar{\mathcal{Y}}(x_1), \bar{\mathcal{Y}}(x_2))^\top$ at the two design points; $\mathbf{C}_{12}$ is a 2×1 vector and $\mathbf{C}_{21} = \mathbf{C}_{12}^\top$; $\mathbf{C}_{22} = \tau^2(2\theta + s_{11}^2)$ in this case is a scalar. Correspondingly, we have the 3×1 vector of spatial covariances

$$\boldsymbol{\Sigma}_{\mathsf{M}_+}(x_0, \cdot) = \begin{pmatrix} \mathrm{Cov}[\mathsf{Y}(x_0), \mathsf{Y}(x_1)] \\ \mathrm{Cov}[\mathsf{Y}(x_0), \mathsf{Y}(x_2)] \\ \mathrm{Cov}[\mathsf{Y}(x_0), \mathsf{D}^1(x_1)] \end{pmatrix} = \tau^2 \begin{pmatrix} e^{-\theta x_0^2} \\ e^{-\theta(x_0-1)^2} \\ 2\theta x_0 e^{-\theta x_0^2} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathsf{M}}(x_0, \cdot) \\ \mathrm{C}_3 \end{pmatrix}.$$

We see that $\boldsymbol{\Sigma}_{\mathsf{M}}(x_0, \cdot)$ is a 2×1 vector that contains the spatial covariances between the true responses at x_0 and those at design points x_1 and x_2. In this case, the scalar $\mathrm{C}_3 = 2\theta\tau^2 x_0 \exp\{-\theta x_0^2\}$ denotes the spatial covariance between the true response at x_0 and the gradient of the response surface at x_1.

We are now ready to study the MSE of the MSE-optimal predictor $\hat{\mathsf{Y}}(x_0)$ at a prediction point x_0. First we give an expression for the MSE when incorporating gradient estimators into a stochastic kriging metamodel (call it MSE_+):

$$\begin{aligned} \mathrm{MSE}_+ &= \boldsymbol{\Sigma}_{\mathsf{M}_+}(x_0, x_0) - \boldsymbol{\Sigma}_{\mathsf{M}_+}(x_0, \cdot)^\top \boldsymbol{\Sigma}_{\mathcal{Y}_+}^{-1} \boldsymbol{\Sigma}_{\mathsf{M}_+}(x_0, \cdot) \\ &= \tau^2 - (\boldsymbol{\Sigma}_{\mathsf{M}}(x_0, \cdot)\ \mathrm{C}_3) \begin{pmatrix} \boldsymbol{\Sigma}_2 & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix}^{-1} \cdot \begin{pmatrix} \boldsymbol{\Sigma}_{\mathsf{M}}(x_0, \cdot) \\ \mathrm{C}_3 \end{pmatrix}. \end{aligned} \tag{20}$$

Let $q = (\mathbf{C}_{22} - \mathbf{C}_{21}\boldsymbol{\Sigma}_2^{-1}\mathbf{C}_{12})^{-1}$. Then Equation (20) can be shown to reduce to

$$\begin{aligned} \mathrm{MSE}_+ = &\left\{\tau^2 - \boldsymbol{\Sigma}_{\mathsf{M}}(x_0, \cdot)^\top \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\Sigma}_{\mathsf{M}}(x_0, \cdot)\right\} \\ &- q\left\{\boldsymbol{\Sigma}_{\mathsf{M}}(x_0, \cdot)^\top \boldsymbol{\Sigma}_2^{-1} \mathbf{C}_{12} - \mathrm{C}_3\right\}^2. \end{aligned} \tag{21}$$

We recognize that $\tau^2 - \boldsymbol{\Sigma}_{\mathsf{M}}(x_0, \cdot)^\top \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\Sigma}_{\mathsf{M}}(x_0, \cdot)$ is the MSE of the MSE-optimal predictor $\hat{\mathsf{Y}}(x_0)$ obtained by stochastic kriging without using gradient estimators when β_0 is known. Hence to see whether incorporating the gradient estimators into a stochastic kriging metamodel reduces the MSE, it is necessary to check whether $q = (\mathbf{C}_{22} - \mathbf{C}_{21}\boldsymbol{\Sigma}_2^{-1}\mathbf{C}_{12})^{-1} > 0$ holds. This is proven in EC.2 of the online companion; hence, we conclude that incorporating gradient estimators into a stochastic kriging metamodel improves prediction performance for any $x_0 \in [x_1, x_2]$, at least for this simplified two-design-point problem. Next we examine the effect of gradient–estimator variability and correlation between the gradient estimator and the response via the parameters s_{11} and ϱ, respectively; the proofs are in Appendices EC.3 and EC.4 (in the online companion).

3.1.1. The Effect of ϱ on MSE. The parameter ϱ represents the correlation between the response and gradient estimator at x_1. The greater $|\varrho|$ is, the greater the reduction

in MSE. However, for the same value of $|\varrho|$, $\varrho < 0$ leads to greater reduction than $\varrho > 0$ in the MSE at those x_0 that are not very close to $x_1 = 0$. This correlation is only under our control through the choice of the gradient estimator. If the gradient estimator is at $x_2 = 1$ instead of at $x_1 = 0$, then the same line of reasoning results in a parallel conclusion: positive correlation between the simulation response and the gradient estimator helps more than negative correlation in reducing the MSE at those x_0 that are not very close to $x_2 = 1$.

To develop an intuitive understanding of the preferred sign for ϱ, we use the following simple argument. Suppose $Y(x_0)$ is predicted through the first-order Taylor series approximation

$$\hat{g}(x_0) = \bar{Y}(x_i) + \bar{\mathcal{D}}^1(x_i)(x_0 - x_i).$$

Because

$$\begin{aligned}\mathrm{Var}[\hat{g}(x_0)] &= \mathrm{Var}[\bar{Y}(x_i)] + (x_0 - x_i)^2\mathrm{Var}[\bar{\mathcal{D}}^1(x_i)] \\ &\quad + 2(x_0 - x_i)\mathrm{Cov}[\bar{Y}(x_i), \bar{\mathcal{D}}^1(x_i)] \\ &= \mathrm{Var}[\bar{Y}(x_i)] + (x_0 - x_i)^2\mathrm{Var}[\bar{\mathcal{D}}^1(x_i)] \\ &\quad + 2(x_0 - x_i)\varrho\sigma_{i0}\sigma_{i1},\end{aligned}$$

it becomes clear that $(x_0 - x_i)\varrho < 0$ reduces $\mathrm{Var}[g(x_0)]$. Hence in the two-design-point model, when the gradient estimator is at x_1, predicting at $x_0 > x_1 = 0$ means that $\varrho < 0$ is preferred at x_1; and it is the other way around if the gradient estimator is at x_2.

3.1.2. The Effect of s_{11} on MSE. The parameter s_{11} represents the sampling noise of the gradient estimator relative to the variation in the response surface itself. The effect of s_{11} on MSE is not as clear-cut as ϱ's effect because it depends on the values of other parameters and the location of x_0. Nevertheless, in EC.4 of the online companion we show that when the correlation between the gradient estimator and the simulation response is very weak, i.e., $|\varrho| \approx 0$, then smaller s_{11} leads to greater reduction in MSE. This makes intuitive sense as smaller s_{11} means a more precise gradient estimator.

3.2. A k-Design-Point Model

In this section we analyze a k-design-point stochastic kriging metamodel with gradient estimators. To make the analysis tractable, we employ the following assumptions: (A) Together with the responses, the estimators for the mth gradient component at all k design points are included in the stochastic kriging metamodel; without loss of generality, we assume $m = 1$. (B) The trend model is $\mathbf{f}(\mathbf{x})^\top\boldsymbol{\beta}$, and all of the parameters $\boldsymbol{\beta}, \tau^2, \{\theta_j\}_{j=1}^d$ are known. And finally, (C) the design points are so widely spread in the design space $\Re^d$ that the spatial correlations of the observations (both the simulation responses and the gradient estimators) at distinct design points are approximately 0; i.e., $\exp\{-\sum_{j=1}^d \theta_j(x_{ij} - x_{hj})^2\} \approx 0$ for $i \neq h$, $i, h = 1, 2, \ldots, k$.

Notice that Assumption A is general in the sense of having gradient estimators at all k design points, but is a simplification in that we have gradient information only in one coordinate direction. Assumption C is motivated by Mitchell et al. (1994), in which progressively weaker correlations are used to find asymptotically optimal experimental designs. In our case this allows us to isolate the impact of incorporating gradient estimators from the complex spatial dependence in a general k-design-point experiment design.

As in §3.1, let $\varrho_i = \mathrm{Corr}[\bar{Y}(\mathbf{x}_i), \bar{\mathcal{D}}^1(\mathbf{x}_i)]$ denote the correlation between the simulation response and the estimator of the first gradient component at the ith design point. Denote the variance of the simulation response at $\mathbf{x}_i$, $i = 1, 2, \ldots, k$ by σ_{i0}^2 and the variance of the estimator of the first gradient component at $\mathbf{x}_i$ by σ_{i1}^2. Define $s_{i1} = \sigma_{i1}/(\sqrt{n_i}\tau)$ and $s_{i0} = \sigma_{i0}/(\sqrt{n_i}\tau)$, $i = 1, 2, \ldots, k$ as in §3.1.

Under these assumptions, the observation vector $\bar{Y}_+$ is the $2k \times 1$ vector of the average simulation responses and the average gradient estimators, specifically,

$$\begin{aligned}\bar{Y}_+ = \big(&\bar{Y}(\mathbf{x}_1), \bar{Y}(\mathbf{x}_2), \ldots, \bar{Y}(\mathbf{x}_k), \bar{\mathcal{D}}^1(\mathbf{x}_1), \\ &\bar{\mathcal{D}}^1(\mathbf{x}_2), \ldots, \bar{\mathcal{D}}^1(\mathbf{x}_k)\big)^\top,\end{aligned}$$

where $\bar{\mathcal{D}}^1(\mathbf{x}_i) = \partial Y(\mathbf{x}_i)/\partial x_{i1} + \bar{\zeta}^1(\mathbf{x}_i)$. The random vector $\bar{Y}_+$ has mean $\mathbf{F}_+\boldsymbol{\beta}$ and variance–covariance matrix $\boldsymbol{\Sigma}_{k+}$. Specifically, we have

$$\mathbf{F}_+ = \big(\mathbf{f}(\mathbf{x}_1)^\top, \ldots, \mathbf{f}(\mathbf{x}_k)^\top, (\partial\mathbf{f}(\mathbf{x}_1)/\partial x_1)^\top, \ldots, (\partial\mathbf{f}(\mathbf{x}_k)/\partial x_1)^\top\big)^\top$$

and

$$\boldsymbol{\Sigma}_{k+} = \begin{pmatrix} \boldsymbol{\Sigma}_k & \boldsymbol{\Sigma}_{01} \\ \boldsymbol{\Sigma}_{10} & \boldsymbol{\Sigma}_{11} \end{pmatrix}. \tag{22}$$

We break $\boldsymbol{\Sigma}_{k+}$ into four $k \times k$ matrices for ease of exposition: $\boldsymbol{\Sigma}_k = \tau^2\,\mathrm{diag}\{1 + s_{i0}^2\}_{i=1}^k$, $\boldsymbol{\Sigma}_{01} = \boldsymbol{\Sigma}_{10} = \tau^2\,\mathrm{diag}\{\varrho_i s_{i0} s_{i1}\}_{i=1}^k$ and $\boldsymbol{\Sigma}_{11} = \tau^2\,\mathrm{diag}\{s_{i1}^2 + 2\theta_1\}_{i=1}^k$.

Given any prediction point $\mathbf{x}_0 \in \Re^d$, the $2k \times 1$ vector of spatial covariances $\boldsymbol{\Sigma}_{\mathsf{M}_0}(\mathbf{x}_0, \cdot)$ consists of the covariances between the true response at $\mathbf{x}_0$ and the responses at $\{\mathbf{x}_i\}_{i=1}^k$ as well as the covariances between the response at $\mathbf{x}_0$ and the first gradient components at $\{\mathbf{x}_i\}_{i=1}^k$. Specifically, $\boldsymbol{\Sigma}_{\mathsf{M}_+}(\mathbf{x}_0, \cdot) = (\boldsymbol{\Sigma}_{\mathsf{M}_0}(\mathbf{x}_0, \cdot)^\top, \boldsymbol{\Sigma}_{\mathsf{M}_1}(\mathbf{x}_0, \cdot)^\top)^\top$, where

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathsf{M}_0}(\mathbf{x}_0, \cdot) = \tau^2\big(&e^{-\sum_{j=1}^d \theta_j(x_{0j} - x_{1j})^2}, e^{-\sum_{j=1}^d \theta_j(x_{0j} - x_{2j})^2}, \ldots, \\ &e^{-\sum_{j=1}^d \theta_j(x_{0j} - x_{kj})^2}\big)^\top \quad \text{and}\end{aligned}$$

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathsf{M}_1}(\mathbf{x}_0, \cdot) = \tau^2(2\theta_1)\big(&(x_{01} - x_{11})e^{-\sum_{j=1}^d \theta_j(x_{0j} - x_{1j})^2}, \ldots, \\ &(x_{01} - x_{k1})e^{-\sum_{j=1}^d \theta_j(x_{0j} - x_{kj})^2}\big)^\top.\end{aligned}$$

Having established the notation, we are ready to give the MSE of the MSE-optimal predictor $\hat{Y}(x_0)$ at $\mathbf{x}_0$ for this simplified k-design-point model. Let $\mathbf{Q} = (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{10}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_{01})^{-1}$. Then

$$\begin{aligned}\mathrm{MSE}_+ &= \boldsymbol{\Sigma}_{\mathsf{M}_+}(\mathbf{x}_0, \mathbf{x}_0) - \boldsymbol{\Sigma}_{\mathsf{M}_+}(\mathbf{x}_0, \cdot)^\top\boldsymbol{\Sigma}_{k+}^{-1}\boldsymbol{\Sigma}_{\mathsf{M}_+}(\mathbf{x}_0, \cdot) \\ &= \big\{\tau^2 - \boldsymbol{\Sigma}_{\mathsf{M}_0}(\mathbf{x}_0, \cdot)^\top\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_{\mathsf{M}_0}(\mathbf{x}_0, \cdot)\big\} - \xi^\top\mathbf{Q}\xi, \end{aligned} \tag{23}$$

where $\xi = \mathbf{\Sigma}_{10}\mathbf{\Sigma}_k^{-1}\mathbf{\Sigma}_{\mathsf{M}_0}(\mathbf{x}_0,\cdot) - \mathbf{\Sigma}_{\mathsf{M}_1}(\mathbf{x}_0,\cdot)$. We show in EC.5 of the online companion that under the assumptions for this simplified k-design-point model, $\mathbf{Q}$ is positive definite. Therefore, we conclude that the MSE is reduced if gradient estimators are incorporated into stochastic kriging for this k-design-point model.

3.2.1. The Effects of ϱ_i, s_{i0}, and s_{i1} on MSE at $\mathbf{x}_0$. For the sake of brevity, we summarize the effect of ϱ_i on MSE and compare its role with s_{i1}'s when $s_{l0} \gg s_{l1}$ for this k-design-point model below and refer the reader to Appendices EC.6 and EC.7 for the effects of s_{i0} and s_{i1} and the proofs.

1. When one predicts at some design point $\mathbf{x}_i$, we conclude that the sign of ϱ_i no longer matters; the greater $|\varrho_i|$ is, the smaller MSE becomes.

2. When one predicts at some point $\mathbf{x}_0$ other than the k design points, a conclusion similar to the one given in §3.1 regarding the effect of ϱ_i can be drawn: given that the gradient estimators in the direction of the first coordinate are included into stochastic kriging, if $x_{01} > x_{i1}$ at design point $\mathbf{x}_i$, we prefer the correlation $\varrho_i = \mathrm{Corr}[\bar{\mathscr{Y}}(x_i), \bar{\mathscr{D}}^1(x_i)]$ to be negative. On the other hand, if $x_{01} < x_{i1}$, a positive ϱ_i is preferred. In both favorable cases, the greater $|\varrho_i|$ is, the smaller MSE is. An intuitive understanding of the preferred sign of ϱ_i can be given following a similar argument given as in §3.1.

3. When the simulation response is much noisier than the gradient estimator at a design point, i.e., $s_{l0} \gg s_{l1}$ at design point $\mathbf{x}_l$, we show in EC.7 of the online companion that the variability of the gradient estimator plays a more important role in determining the reduction amount in MSE at $\mathbf{x}_0$. Specifically, a gradient estimator with a small variance and a low correlation is more effective in reducing the MSE than another one with a larger variance and a higher correlation with the simulation response at a design point. As we show below, we are more likely to encounter this situation when we use IPA.

3.2.2. The Effects of ϱ_i, s_{i0}, and s_{i1} on IMSE. In §3.2.1, we studied the effect of ϱ_i on MSE at any given prediction point $\mathbf{x}_0 \in \Re^d$ and found that the effect depends on the location of the prediction point to some extent. In this section we investigate the effect over a prediction region denoted by $W \subseteq \Re^d$ following an approach similar to Šimák (2002). Specifically, we use a more comprehensive performance measure, the integrated mean squared error over the region W:

$$\begin{aligned}\mathrm{IMSE}_W &= \int_W \mathrm{MSE}(\mathbf{x})\,d\mathbf{x}\\ &= \int_W \left(\mathbf{\Sigma}_{\mathsf{M}_+}(\mathbf{x},\mathbf{x}) - \mathbf{\Sigma}_{\mathsf{M}_+}(\mathbf{x},\cdot)^\top \mathbf{\Sigma}_{k+}^{-1}\mathbf{\Sigma}_{\mathsf{M}_+}(\mathbf{x},\cdot)\right) d\mathbf{x}\\ &= \int_W \left(\tau^2 - \mathbf{\Sigma}_{\mathsf{M}_+}(\mathbf{x},\cdot)^\top \mathbf{\Sigma}_{k+}^{-1}\mathbf{\Sigma}_{\mathsf{M}_+}(\mathbf{x},\cdot)\right) d\mathbf{x}, \qquad (24)\end{aligned}$$

where $d\mathbf{x} = dx_1,\ldots,dx_d$ represents a volume element of $\Re^d$. We see from Equation (24) that studying IMSE as a function of ϱ_i, s_{i0}, and s_{i1} only requires studying the reduction

$$\Delta_{\mathrm{IMSE}}(W) = \int_W \left(\mathbf{\Sigma}_{\mathsf{M}_+}(\mathbf{x},\cdot)^\top \mathbf{\Sigma}_{k+}^{-1}\mathbf{\Sigma}_{\mathsf{M}_+}(\mathbf{x},\cdot)\right) d\mathbf{x}. \qquad (25)$$

Assuming that the prediction region W is sufficiently large with respect to the practical range of the Gaussian correlation function, we can approximate Equation (25) as an integral over $\Re^d$ and write $\Delta_{\mathrm{IMSE}}(W)$ as

$$\Delta_{\mathrm{IMSE}} = \int_{\Re^d} (\mathbf{\Sigma}_{\mathsf{M}_+}(\mathbf{x},\cdot)^\top \mathbf{\Sigma}_{k+}^{-1}\mathbf{\Sigma}_{\mathsf{M}_+}(\mathbf{x},\cdot))\,d\mathbf{x} \qquad (26)$$

given that the integral exists. We give an expression for Δ_{IMSE} below and leave the derivations to EC.8 of the online companion:

$$\begin{aligned}\Delta_{\mathrm{IMSE}} = \tau^2\left(\frac{\pi}{2}\right)^{d/2}\prod_{j=1}^{d}\theta_j^{-1/2}\\ \cdot\sum_{i=1}^{k}\frac{2(s_{i1}^2+2\theta_1)+\theta_1(1+s_{i0}^2)}{2[(s_{i1}^2+2\theta_1)(1+s_{i0}^2)-(\varrho_i s_{i0}s_{i1})^2]}. \qquad (27)\end{aligned}$$

The following remarks can be made regarding the effects of ϱ_i, s_{i0}^2, and s_{i1}^2 on IMSE of prediction over $\Re^d$.

The effect of ϱ_i. It is obvious from Equation (27) that the larger $|\varrho_i|$ is, the larger Δ_{IMSE} is and hence the smaller the IMSE.

The effect of s_{i0}^2. Some algebraic manipulation gives

$$\begin{aligned}\frac{\partial\Delta_{\mathrm{IMSE}}}{\partial s_{i0}^2} = \tau^2\left(\frac{\pi}{2}\right)^{d/2}\prod_{j=1}^{d}\theta_j^{-1/2}\\ \cdot\sum_{i=1}^{k}\frac{2s_{i1}^4(-1+\varrho_i^2)+s_{i1}^2\theta_1(-8+5\varrho_i^2)-8\theta_1^2}{4[(s_{i1}^2+2\theta_1)(1+s_{i0}^2)-(\varrho_i s_{i0}s_{i1})^2]^2}.\end{aligned}$$

Because $|\varrho_i| \leqslant 1$, we have $\partial\Delta_{\mathrm{IMSE}}/\partial s_{i0}^2 < 0$, $i = 1,2,\ldots,k$. Therefore, the greater s_{i0}^2 is, the smaller Δ_{IMSE} is and hence the larger IMSE. In words, more variable simulation responses at the design points result in larger IMSE.

The effect of s_{i1}^2. The expression for $\partial\Delta_{\mathrm{IMSE}}/\partial s_{i1}^2$ can be rewritten as

$$\begin{aligned}\frac{\partial\Delta_{\mathrm{IMSE}}}{\partial s_{i1}^2} = \tau^2\left(\frac{\pi}{2}\right)^{d/2}\prod_{j=1}^{d}\theta_j^{-1/2}\\ \cdot\sum_{i=1}^{k}\theta_1\frac{s_{i0}^4(-1+\varrho_i^2)+(-1+(-2+5\varrho_i^2)s_{i0}^2)}{4[(s_{i1}^2+2\theta_1)(1+s_{i0}^2)-(\varrho_i s_{i0}s_{i1})^2]^2}.\end{aligned}$$

When $\varrho_i^2 \leqslant 2/5$, i.e., $-2+5\varrho_i^2 \leqslant 0$, then $\partial\Delta_{\mathrm{IMSE}}/\partial s_{i1}^2 < 0$. In this case, Δ_{IMSE} decreases (or equivalently, IMSE increases) as s_{i1}^2 increases. However, when $\varrho_i^2 > 2/5$, as long as $s_{i0}^2 < (5\varrho_i^2-2)^{-1}$ so that $(-2+5\varrho_i^2)s_{i0}^2 < 1$, then Δ_{IMSE} decreases (or equivalently, IMSE increases) as s_{i1}^2 increases. In particular, the upper bound $(5\varrho_i^2-2)^{-1}$ on s_{i0}^2 for this result to hold is a decreasing function of ϱ_i^2. It follows that $\varrho_i^2 = 1$ gives the tightest upper bound $1/3$ on s_{i0}^2. In a stochastic simulation experiment, $s_{i0}^2 = n^{-1}\sigma_{i0}^2/\tau^2 < 1/3$ will hold when a sufficient number of simulation replications n_i is applied at design point $\mathbf{x}_i$.

We conclude that when s_{i0}^2 is sufficiently small, Δ_{IMSE} decreases (or equivalently, IMSE increases) as s_{i1}^2 increases. Hence, the noisier the gradient estimators, the larger the IMSE. Last but not the least, it is also shown at the end of EC.8 of the online companion that when $s_{i0} \gg s_{i1}$, the variability of the gradient estimator plays a more important role in determining the reduction amount in Δ_{IMSE} than the correlation does.

The insights gained in this section on IMSE over $\Re^d$ are relatively simple to summarize. The greater correlation $|\varrho_i|$ between simulation response and gradient estimator and the less variable the gradient estimators and the responses are at all the design points, the better prediction performance is achieved over a relatively large prediction region for this vanishing intersite correlation k-design-point model. When there exist competing gradient estimators to do prediction, the ones with smaller variance are preferred.

4. IPA and LR/SF Gradient Estimators in Stochastic Kriging Metamodels

In this section, we review some stochastic gradient estimation techniques, focusing on the infinitesimal perturbation analysis and the likelihood ratio/score function methods, but briefly discussing three other gradient estimation methods. Rather than provide a comprehensive treatment of this topic, we give some key features of the first two gradient estimation techniques that benefit our discussion that follows. Refer to L'Ecuyer (1990), Fu (2006), and Rubinstein and Shapiro (1993, Chapter 2) for details.

Suppose that a stochastic simulation model is parameterized by a vector of design variables $\mathbf{x} \in \mathscr{X}$, where $\mathscr{X}$ is some open subset in $\Re^d$. We are interested in estimating the gradient of some real-valued (differentiable) function $\alpha(\mathbf{x})$, $\nabla\alpha(\mathbf{x})$, where $\alpha(\mathbf{x}) = \mathrm{E}[\mathscr{Y}(\mathbf{x})]$ with respect to some probability measure $\mathrm{P}_{\mathbf{x}}$ over some measurable space $(\Omega, \mathscr{G})$. Recall that $\mathscr{Y}(\mathbf{x})$ is the stochastic simulation response at point $\mathbf{x}$. To emphasize its dependence on a given "sample path" ω, we note that $\mathscr{Y}(\mathbf{x}) = h(\mathbf{x}, \omega)$, where $h(\cdot, \cdot)$ is some function of the vector $\mathbf{x}$ and the sample path $\omega \in \Omega$ and $h(\mathbf{x}, \cdot)$ is $\mathscr{G}$-measurable.

Suppose that there is a probability measure G on the same measurable space that is independent of $\mathbf{x}$, and for $\mathbf{x} \in \mathscr{X}$, $\mathrm{P}_{\mathbf{x}}$ is absolutely continuous with respect to G (for any set $A \in \mathscr{G}$, $G(A) = 0 \Rightarrow \mathrm{P}_{\mathbf{x}}(A) = 0$). In this case,

$$\alpha(\mathbf{x}) = \mathrm{E}[\mathscr{Y}(\mathbf{x})] = \int_{\Omega} h(\mathbf{x}, \omega) W(G, \mathbf{x}, \omega)\, dG(\omega),$$

where $W(G, \mathbf{x}, \omega) = (d\mathrm{P}_{\mathbf{x}}/dG)(\omega)$ is the Radon-Nikodym derivative of $\mathrm{P}_{\mathbf{x}}$ with respect to G.

Let

$$\begin{aligned}\psi_r(\mathbf{x}, \omega) &= \frac{\partial[h(\mathbf{x}, \omega) W(G, \mathbf{x}, \omega)]}{\partial x_r} \\ &= W(G, \mathbf{x}, \omega) \frac{\partial h(\mathbf{x}, \omega)}{\partial x_r} \\ &\quad + h(\mathbf{x}, \omega) \frac{\partial W(G, \mathbf{x}, \omega)}{\partial x_r} \end{aligned} \tag{28}$$

and $\psi(\mathbf{x}, \omega) = (\psi_1(\mathbf{x}, \omega), \psi_2(\mathbf{x}, \omega), \ldots, \psi_d(\mathbf{x}, \omega))^{\top}$. Then under some regularity conditions such that the interchange of differentiation and expectation is permitted, it follows that the gradient evaluated at point $\mathbf{x}$ is $\nabla\alpha(\mathbf{x}) = \int_{\Omega} \psi(\mathbf{x}, \omega)\, dG(\omega)$, where for $r = 1, 2, \ldots, d$,

$$\frac{\partial \alpha(\mathbf{x})}{\partial x_r} = \int_{\Omega} \psi_r(\mathbf{x}, \omega)\, dG(\omega).$$

Therefore, if we sample ω from G, then $\psi_r(\mathbf{x}, \omega)$ is an unbiased estimator of $\partial\alpha(\mathbf{x})/\partial x_r$.

When we want to estimate $\partial\alpha(\mathbf{x}^*)/\partial x_r$ at some point $\mathbf{x}^*$, a convenient choice is to let $G = \mathrm{P}_{\mathbf{x}^*}$; if it happens that $h(\mathbf{x}, \omega)$ is independent of $\mathbf{x}$, then the estimator of $\partial\alpha(\mathbf{x})/\partial x_r$ reduces to

$$\psi_r(\mathbf{x}^*, \omega) = h(\mathbf{x}^*, \omega) S_r(\mathbf{x}^*, \omega), \tag{29}$$

where $S_r(\mathbf{x}^*, \omega) = [\partial W(\mathrm{P}_{\mathbf{x}^*}, \mathbf{v}, \omega)/\partial x_r]|_{\mathbf{v}=\mathbf{x}^*}$. Equation (29) is one choice for $\mathscr{D}^r(\mathbf{x})$, namely, $\mathscr{D}^r_{\text{LR}}(\mathbf{x}) = h(\mathbf{x}, \omega) S_r(\mathbf{x}, \omega)$. Let $S(\mathbf{x}, \omega) = (S_1(\mathbf{x}, \omega), \ldots, S_d(\mathbf{x}, \omega))^{\top}$; then the gradient estimator is $\psi(\mathbf{x}, \omega) = h(\mathbf{x}, \omega) S(\mathbf{x}, \omega)$, which is familiar as the pure form of the likelihood ratio (LR) gradient estimator.

In stochastic simulation, it is sometimes convenient to consider ω as a sequence of independent $\mathscr{U}(0, 1)$ random variables. In this case, $\mathrm{P}_{\mathbf{x}}$ is independent of $\mathbf{x}$, so that $W(G, \mathbf{x}, \omega) = 1$; and under appropriate conditions, Equation (28) reduces to the infinitesimal perturbation analysis estimator of $\partial\alpha(\mathbf{x})/\partial x_r$, $r = 1, 2, \ldots, d$, specifically

$$\psi_r(\mathbf{x}, \omega) = \frac{\partial h(\mathbf{x}, \omega)}{\partial x_r}, \tag{30}$$

provided that $h(\mathbf{x}, \omega)$ exists for almost all ω. Notice that Equation (30) offers a second candidate choice for $\mathscr{D}^r(\mathbf{x})$, $\mathscr{D}^r_{\text{IPA}}(\mathbf{x}) = \partial h(\mathbf{x}, \omega)/\partial x_r$; the vector of such estimators for partial derivatives forms an IPA gradient estimator.

As long as the condition for Equation (28) to hold is present, both IPA and LR/SF approaches give unbiased gradient estimators. One advantage of the IPA gradient estimator is that its variance typically does not increase with the simulation run length as discussed in L'Ecuyer (1990), which often makes its variance much smaller than the variance of the LR/SF gradient estimator. On the other hand, deriving $\nabla_{\mathbf{x}} h(\mathbf{x}, \omega)$ can sometimes be quite difficult for IPA; by comparison, the LR/SF gradient estimator seems to apply more widely and easily. Fortunately, the IPA gradient estimator has been derived for several large classes of problems. In both cases implementing the estimator typically involves little more than accumulating information already being generated by the simulation and performing some simple calculations.

Another approach is finite-difference gradient estimation (FD). FD estimation is not computationally efficient and the FD estimator may be quite biased (if the finite difference used is large) or variable (if the finite difference is

small). If we have k design points, then to do FD gradient estimation at all of them requires at least kd additional simulation runs. Because spatial correlation metamodeling leverages design points that are closest in space to the prediction point, we would almost certainly be better off filling the space more fully with kd additional design points as opposed to estimating gradients at only k design points. The same is likely true of weak-derivative gradient estimation (also called measured-value differentiation; see Heidergott et al. 2010); although it provides unbiased gradient estimators, the computational effort grows with the dimension of $\mathbf{x}$. And simultaneous perturbation gradient estimation (Spall 2003), like finite differences, introduces bias. However, our general framework can accommodate any of these alternative gradient estimators if used.

In experiments on two examples taken from L'Ecuyer (1990), a k-out-of-N reliability system with parameter-dependent component-lifetime densities and an M/M/1 queue with parameter-dependent service time, we made the following observations:

- The estimated correlation between the gradient estimator and the response at a given design point is typically larger in absolute value when using IPA as compared to LR/SF.
- The variances of the gradient estimators are typically much smaller for IPA than LR/SF. In fact, the responses are much noisier than the IPA gradient estimators; whereas for LR/SF, it is the other way around.
- The prediction performance is better when IPA gradient estimation is applied with stochastic kriging.

In §§4.1 and 4.2, through analysis of IPA and LR/SF gradient estimators that assume a particular form, we link the properties of the IPA and LR/SF gradient estimators to those studied with the simplified stochastic kriging metamodels described in §3 and the observations mentioned above. This analysis is useful for situations in which both IPA and LR/SF gradient estimators are available, such as the example given in §5. We focus on analysis of a one-dimensional ($d = 1$) design space, but the results can easily be generalized to higher dimensions.

4.1. A Closer Look at IPA Gradient Estimators

Let h be a function of random variables $Z_l(x)$, $l = 1, 2, \ldots, s$. We use $h(\mathbf{Z}(x))$ as a short form for $h(Z_1(x), Z_2(x), \ldots, Z_s(x))$. In our context, $h(\mathbf{Z}(x))$ is the sample path performance for a discrete-event stochastic system over a finite horizon and the $Z_l(x)$'s are the inputs to the system. For the "sample path" ω, $\mathcal{Y}(x) = h(\mathbf{Z}(x))$ at the design point x. Notice that to ease notation ω is omitted. Assuming that each of the $Z_l(x)$'s is almost surely differentiable with respect to x, then under some mild conditions $h(\mathbf{Z}(x))$ is also almost surely differentiable with respect to x. Therefore,

$$\mathcal{D}^1_{\mathrm{IPA}}(x) = \frac{dh(\mathbf{Z}(x))}{dx} = \sum_{l=1}^{s} \frac{\partial h}{\partial Z_l}\frac{dZ_l}{dx} \qquad \text{a.s.} \tag{31}$$

We next give a closed-form expression for the covariance between the IPA gradient estimator and the observed response at a design point $x \in \mathcal{X} = [a, b]$ (leaving the proof for EC.9 of the online companion):

$$\begin{aligned}
&\mathrm{Cov}[\mathcal{Y}(x), \mathcal{D}^1_{\mathrm{IPA}}(x)] \\
&\quad = \mathrm{E}[\mathcal{Y}(x)\mathcal{D}^1_{\mathrm{IPA}}(x)] - \mathrm{E}[\mathcal{Y}(x)]\mathrm{E}[\mathcal{D}^1_{\mathrm{IPA}}(x)] \\
&\quad = \mathrm{E}\left[h(\mathbf{Z}(x)) \cdot \frac{dh(\mathbf{Z}(x))}{dx}\right] - \alpha(x)\alpha'(x) \\
&\quad = \frac{1}{2}\frac{d}{dx}(\mathrm{Var}[\mathcal{Y}(x)]).
\end{aligned} \tag{32}$$

It follows that

$$\begin{aligned}
&\mathrm{Corr}[\mathcal{Y}(x), \mathcal{D}^1_{\mathrm{IPA}}(x)] \\
&\quad = \frac{1}{2} \cdot \frac{d}{dx} \ln(\mathrm{Var}[\mathcal{Y}(x)]) \cdot R^{-1}_{\mathrm{IPA}}(x),
\end{aligned} \tag{33}$$

where $R_{\mathrm{IPA}}(x) = (\mathrm{Var}[\mathcal{D}^1_{\mathrm{IPA}}(x)]/\mathrm{Var}[\mathcal{Y}(x)])^{1/2}$ is defined as the square root of the ratio of the variance of the gradient estimator to the variance of the stochastic simulation response at design point x.

Given that the aforementioned conditions for the particular form of sample path gradient estimator are satisfied, and also that the interchange of differentiation and expectation is permissible, we see that the larger $R_{\mathrm{IPA}}(x)$ is, the closer $|\mathrm{Corr}[\mathcal{Y}(x), \mathcal{D}^1_{\mathrm{IPA}}(x)]|$ is to zero. And the sign of this correlation depends only on the derivative of $\ln(\mathrm{Var}[\mathcal{Y}(x)])$ with respect to x. Typically, $R_{\mathrm{IPA}}(x)$ for an IPA gradient estimator is relatively small, because the gradient estimator is much less noisy than its corresponding simulation response at a given design point x. Hence one should expect a strong correlation between the gradient estimator and the simulation response obtained at the same design point.

4.2. A Closer Look at LR/SF Gradient Estimators

Using the notation established in previous sections, suppose that $Z_l(x)$, $l = 1, 2, \ldots, s$ are independent inputs to the system whose distributions depend on x with probability density function $f_l(\cdot)$. If $\mathcal{Y}(x) = h(\mathbf{Z}(x), \omega)$ is independent of x given $\mathbf{Z}(x)$, then the LR/SF estimator of $d\alpha(\mathbf{x})/dx$ takes the simple form

$$\mathcal{D}^1_{\mathrm{LR}}(x) = \mathcal{Y}(x) \cdot S(x, \omega),$$

where $S(x, \omega) = (d/dx)\sum_{l=1}^{s} \ln f_l(Z_l(x))$.

Letting $\alpha'(x) = d\alpha(x)/dx$ for short, we have $\alpha(x) = \mathrm{E}[\mathcal{Y}(x)]$ and $\alpha'(x) = \mathrm{E}[\mathcal{D}^1_{\mathrm{LR}}(x)]$. The covariance of the gradient estimator and the response at x can be expressed as

$$\begin{aligned}
\mathrm{Cov}[\mathcal{Y}(x), \mathcal{D}^1_{\mathrm{LR}}(x)] &= \mathrm{E}[\mathcal{Y}(x)\mathcal{D}^1_{\mathrm{LR}}(x)] - \alpha(x)\alpha'(x) \\
&= \mathrm{E}[(\mathcal{Y}(x))^2 \cdot S(x, \omega)] - \alpha(x)\alpha'(x) \\
&= \frac{d}{dx}\mathrm{Var}[\mathcal{Y}(x)] + \alpha(x)\alpha'(x).
\end{aligned} \tag{34}$$

The correlation of the gradient estimator and the response at x is

$$\begin{aligned}\mathrm{Corr}[\mathcal{Y}(x), \mathcal{D}^1_{\mathrm{LR}}(x)] &= \frac{\mathrm{Cov}[\mathcal{Y}(x), \mathcal{D}^1_{\mathrm{LR}}(x)]}{\sqrt{\mathrm{Var}[\mathcal{Y}(x)]}\sqrt{\mathrm{Var}[\mathcal{D}^1_{\mathrm{LR}}(x)]}} \\ &= \frac{d}{dx}\ln(\mathrm{Var}[\mathcal{Y}(x)]) \cdot R^{-1}_{\mathrm{LR}}(x) \\ &\quad + \frac{\alpha(x)\alpha'(x)}{\sqrt{\mathrm{Var}[\mathcal{Y}(x)]}\sqrt{\mathrm{Var}[\mathcal{D}^1_{\mathrm{LR}}(x)]}}, \quad (35)\end{aligned}$$

where $R_{\mathrm{LR}}(x)$ is defined analogously to $R_{\mathrm{IPA}}(x)$.

It is interesting to compare Equations (32) and (34) for IPA and LR/SF gradient estimators. At a given design point x, it follows that $R_{\mathrm{LR}}(x)$ is much larger than $R_{\mathrm{IPA}}(x)$, because $\mathrm{Var}[\mathcal{D}^1_{\mathrm{LR}}(x)]$ is typically large compared to $\mathrm{Var}[\mathcal{D}^1_{\mathrm{IPA}}(x)]$. The consequence is that the correlation between the gradient estimator and the simulation response at a given design point is often smaller in magnitude for a LR/SF gradient estimator.

We conclude this section with some observations that are based on the analysis in §§3.1–3.2, which reveals what affects the MSE of prediction for stochastic kriging enhanced with gradient estimators—the comparison of gradient estimators in §§4.1–4.2—which relates the analysis to IPA and LR gradient estimators and the experiments we have conducted, some of which are described below. Typically, the correlation between the gradient estimator and the simulation response at a given design point is larger in magnitude for an IPA gradient estimator, and the variance of an IPA gradient estimator is often smaller compared to its LR/SF counterpart. There do exist exceptions, however, in which higher correlation is obtained for an LR/SF gradient estimator (see the Black-Scholes call option pricing model in §5). Nevertheless, the variance of a gradient estimator plays the dominant role in prediction performance. Therefore, if both types of gradient estimators are available, the IPA gradient estimation approach is recommended for stochastic kriging.

5. A Tractable Example: The Black-Scholes Call Option Pricing Model

Through an experiment on pricing a call option, we demonstrate the ability of stochastic kriging with gradient estimators to deliver better prediction results "on demand." In this experiment all parameters of the metamodel are unknown and we estimate them from simulation output data. However, we know the true response surface and exploit it in our evaluation.

5.1. The Model: Dynamics and Simulation

The Black-Scholes model is described by the following stochastic differential equation (SDE): $dS_t = rS_t\,dt + \sigma S_t\,dW_t$, where r is the risk-free rate and σ is the volatility of the stock price S_t. Usually, r is obtained from certain benchmark interest rates such as treasury bond rates or LIBOR and σ can be calculated from historical data. This SDE admits a closed-form solution that is

$$S_t = S_0 \exp\left\{\left(\frac{r-\sigma^2}{2}\right)t + \sigma W_t\right\},$$

where $W_t \sim \mathcal{N}(0, t)$.

The European call option is a right to buy a stock at the prespecified date, called the option maturity T, at the prespecified price called the option strike K. The value of this option is the net present value calculated under the model above and it is given by

$$C(T, K; S_0, r, \sigma) = \mathrm{E}[P] = \mathrm{E}[e^{-rT}(S_T - K)^+]. \quad (36)$$

We can obtain price estimates for different (T, K) pairs by the Monte Carlo method. The sensitivities of interest are the partial derivatives of C with respect to the parameters (S_0, r, σ). In this experiment we focus on the sensitivity regarding the underlying stock price S_0 with all other parameters fixed. The IPA and LR gradient estimators are, respectively,

$$\frac{dP}{dS_0} = e^{-rT}\left(\frac{S_T}{S_0}\right) \cdot \mathbf{1}\{S_T \geqslant K\}$$

$$\frac{dP}{dS_0} = e^{-rT}(S_T - K)^+ d(S_T)(S_0\sigma\sqrt{T})^{-1}$$

where

$$d(y) = \left(\sigma\sqrt{t}\right)^{-1}\left(\ln\left(\frac{y}{S_0}\right) - \left(r - \frac{\sigma^2}{2}\right)t\right).$$

Refer to Glasserman (2004, Chapter 7) for details.

5.2. Experimental Design and Results

We compare the prediction performance of stochastic kriging with and without gradient estimators. The example is a one-dimensional European call option model. We treat $C(T, K; S_0, r, \sigma)$ as an unknown response of the design variable S_0 while holding (T, K, r, σ) fixed. The parameter configuration is shown in Table 1.

The experiment design is as follows. An equally spaced grid design of k design points $S_0 \in [80, 120]$ is used with $k \in \{7, 13, 25\}$; notice that the two end points 80 and 120 are always included. To assess the impact of the stochastic simulation noise, we use an equal number of simulation replications n at each design point and vary $n \in$

Table 1. Parameters for the Black-Scholes European call option model.

S_0	K	T	r	σ
[80, 120]	100	1 year	3%	40%

$\{500, 2{,}000, 8{,}000\}$. Pretending that we have little information about the true response surface, a constant trend model $\mathbf{f}(S_0)^\top \boldsymbol{\beta} = \beta_0$ is chosen for stochastic kriging. We consider three types of stochastic kriging metamodels, stochastic kriging without gradient estimators (SK), stochastic kriging with IPA gradient estimators (SK + IPA), and stochastic kriging with LR/SF gradient estimators (SK + LR). For each type of metamodel, we fit it using maximum likelihood estimates for the model parameters β_0, τ^2, and θ. We then use $\hat{\beta}_0$, $\hat{\tau}^2$, and $\hat{\theta}$ to do prediction at $N = 193$ equally spaced points in $[80, 120]$. In particular, to confirm our earlier conjecture on the detrimental effect of CRN on prediction, both CRN and independent sampling are employed to drive the simulations. As to the method of comparison, we use the estimated root mean squared error of prediction (ERMSE) over the grid of $N = 193$ prediction points,

$$\text{ERMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(C(S_0^i) - \hat{C}(S_0^i))^2}, \tag{37}$$

where $C(S_0^i)$ denotes the true option value at S_0^i, and $\hat{C}(S_0^i)$ represents the predicted value at the same point. It is known that $C(S_0)$ as defined in Equation (36) is

$$\begin{aligned} C(S_0) &= C(T, K; S_0, r, \sigma) \\ &= S_0\Phi\big(-d(K) + \sigma\sqrt{T}\big) - e^{-rT}K\Phi(-d(K)), \end{aligned}$$

where $\Phi(\cdot)$ stands for the CDF of the standard normal random variable.

We ran the experiment for 100 macroreplications. We first compare the ERMSEs obtained by SK, SK + IPA, and SK + LR with $k = 25$ design points, given that independent sampling and CRN are used in driving the simulation. In Figure 4, the left panel is for independent sampling and the right panel is for sampling using CRN. Inside each panel, three groups of box plots are shown from left to right according to an increasing number of simulation replications $n = 500, 2{,}000$, and $8{,}000$. Inside each group of three box plots, ERMSEs obtained by SK, SK + IPA, and SK + LR are shown from left to right. The box is formed by the 25th, 50th, and 75th percentiles; the whiskers extend plus-or-minus 1.5× the interquartile range beyond the box; and the diamonds are observations beyond that range. That the use of CRN increases ERMSE is obvious.

Having seen the disadvantage of using CRN with all three metamodels in Figure 4, Figures 5(a)–5(c) focus on the ERMSEs obtained by SK, SK + IPA, and SK + LR given that independent sampling is used to drive the simulation. The figures are ordered in an increasing number of design points k; in each figure, three groups of box plots are shown according to $n = 500, 2{,}000$, and $8{,}000$.

We summarize the findings in Figures 4 and 5(a)–5(c) as follows.

- CRN is detrimental. In particular, SK + IPA seems to be the most vulnerable to the adverse effect of CRN, as the increase in its corresponding ERMSE is the most dramatic among the three. We recommend using independent sampling in driving the simulation when stochastic kriging with gradient estimators is used for prediction.

Figure 4. The Black-Scholes model: Box plots of ERMSE when independent sampling (left) and CRN (right) are used with number of design points $k = 25$.

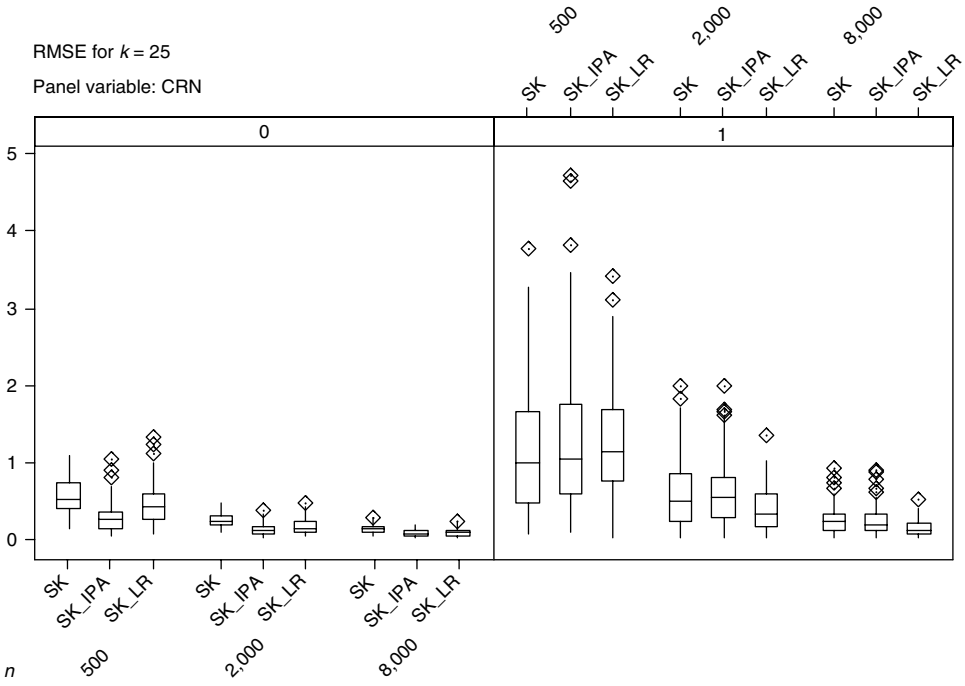
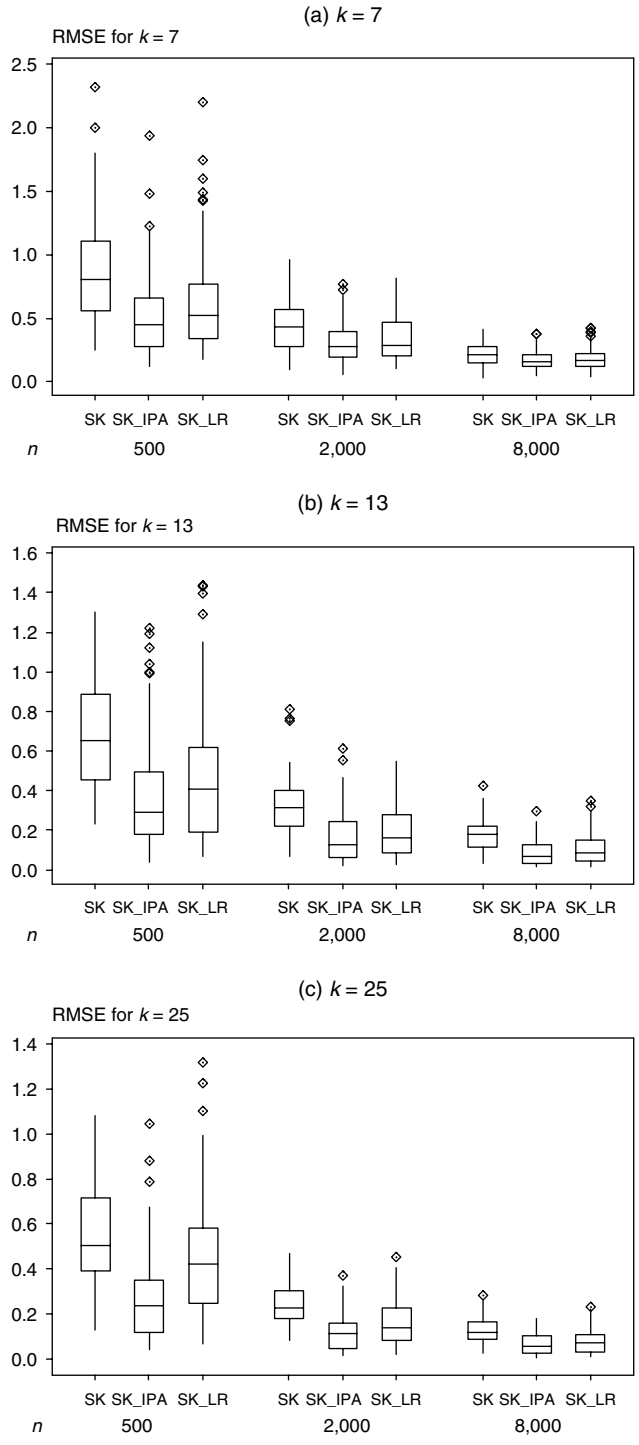


Figure 5. The Black Scholes model: Box plots of ERMSE of prediction over the 193 checkpoints when CRN is not used.



• Employing stochastic kriging with gradient estimators (IPA or LR) gives lower ERMSE of prediction as compared to not using the gradient estimators, as shown in Figures 5(a)–5(c) and easily significant with 95% confidence.

• In Figures 5(a)–5(c), we observe that under identical simulation conditions, SK + IPA gives the smallest ERMSEs, followed by SK + LR and SK for $k = 7, 13$ and 25. The advantage of SK + IPA in prediction over the other two approaches diminishes as the number of design points k increases; in particular, the ERMSEs of SK + IPA and SK + LR become indistinguishable. In this experiment, we observe that the variances of the gradient estimators are smaller for IPA; the ratio of the variance of an LR gradient estimator to the variance of its counterpart IPA gradient estimator ranges from 3.5 to 8. The correlations between the gradient estimators and the simulation responses are higher for LR gradient estimators than those of IPA's though. Specifically, the observed correlations for LR gradient estimators are between 0.9 and 0.97, whereas for IPA the correlations are between 0.82 and 0.88. It is worth mentioning that the simulation responses at all design points are very noisy, with variances between 226 and 487 times of those of the corresponding LR gradient estimators. These observations together with our analysis in §§3.2.1 and 3.2.2 explain to some extent why SK + IPA dominates SK + LR in prediction performance despite the fact that LR gradient estimators have higher correlations and not especially large variances in this experiment. It is also worth mentioning that as the number of replications increases, the ERMSEs of SK + IPA and SK + LR become increasingly closer, which is consistent with the previous analysis in §3.2.1. Indeed, in the other two examples mentioned at the end of §4, we have observed more significant differences in the variances and correlations associated with the IPA and LR gradient estimators, and the superiority of SK + IPA is more evident. This experiment confirms our earlier result that the variability of a gradient estimator plays a more important role in affecting the prediction performance than the correlation does.

• Increasing the number of simulation replications or the number of design points effectively reduces the ERMSE for all cases.

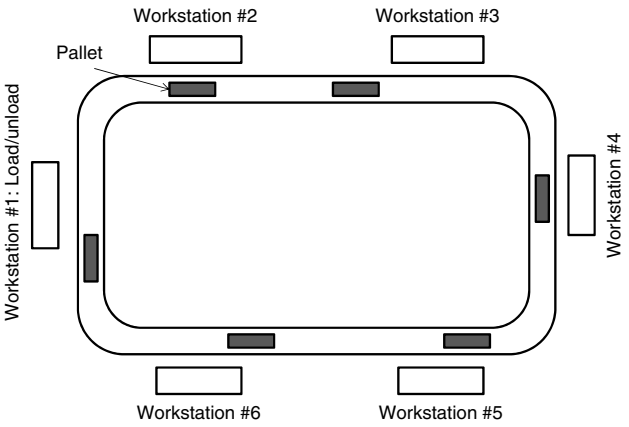
6. A Realistic Example: A Closed-Loop Flexible Assembly System

In this section we present a realistic example to illustrate what an analyst might face, and what results they might expect to achieve, if they enhanced a stochastic kriging metamodel by incorporating gradient estimators. Specifically, we consider the closed-loop flexible assembly system (CLFAS) described in Suri and Leung (1987). Flexible assembly systems are known for features such as reducing production lead times, lowering cost, and increasing flexibility. However, building a CLFAS involves a significant investment of time and resources, and hence studying approaches that can provide good and quick approximations to the system performance becomes important.

6.1. The Model: Dynamics and Simulation

The CLFAS under consideration consists of six automatic workstations connected by a conveyor as shown in Figure 6.

Figure 6. Schematic diagram of a closed-loop flexible assembly system.



The number of pallets in the system is also six. We chose six of each simply to match Suri and Leung (1987); our approach is no more complicated if there are more.

Unfinished parts enter and leave the CLFAS through station 1 and make one circuit through the system on the pallets. If the times between workstations are negligible, and no station is ever blocked, then the total assembly time equals the total operation time $\sum_{r=1}^{6} T_r$, where T_r represents the operation time at station $r = 1, 2, \ldots, 6$. The operation time T_r consists of the deterministic machine cycle time of x_r minutes and possibly an additional random time R_r to clear the machine if it jams. Therefore, $T_r = x_r + I\{\text{jam at station } r\}R_r$, where $I\{\cdot\}$ is the indicator function. The probability that a part causes station r to jam is α_r. In our simulation we took $\alpha_r = 0.005$, $r = 1, 2, \ldots, 6$, and let the R_r's be i.i.d. $U(0.1, 1.1)$ minutes, as suggested in the paper. Altering the jam probabilities and repair distribution will change the results but not our method.

Because the operation times are random, queueing does occur. If the queue in front of, say, station 3 is full then station 2 is blocked, meaning it cannot release a finished part. In our simulation there is space for one part to queue in front of each station. We are interested in predicting the expected throughput $\mathsf{Y}(\mathbf{x})$ of the first 5,000 parts completed by the CLFAS as a function of the station cycle times $\mathbf{x} = (x_1, x_2, \ldots, x_6)^\top$. This helps us to identify the bottleneck workstation(s) so that more resources can be devoted to improve the expected throughput.

IPA gradient estimators have been derived for many queueing systems, including closed, tandem, single-class networks. The algorithm given by Suri and Leung (1987) for estimating the throughput $\mathsf{Y}(\mathbf{x})$ and its gradient $\mathsf{D}^r(\mathbf{x}) = \partial \mathsf{Y}(\mathbf{x})/\partial x_r$, $r = 1, 2, \ldots, 6$ is restated below. Notice that this algorithm simply adds some accumulator variables that are easily updated during the course of the simulation without disrupting how it normally executes.

IPA Algorithm

1. Initialize $A_{i,r} \leftarrow 0$ for i, $r = 1, 2, \ldots, 6$. These are the accumulator variables for calculating the gradient.

2. At the end of an operation at station i with total operation time T_i, let $A_{i,i} \leftarrow A_{i,i} + dT_i/dx_i$, where dT_i/dx_i denotes the sample gradient of the random variable T_i. Because x_i is a location parameter of the distribution of T_i, $dT_i/dx_i = 1$, $i = 1, 2, \ldots, 6$.

3. If a pallet leaving station i going to station k terminates an idle period of station k, then set $A_{k,r} \leftarrow A_{i,r}$, $r = 1, 2, \ldots, 6$.

4. If a pallet leaving station i going to station k terminates a blocked period of station i, then set $A_{i,r} \leftarrow A_{k,r}$, $r = 1, 2, \ldots, 6$.

5. At the end of the simulation, let P denote the total number of parts completed and L be total length of simulation in time units. Estimate the throughput and its gradient by

$$\mathscr{Y}(\mathbf{x}) = \frac{P}{L},$$

$$\mathscr{D}^r(\mathbf{x}) = -\frac{\mathscr{Y}(\mathbf{x})}{L} A_{6r}, \quad r = 1, 2, \ldots, 6.$$

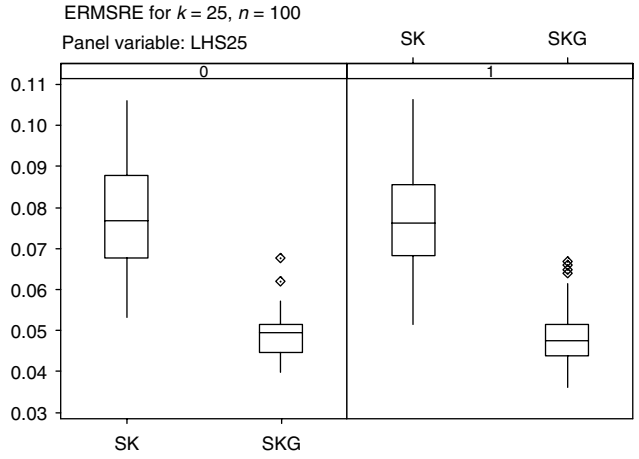
6.2. The Experimental Design and Results

Our goal is to compare the prediction performance of stochastic kriging and stochastic kriging with IPA gradient estimators (SKG) through this realistic manufacturing design problem.

The experimental design space is $\Omega_{\mathbf{x}} = [0.05, 0.15]^6$. For demonstration purposes, we use two different 25-point experiment designs: design 1 is the union of a 17-point maximin Latin-hypercube design and a 2_{III}^{6-3} fractional factorial design; and design 2 is a 25-point maximin Latin-hypercube design. At each design point $\mathbf{x} = (x_1, x_2, \ldots, x_6)^\top$, we simulate $n = 100$ independent replications of the CLFAS with a run length of $P = 5{,}000$ completed parts to obtain the simulation response $\mathscr{Y}_j(\mathbf{x})$ and gradient estimates $\mathscr{D}_j^r(\mathbf{x})$, for $r = 1, 2, \ldots, 6$, $j = 1, 2, \ldots, 100$. These are the only data we need to fit the metamodels and compute predictions.

Assuming (correctly) that we have little information about the true response surface, a constant trend model $\mathbf{f}^\top(\mathbf{x})\boldsymbol{\beta} = \beta_0$ is selected. We fit two stochastic kriging metamodels, one with and one without the IPA gradient estimators, and used maximum likelihood estimation to obtain the stochastic kriging parameters $\hat{\beta}_0$, $\hat{\tau}^2$, $\hat{\theta}_r$, $r = 1, 2, \ldots, 6$. With these parameters we did prediction at $N = 100$ Latin-hypercube sampled checkpoints throughout $\Omega_{\mathbf{x}}$. Because the true throughputs at the checkpoints are unknown, we approximated them by simulating 1,000 replications of the CLFAS at each checkpoint. The simulation model and experiment were programmed in VBA, and the metamodel estimation and prediction were done in Matlab using the simulation data.

Figure 7. The CLFAS model: Box plots of ERMSRE over the 100 checkpoints: design 1 (left panel) and design 2 (right panel).



For a summary presentation of the prediction performances of SK and SKG, we computed the estimated root mean squared relative error of prediction (ERMSRE) over the $N = 100$ checkpoints, which is defined as

$$\text{ERMSRE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\Upsilon(\mathbf{x}_i) - \hat{\Upsilon}(\mathbf{x}_i)}{\Upsilon(\mathbf{x}_i)}\right)^2}. \tag{38}$$

We reran the experiment for 50 macro-replications, computing the ERMSRE using the $N = 100$ checkpoints in each. Figure 7 contains box plots of these 50 ERMSRE values. The left panel is for design 1 and the right panel is for design 2. Inside each panel, the left box plot of ERMSRE is for SK and the right is for SKG. It is evident that SKG leads to much smaller ERMSRE than SK and hence better global prediction performance. Notice that the two experimental designs do not make a significant difference. We had assumed that the 2^{6-3}_{III} fractional factorial design would be helpful for SK by avoiding extrapolation error in $\Omega_{\mathbf{x}} = [0.05, 0.15]^6$; SKG is more resistant to extrapolation error.

7. Conclusion

In this paper, we introduced the idea of incorporating gradient estimators into stochastic kriging metamodels to improve response surface prediction, and we evaluated the idea via mathematical analysis and two experiments. The experiment results demonstrated the advantages of the enhanced metamodel over stochastic kriging in providing better prediction performance. The results also showed that using CRN degrades the prediction performance of stochastic kriging with gradient estimators; in particular, stochastic kriging with IPA gradient estimators seems more susceptible to the adverse effect of CRN than stochastic kriging with LR gradient estimators. In general, we recommend to use stochastic kriging with IPA gradient estimators, when available, and independent sampling to drive the simulation. Research on experiment design for stochastic kriging with gradient estimators is an obvious next step.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/opre.1120.1143.

Acknowledgments

The authors thank the referees and editors for comments and suggestions that improved the paper. This paper is based upon work supported by the National Science Foundation [Grant CMMI-0900354].

References

Ankenman BE, Nelson BL, Staum J (2010) Stochastic kriging for simulation metamodeling. *Oper. Res.* 58(2):371–382.

Chen X, Ankenman B, Nelson BL (2010) Common random numbers and stochastic kriging. Johansson B, Jain S, Montoya-Torres J, Hugan, J, Yucesan E, eds. *Proc. 2010 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 947–956.

Chen X, Ankenman B, Nelson BL (2012) The effects of common random numbers on stochastic kriging metamodels. *ACM Trans. Modeling Comput. Simulation* 22:7/1–7/20.

Forrester AIJ, Keaney AJ (2009) Recent advances in surrogate-based optimization. *Progress in Aerospace Sci.* 45:50–79.

Fu MC (2006) Stochastic gradient estimation. Henderson SG, Nelson BL, eds. *Elsevier Handbooks in Operations Research and Management Science: Simulation* (Elsevier, New York), 575–616.

Glasserman P (1991) *Gradient Estimation via Perturbation Analysis* (Kluwer, Boston).

Glasserman P (2004) *Monte Carlo Methods in Financial Engineering* (Springer, New York).

Heidergott B, Vázquez-Abad FJ, Pflug F, Farenhorst-Yuan T (2010) Gradient estimation for discrete-event systems by measure-valued differentiation. *ACM Trans. Modeling Comput. Simulation* 20:5/1–5/28.

L'Ecuyer P (1990) A unified view of the IPA, SF, and LR gradient estimation techniques. *Management Sci.* 36(11):1364–1383.

Mitchell T, Morris M, Ylvisaker D (1994) Asymptotically optimum experimental designs for prediction of deterministic functions given derivative information. *J. Statist. Planning and Inference* 41:377–389.

Morris MD, Mitchell TJ, Ylvisaker D (1993) Bayesian design and analysis of computer experiments: Use of derivatives in surface prediction. *Technometrics* 35:243–255.

Näther W, Šimák J (2003) Effective observation of random processes using derivatives. *Metrika* 58:71–84.

Parzen E (1962) *Stochastic Processes* (Holden-Day, San Francisco).

Rubinstein RY, Shapiro A (1993) *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method* (John Wiley & Sons, New York).

Sacks J, Welch WJ, Mitchell TJ, Wynn HP (1989) Design and analysis of computer experiments. *Statist. Sci.* 4:409–423.

Santner TJ, Williams BJ, Notz WI (2003) *The Design and Analysis of Computer Experiments* (Springer, New York).

Siem AYD (2008) Property preservation and quality measures in metamodels. Open access publications, Tilburg University urn:nbn:nl:ui:12-364657. Tilburg University, Tilburg, The Netherlands.

Šimák J (2002) On experimental designs for derivative random fields. Ph.D. thesis. TU Bergakademie Freiberg, Freiberg, Germany.

Spall J (2003) *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control* (John Wiley & Sons, Inc., New York).

Stein ML (1999) *Interpolation of Spatial Data: Some Theory for Kriging* (Springer, New York).

Stephenson G (2010) Using derivative information in the statistical analysis of computer models. Ph.D. thesis. School of Ocean and Earth Science, University of Southampton, Southhampton UK.

Suri R, Leung YT (1987) Single run optimization of a SIMAN model for closed loop flexible systems. Thesen A, Grant H, Kelton WD, eds. *Proc. 1987 Winter Simulation Conf.* (ACM, New York), 738–748.

Yamazaki W, Rumpfkeil MP, Mavriplis DJ (2010) Design optimization utilizing gradient/Hessian enhanced surrogate model. AIAA Paper 2010-4363, *40th Fluid Dynamics Conference and Exhibit, Chicago, IL.*

Xi Chen is an assistant professor in the Department of Statistical Sciences and Operations Research at Virginia Commonwealth University. Her research interests include but are not limited to stochastic modeling and simulation, applied probability and statistics, computer experiment design and analysis, and simulation optimization.

Bruce E. Ankenman is a Charles Deering McCormick Professor of Teaching Excellence and an associate professor in the Department of Industrial Engineering and Management Sciences at Northwestern University. His research interests primarily deal with the design and analysis of experiments that are used to build models for physical systems or metamodels for simulated systems. He is the codirector of the Segal Design Institute.

Barry L. Nelson is the Walter P. Murphy Professor and Chair of the Department of Industrial Engineering and Management Sciences at Northwestern University. His research addresses statistical issues in the design and analysis of stochastic computer simulation experiments, including metamodeling, multivariate input modeling, simulation optimization, input uncertainty quantification, and variance reduction. He is a Fellow of INFORMS and IIE.