Stochastic Kriging with Biased Sample Estimates

Xi Chen∗

Virginia Commonwealth University

and

Kyoung-Kuk Kim†

Korea Advanced Institute of Science and Technology

Nov 2013

Abstract

Stochastic kriging has been studied as an effective metamodeling technique for approximating response surfaces in the context of stochastic simulation. In a simulation experiment an analyst typically needs to estimate relevant metamodel parameters and further do prediction, therefore the impact of parameter estimation on the performance of the metamodel-based predictor has drawn some attention in the literature. However, how the standard stochastic kriging predictors are affected by the presence of bias in finite-sample estimates has not yet been fully investigated. In this paper, we study prediction performances and investigate optimal budget allocation rules subject to a fixed computational budget constraint. Furthermore, we extend the analysis to two-level or nested simulation, which has been recently documented in the risk management literature, with biased estimators.

Keywords: Stochastic kriging; Simulation experimental design; Nested simulation; Optimal budget allocation

1 Introduction

Simulation models, both deterministic and stochastic, can provide high-fidelity predictions of the behavior of complex systems at different settings of their controllable factors or decision variables. However, simulation experiments may be (and frequently are) time-consuming or costly to execute, potentially limiting the usefulness of simulation in some situations, including support for real-time decision making and system optimization. To mitigate this deficiency, carefully designed simulation experiments can be employed to fit metamodels—approximations of some aspects of system performance—and fitted metamodels may be exercised in real time. Ideally, they provide the fidelity of full simulation models with the ease of use of, say, a spreadsheet model.

To be useful, metamodels built on simulation estimates need to be accurate. However, many reasons can lead to metamodel inadequacy, and it is always a challenge to detect and subsequently compensate for

∗Department of Statistical Sciences and Operations Research, E-mail: xchen4@vcu.edu
†Corresponding author, Department of Industrial and Systems Engineering, E-mail: catenoid@kaist.ac.kr
it. Metamodel inadequacy has been investigated in the deterministic computer experiment literature. For example, Kennedy and O’Hagan [2001] present computer model calibration via a Bayesian approach, in which a list of sources of computer model uncertainties is given, including observation errors. In the context of stochastic simulation, a plethora of examples exist that illustrate simulation response estimates obtained with a finite computational budget are biased: In a steady-state simulation of a queueing system with initialization bias, it is well-known that even if all the computational budget is dedicated to simulating a single sample path, the sample-path average waiting time as an estimator is biased for the steady-state expected waiting time. Quantiles of an empirical distribution constructed based on a finite number of simulation response estimates are biased for the true quantiles of the underlying distribution.

There is another recent and interesting simulation context, which has been quite popular in the risk management literature, namely, two-level simulation or nested simulation. Two-level simulation involves an outer-level simulation (e.g., scenario generation) and an inner-level simulation (e.g., computing expectations conditional on a generated scenario) to obtain a single simulation output. Mathematically, it can be written as \( \mathbb{E}[g(X|Z)] \) where \( \mathbb{E} \) is mathematical expectation and \( g(X|Z) \) is a functional of the distribution function of a random variable (or vector) \( X \) conditional on a simulated value \( Z \). For instance, \( g(X|Z) = \mathbb{E}[f(X)|Z] \) for some function \( f \). See Broadie et al. [2011], Gordy and Juneja [2010], Lan et al. [2010], and Sun et al. [2011]. To our best knowledge, two-level simulation has not been studied in the stochastic kriging context except by Liu and Staum [2010] and Chen et al. [2012b]; however, neither paper considers the impact of bias in the simulation response estimates nor do they provide an in-depth analysis of computational budget allocation problems. Acknowledging that the aforementioned simulation experiments typically take considerable computational time to get reasonably accurate estimates, metamodels would seem to be a natural candidate tool to help.

The objective of this paper is to get more benefits from a simulation experiment that is designed for building a stochastic kriging metamodel for prediction with a fixed computational budget, while taking into account possible bias in simulation response estimates. We have in mind a situation where a set of design points and a computational budget at each of them have been specified to exercise the simulation model, however, a budget allocation rule needs to be determined for good performance of the resulting metamodel predictor. Here is a summary of our findings: given a fixed but sufficiently large computational budget \( N \) to expend at a design point which consists of the number of replications \( n \) and simulation efforts on each of the replications, typically a moderate number of simulation replications \( n (n \leq 50) \) works best; hence there is no need to increase \( n \) even though an additional budget becomes available. Furthermore, in the two-level simulation context, it is optimal to make the number of inner level simulation runs of order \( O(N^{1/3}) \) and set the product of the number of outer level simulated scenarios and the number of replications to be of order \( O(N^{2/3}) \), in which case the optimal number of replications \( n \) to use is typically found to be moderate as well.

The remainder of this paper is organized as follows. In Section 2, a brief review of kriging and standard stochastic kriging metamodels is presented. In Section 3, we extend stochastic kriging in recognition of bias present in simulation response estimates and analyze the associated impact on the mean squared error of standard stochastic kriging predictors. We then study optimal budget allocation rules when parameters, in particular, the trend model vector \( \beta \) and the intrinsic variance-covariance matrix \( \Sigma_e \), which quantifies the sampling variability inherent in stochastic simulation, are estimated. Two numerical examples follow for an illustration. In Section 4, through a parallel analysis similar to Section 3 we arrive at optimal budget allocation rules for two-level simulation experiments. Numerical results are presented for an example of a
portfolio of financial derivatives. Section 5 concludes the paper and provides some perspectives on future research topics.

2 A Review on Kriging and Stochastic Kriging

In the past few decades, kriging metamodels were applied to approximate the outputs of deterministic computer experiments. As a useful tool for engineering design, kriging has become increasingly popular in recent years [e.g. Sacks et al., 1989, Santner et al., 2003, Kleijnen, 2008]. In a deterministic computer experiment, responses are obtained without noise and a kriging metamodel can be developed after observing $Y(x_i)$ at a set of design points $x_i$, with $x_i \in \mathbb{R}^d$, $i = 1, 2, \ldots, k$. The unknown response surface is modeled by kriging as

$$Y(x) = f(x)^T \beta + M(x),$$

where $f(x)$ and $\beta$ are, respectively, a $p \times 1$ vector of known functions of $x$ and a $p \times 1$ vector of unknown parameters. In kriging it is assumed that $M(x)$ is a realization of a mean zero stationary Gaussian random field with $\mathbb{E}[|M(x)|^2] < \infty$ for all $x \in \mathbb{R}^d$. The response surface $Y(x)$ can be thought of as being sampled from a space of functions mapping $\mathbb{R}^d \to \mathbb{R}$, in which functions are assumed to exhibit spatial correlation. Specifically, the spatial covariance function between the responses at two points is usually described as follows:

$$\Sigma_M(x, w) = \text{Cov}(M(x), M(w)) = \tau^2 R(x-w; \theta),$$

where $\tau^2$ denotes the spatial variance of the random process. The correlation function $R(\cdot, \theta)$ satisfies $\lim_{\|x-w\|\to \infty} R(x-w; \theta) = 0$ and $R(0; \theta) = 1$, and the $d \times 1$ vector $\theta$ contains the spatial correlation parameter $\theta_i$ which controls how quickly the correlation decreases as two points become farther apart in the direction of the $i$th coordinate, $i = 1, 2, \ldots, d$. One particular choice of the correlation function that is quite popular in practice is the Gaussian correlation function

$$R(x-w; \theta) = \exp \left\{ -\sum_{j=1}^{d} \theta_j |x_j - w_j|^2 \right\},$$

where the parameters $\theta_j \geq 0$, $j = 1, 2, \ldots, d$ for different coordinate directions are not necessarily equal. The sample paths of a spatial process are infinitely differentiable if the Gaussian correlation function is used; more details can be found in Section 2.1.4 of Šimáč [2002], Section 2.3.4 of Santner et al. [2003] and Chapter 3 of Parzen [1962]. For convenience, we use the Gaussian correlation function for all the numerical examples demonstrated in this paper.

Recently, Ankenman et al. [2010] proposed stochastic kriging methodology for stochastic simulation experiments. They distinguish two different sources of uncertainty involved; one is the stochastic nature from the simulation itself, and the other is the imposed stochastic nature of the aforementioned spatial random field built for a given problem. Ankenman et al. [2010] refer to them as intrinsic and extrinsic uncertainties, respectively, and suggest to fully account for both types of uncertainties. They show that the joint modeling approach of the two has an impact on experimental design, estimation, and inference. The remainder of this section briefly summarizes the basic ingredients of stochastic kriging methodology. We refer the reader to Ankenman et al. [2010] for full details.
The stochastic kriging metamodel for the simulation output at design point \( x_i \) on the \( j \)th simulation replication is given as

\[
\tilde{Y}_j(x_i) = Y(x_i) + \varepsilon_j(x_i) \\
= f(x_i)\top\beta + M(x_i) + \varepsilon_j(x_i), \quad j = 1, 2, \ldots, n_i, \quad i = 1, 2, \ldots, k. \tag{1}
\]

In (1), it is assumed that the simulation experiment has been run at \( k \) design points \( \{x_i\}_{i=1}^k \) and \( n_i \) simulation replications are made at design point \( x_i \). The last term \( \varepsilon_j(x_i) \) accounts for zero-mean simulation error realized at the \( i \)th design point on the \( j \)th simulation replication. Furthermore, it is independent of \( M(x_i) \) for \( i = 1, 2, \ldots, k \). Regarding selecting the design points, space-filling designs are often used for building metamodels with outputs from stochastic simulation experiments, similar to the practice in deterministic computer experiments. One popular choice is the Latin hypercube sampling design (LHS). For an introduction to LHS, see, for example, Section 4.5.1 of Kleijnen [2008] and Section 5.2.2 of Santner et al. [2003].

Let the sample mean of the simulation output at design point \( x_i \) be

\[
\bar{Y}(x_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_j(x_i) = Y(x_i) + \frac{1}{n_i} \sum_{j=1}^{n_i} \varepsilon_j(x_i), \quad i = 1, 2, \ldots, k,
\]

and denote the \( k \times 1 \) vector of the sample averages as \( \bar{Y} = (\bar{Y}(x_1), \ldots, \bar{Y}(x_k))\top \). Write \( \Sigma_M \) for the \( k \times k \) spatial covariance matrix of \( M \) at the \( k \) design points \( \{x_i\}_{i=1}^k \). Suppose that we make a prediction at point \( x_0 \). Define \( \Sigma_M(x_0, \cdot) \) to be the \( k \times 1 \) vector that contains the spatial covariances between \( x_0 \) and each of the \( k \) design points, that is,

\[
\Sigma_M(x_0, \cdot) = \left( \Sigma_M(x_0, x_1), \ldots, \Sigma_M(x_0, x_k) \right)\top.
\]

Regarding the simulation errors \( \varepsilon_j(x_i), j = 1, 2, \ldots, n_i, i = 1, 2, \ldots, k \), we assume that they are independent across replications and design points. Originally in Ankenman et al. [2010], the possibility of correlated random errors across design points was permitted so that one can implement common random numbers (CRN). However, as shown in Chen et al. [2012a], the use of CRN may degrade the performance of stochastic kriging predictors. In this paper, therefore, we assume that CRN is not used across design points in simulation experiments. It follows that without using CRN the \( k \times 1 \) vector of the averaged simulation errors \( \bar{\varepsilon} = (\bar{\varepsilon}(x_1), \bar{\varepsilon}(x_2), \ldots, \bar{\varepsilon}(x_k)) \) has a variance-covariance matrix (denoted by \( \Sigma_\varepsilon \)) that takes a diagonal form. Its diagonal entries are as follows

\[
(\Sigma_\varepsilon)_{ii} = \text{Var}\left[ \frac{1}{n_i} \sum_{j=1}^{n_i} \varepsilon_j(x_i) \right], \quad i = 1, 2, \ldots, k.
\]

Appendix EC.1 of Ankenman et al. [2010] shows that when \( \beta, \Sigma_\varepsilon \) and the spatial parameters \( \tau^2, \theta \) are known, the MSE-optimal predictor of \( Y(x_0) \) (also known as the best linear predictor (BLP)) is

\[
\tilde{Y}(x_0) = f(x_0)\top\beta + \Sigma_M(x_0, \cdot)\top \Sigma^{-1} (\bar{Y} - F\beta), \tag{2}
\]

and its corresponding mean squared error of prediction (MSE) is

\[
\text{MSE}(\tilde{Y}(x_0)) = \Sigma_M(x_0, x_0) - \Sigma_M(x_0, \cdot)\top \Sigma^{-1} \Sigma_M(x_0, \cdot), \tag{3}
\]
where $\Sigma = \Sigma_M + \Sigma_e$, and $F = (f(x_1)^T, \ldots, f(x_k)^T)^T$. When the parameter vector $\beta$ is unknown and needs to be estimated, Appendix A.1 of Chen et al. [2012a] shows that the MSE-optimal predictor of $Y(x_0)$ (also known as the best linear unbiased predictor (BLUP)) is given by

$$\tilde{Y}(x_0) = f(x_0)^T \hat{\beta} + \Sigma_M(x_0, \cdot)^T \Sigma^{-1} \left( \tilde{Y} - F \hat{\beta} \right),$$

(4)

where

$$\hat{\beta} = \left( F^T \Sigma^{-1} F \right)^{-1} F^T \Sigma^{-1} \tilde{Y}$$

(5)

is the generalized least squares estimator of $\beta$. The corresponding MSE equals

$$\text{MSE}(\tilde{Y}(x_0)) = \Sigma_M(x_0, x_0) - \Sigma_M(x_0, \cdot)^T \Sigma^{-1} \Sigma_M(x_0, \cdot) + \eta^T (F^T \Sigma^{-1} F)^{-1} \eta,$$

(6)

where $\eta = f(x_0) - F^T \Sigma^{-1} \Sigma_M(x_0, \cdot)$. Notice that the last term in (6) is the amount by which the MSE is inflated due to the estimation of $\beta$.

For an actual implementation of stochastic kriging methodology, one needs to estimate the necessary parameters, i.e., $\beta$, $\Sigma_e$ and the spatial parameters $\theta$ and $\tau^2$. With an additional assumption that the simulation errors $\epsilon_j(x_i), j = 1, 2, \ldots, n_i$ at design point $x_i$ are i.i.d. normal random variables, Theorem 1 in Ankenman et al. [2010] proves that the plug-in predictor $\tilde{Y}(x_0)$ obtained by replacing $\Sigma_e$ with its sample counterpart $\hat{\Sigma}_e$ is unbiased for $Y(x_0)$. The suggested approach is to first estimate $\hat{\Sigma}_e$ using the sample covariance matrix $\hat{\Sigma}_e$ whose diagonal entries can be specified as

$$(\hat{\Sigma}_e)_{ii} = \frac{1}{n_i(n_i - 1)} \sum_{j=1}^{n_i} (Y_j(x_i) - \tilde{Y}(x_i))^2, \quad i = 1, 2, \ldots, k,$$

and then substitute $\hat{\Sigma}_e$ into the log-likelihood function and do optimization to get the maximum likelihood estimates (MLEs) for $\theta$, $\tau^2$ and $\beta$.

3 Stochastic Kriging with Biased Finite-Sample Estimates

3.1 The model modification

We begin our discussion of extending stochastic kriging to incorporate biased sample estimates with a simple example. Suppose that we are interested in obtaining a point estimator of the $p$-quantile $v_p$ for an unknown distribution of a random variable $L$ using i.i.d. simulated responses. For example, $L_1, L_2, \ldots, L_m$ are a sample of size $m$. Consider two choices of point estimator. The first one is a single sample estimator that is an order statistic from a single large sample:

$$\hat{v}_p^{(S)} = L_{[pm]:m},$$

which is the $[pm]$th order statistic from $m$ observations. Here, $[x]$ denotes the smallest integer greater than or equal to $x$. The second one is a multi-sample estimator that is the average of the appropriate order statistics from each of several smaller samples. Specifically, one can divide the $m$ observations of $L_i$’s
into \( n \) independent samples of size \( n_s \) (suppose that \( m \) is a multiple of \( n \)), say \( L_1^{(h)}, L_2^{(h)}, \ldots, L_{n_s}^{(h)} \) for \( h = 1, 2, \ldots, n \), and then construct the estimator as

\[
\tilde{Y}_p^{(M)} = \frac{1}{n} \sum_{h=1}^{n} L_{[pm_s]:n_s}^{(h)}.
\]

Notice that the total sample sizes used to construct the two estimators are the same. However, it is not obvious to determine which estimator is better. This problem has been analyzed in Zelterman [1987] based on asymptotic behaviors of associated MSEs.

Extending this example further into the stochastic kriging context, consider the problem of building a response surface using stochastic kriging to predict the \( p \)-quantile \( Y(x_0) \) of a certain random variable, say \( L(x_0) \). The task at hand is to get a stochastic kriging predictor \( \hat{Y}(x_0) \) for the \( p \)-quantile of \( L(x_0) \) given the simulation responses \( \{Y_j(x_i)\}_{j=1}^{n}, i = 1, 2, \ldots, k \). Here the quantiles are calculated through i.i.d. samples. In this example \( Y_j(x_i) \) denotes a quantile estimate obtained at design point \( x_i \) on the \( j \)th replication; henceforth the simulation experiment described above would be equivalent to implementing the aforementioned multi-sample estimation method at each of the \( k \) design points. The total number of simulation runs (the total computational budget) needed amounts to \( kmn_s \), in case where \( n_i = n \) for all \( i \) for instance. As noted in Section 1, the quantile estimators based on order statistics are biased and for this quantile estimation example, the asymptotic order of bias is known to be \( O(n_s^{-1}) \) if a sample of size \( n_s \) is used.

We modify the metamodel given in (1) into the following form to account for bias present in a simulation response estimate:

\[
Y_j(x_i) = f(x_i)^T \beta + M(x_i) + \varepsilon_j(x_i; n_s) + \xi(x_i; n_s) = Y_j^o + \xi(x_i; n_s), \quad j = 1, 2, \ldots, n_i, \quad i = 1, 2, \ldots, k. \tag{7}
\]

In (7), we use \( Y_j^o \) to represent the unbiased simulation response estimate that would be obtained at design point \( x_i \) if there were no bias present. The term \( \xi(x_i; n_s) \) stands for the bias generated at design point \( x_i \) when the sample size is \( n_s \). Because the bias in a simulation response estimate typically decreases in the sample size and the decreasing rate depends on specific simulation methods implemented, we model \( \xi(x_i; n_s) \) as a deterministic quantity \( B_s n_s^{-\alpha} \). That is, the bias decreases with order \( \alpha \) as \( n_s \) increases; the \( B_s \)'s are unknown constants that may be location dependent but nevertheless independent of \( n_s \). Therefore, we can write the vector that contains averaged unbiased simulation response estimates as \( \hat{Y}^o = \hat{Y} - \xi(n_s) \), in which \( \xi(n_s) \) denotes the vector of bias incurred at \( k \) design points. Notice that we only observe \( \hat{Y} \) through a simulation experiment and that increasing numbers of simulation replications \( \{n_i\}_{i=1}^{k} \) while keeping \( n_s \) fixed do not help reduce bias.

To simplify analysis, throughout this section we assume that the same number of replications \( n \) and the same sample size \( n_s \) are used to obtain a single simulation response estimate \( Y_j(x_i), i = 1, 2, \ldots, k \). Recall that the simulation error induced on the \( j \)th simulation replication is denoted by \( \varepsilon_j(x_i; n_s) \) which is assumed to have mean zero. Let us next make the form of the intrinsic variance-covariance matrix explicit: Given that CRN is not implemented, the intrinsic variance-covariance matrix \( \Sigma_x \) becomes a diagonal matrix with diagonal entries of form \( n^{-1} \sigma_i^2(n_s) \) where \( \sigma_i^2(n_s) = \text{Var}(\varepsilon_j(x_i; n_s)) \). As the sample size \( n_s \) becomes sufficiently large, the intrinsic variance at \( x_i \) is assumed to be \( A_1 n_s^{-1}, i = 1, 2, \ldots, k \), as in the case of the aforementioned quantile estimators [see Gordy and Juneja, 2010; Hong and Liu, 2009, for other examples].
Notice that the $A_i$’s are some positive constants that may be location dependent but independent of $n_s$ and $n$. Consequently, as $n_s$ gets sufficiently large, we have
\[
\Sigma_e \approx \frac{1}{nm_s}\text{diag}\left\{A_1, A_2, \ldots, A_k\right\}.
\]

3.2 Analysis of the modified stochastic kriging metamodel

In the upcoming subsections we study the impact of bias on performance of the standard stochastic kriging predictor, and furthermore obtain the best allocation rule of a fixed computational budget at each design point. The objective is to make the standard stochastic kriging predictor attain the minimum MSE of prediction in the presence of bias. We first start with the case where all parameters are known except $\beta$, and then move onto the case where $\Sigma_e$ further needs to be estimated. Spatial parameters $\tau^2$ and $\theta$ are assumed to be known throughout.

3.2.1 Intrinsic variance-covariance matrix $\Sigma_e$ known

Recall that the standard stochastic kriging predictor with known $\beta$ is given in (2). If there were no bias, the observation vector would be $\tilde{Y}^o$ and it follows from standard arguments under the stochastic kriging framework that the predictor $\tilde{Y}(x_0)^o$ based on unbiased response estimates is the BLP of $Y(x_0)$, with its mean squared error of prediction given by (3). Let us restate it here,
\[
\text{MSE}(\tilde{Y}(x_0)^o) = \Sigma_M(x_0, x_0) - \Sigma_M(x_0, \cdot)^\top \Sigma^{-1} \Sigma_M(x_0, \cdot).
\]
(8)

In contrast, when bias is present in simulation response estimates, the observation vector becomes $\tilde{Y} = \tilde{Y}^o + \xi(n_s)$. This makes the standard stochastic kriging predictor $\tilde{Y}(x_0)$ biased. Also, its MSE gets inflated and can be shown to take the following form
\[
\text{MSE}\left(\tilde{Y}(x_0)\right) = \left(\Sigma_M(x_0, \cdot)^\top \Sigma^{-1} \xi(n_s)\right)^2 + \text{MSE}(\tilde{Y}(x_0)^o).
\]

Now consider the standard stochastic kriging predictor with unknown $\beta$. Recall the estimator $\hat{\beta}$ and the corresponding predictor $\tilde{Y}(x_0)$ are given by (5) and (4), respectively. To ease notation, let $\hat{\beta} = W\hat{Y}$ where $W = (F^\top \Sigma^{-1} F)^{-1} F^\top \Sigma^{-1}$. It follows immediately that $\hat{\beta}$ is no longer unbiased for $\beta$ due to the bias in the simulation response estimates. Specifically, $E\left[\hat{\beta}\right] = W(\beta + \xi(n_s)) = \beta + W\xi(n_s)$ and hence the bias of $\tilde{Y}(x_0)$ can be easily seen through
\[
E\left[\tilde{Y}(x_0)\right] = f(x_0)^\top E\left[\hat{\beta}\right] + \Sigma_M(x_0, \cdot)^\top \Sigma^{-1} \left(F\beta + \xi(n_s) - FE\left[\hat{\beta}\right]\right)
\]
\[
= f(x_0)^\top \beta + f(x_0)^\top W\xi(n_s) + \Sigma_M(x_0, \cdot)^\top \Sigma^{-1} (I - FW) \xi(n_s)
\]
\[
= f(x_0)^\top \beta + w^\top \xi(n_s),
\]
where $w^\top = f(x_0)^\top W + \Sigma_M(x_0, \cdot)^\top \Sigma^{-1} (I - FW)$. The MSE of the predictor $\tilde{Y}(x_0)$ in this case can be expressed as
\[
\text{MSE}\left(\tilde{Y}(x_0)\right) = \left(w^\top \xi(n_s)\right)^2 + \eta^\top (F^\top \Sigma^{-1} F)^{-1} \eta + \text{MSE}(\tilde{Y}(x_0)^o),
\]
(9)
where \( \eta = f(x_0) - F^\top \Sigma^{-1} \Sigma_M(x_0, \cdot) \). Apparently the MSE of the standard stochastic kriging predictor is inflated by \( O(n_s^{-2\alpha}) \) as compared to the unbiased case given a fixed computational budget of \( N = nn_s \) at each design point. Therefore, the larger \( n_s \) is, the smaller the MSE \( \left( \hat{Y}(x_0) \right) \). We note that the MSE optimality in the next proposition is based on (9). The results in other propositions throughout this paper can be understood in a similar fashion.

**Proposition 1** Suppose that the design points are preselected. Also suppose that the intrinsic variance-covariance matrix \( \Sigma_e \) and spatial parameters are known. Then it is MSE-optimal to make only one replication \((n = 1)\) and spend all the computational budget \( N \) to make the sample size \( n_s \) as large as possible, i.e., \( n_s = N \).

A budget allocation rule \((n, n_s)\) determines the way we break a single sample of size \( N \) into \( n \) smaller samples each of size \( n_s \). And as long as all the other parameters except \( \beta \) are known, we should always put all the computational effort into one single simulation replication based on the proposition above. Since parameter estimation is an essential part of implementing the stochastic kriging technique, in the next subsection, we analyze the case of estimating the unknown intrinsic variance-covariance matrix and its impact on the stochastic kriging predictor.

### 3.2.2 Intrinsic variance-covariance matrix \( \Sigma_e \) unknown

In this subsection, we move one step further to study the effect of estimating the matrix \( \Sigma_e \) on the standard stochastic kriging predictor \( \hat{Y}(x_0) \) with unknown \( \beta \) in (4). We continue to assume that all the spatial parameters are given. Recall that without using CRN, the matrix \( \Sigma_e \) is diagonal and we can estimate its diagonal entries by sample variance estimators. Specifically,

\[
(\Sigma_e)_{ii} = \frac{1}{n} \sigma_{i0}^2(n_s), \quad \hat{\sigma}_{i0}^2(n_s) = \frac{1}{n-1} \sum_{j=1}^n (Y_j(x_i) - \bar{Y}(x_i))^2.
\]

For notational simplicity, we write

\[
\sigma = (\sigma_{10}^2, \sigma_{20}^2, \ldots, \sigma_{k0}^2)^\top, \quad \hat{\sigma} = (\hat{\sigma}_{10}^2(n_s), \hat{\sigma}_{20}^2(n_s), \ldots, \hat{\sigma}_{k0}^2(n_s))^\top.
\]

In short we use \( \Sigma_e = n^{-1} \text{diag}\{\sigma\} \) and \( \hat{\Sigma}_e = n^{-1} \text{diag}\{\hat{\sigma}\} \) to denote the true intrinsic variance-covariance matrix and its estimator. In the forthcoming analysis, it is convenient to work with two slightly different predictors constructed from the biased average simulation response vector \( \bar{Y} \) and the unbiased one \( \bar{Y}^o \):

\[
\hat{Y}(x_0; \sigma) = f(x_0)^\top \hat{\beta} + \Sigma_M(x_0, \cdot)^\top \Sigma^{-1} \left( \bar{Y} - F \hat{\beta} \right) = w(\sigma)^\top \bar{Y},
\]

\[
\hat{Y}(x_0; \sigma)^o = f(x_0)^\top \hat{\beta} + \Sigma_M(x_0, \cdot)^\top \Sigma^{-1} \left( \bar{Y}^o - F \hat{\beta} \right) = w(\sigma)^\top \bar{Y}^o,
\]

where \( \bar{Y}^o = \bar{Y} - \xi(n_s) \) denotes the average simulation response vector that would be obtained if there were no bias, and \( w(\sigma)^\top = f(x_0)^\top W + \Sigma_M(x_0, \cdot)^\top \Sigma^{-1} (I - FW) \) is as given in the previous subsection. Here we make the dependence on \( \sigma \) explicit to emphasize that these predictors depend on the true intrinsic covariance vector \( \sigma \). Therefore, if it were possible to obtain simulation responses without any bias, then \( \hat{Y}(x_0; \sigma)^o \) would have been our standard MSE-optimal predictor.
the resulting stochastic kriging predictor \( \hat{Y}(x_0; \hat{\sigma}) \) is obtained by replacing \( \sigma \) with \( \hat{\sigma} \). Corresponding to the two predictors given above, we have

\[
\hat{Y}(x_0; \hat{\sigma}) = w(\hat{\sigma})^T \hat{\gamma}, \quad \hat{Y}(x_0; \hat{\sigma})^o = w(\hat{\sigma})^T \bar{\gamma}^o.
\]

Notice that \( \hat{\sigma} \) does not depend on the bias vector \( \xi(n_s) \), since based on (7),

\[
\hat{\sigma}^2_{n_s}(n_s) = \frac{1}{n-1} \sum_{j=1}^n (Y_j(x_i) - \hat{Y}(x_i))^2 = \frac{1}{n-1} \sum_{j=1}^n (\varepsilon_j(x_i) - \bar{\varepsilon}(x_i))^2.
\]

For the analytical tractability, we impose the following assumption in this subsection.

**Assumption 1** At design point \( x_i \), the simulation errors \( \varepsilon_1(x_i), \varepsilon_2(x_i), \ldots \) are i.i.d. \( \text{N}(0, A_i n_s^{-1}) \), \( i = 1, 2, \ldots, k \).

Notice that the normality assumption on the simulation errors is also used by Ankenman et al. [2010]. The important consequence of this assumption is that the sample variances are independent of \( \bar{\gamma} \) or \( \bar{\gamma}^o \). Utilizing this fact, one can show the following lemma whose proof is deferred to Appendix A.1.

**Lemma 1** Under Assumption 1 on simulation errors, we have the following decomposition of the mean squared error of the predictor \( \hat{Y}(x_0; \hat{\sigma}) \):

\[
\text{MSE} \left( \hat{Y}(x_0; \hat{\sigma}) \right) = \text{MSE} \left( \hat{Y}(x_0; \sigma)^o \right) + \mathbb{E} \left[ (w(\hat{\sigma})^T \xi(n_s))^2 \right] + \mathbb{E} \left[ (\hat{Y}(x_0; \hat{\sigma})^o - \hat{Y}(x_0; \sigma)^o)^2 \right].
\]

(10)

The decomposition above manifests how the bias and the parameter estimation affect the mean squared error of the standard stochastic kriging predictor. Notice that

\[
\hat{Y}(x_0; \hat{\sigma}) - Y(x_0) = \left\{ \hat{Y}(x_0; \sigma)^o - Y(x_0) \right\} + \left\{ \hat{Y}(x_0; \hat{\sigma}) - \hat{Y}(x_0; \sigma)^o \right\} + \left\{ \hat{Y}(x_0; \hat{\sigma})^o - \hat{Y}(x_0; \sigma)^o \right\}.
\]

The terms in the equation above describe the discrepancies due to prediction using the unbiased simulation response estimates at design points, bias, and estimation of the intrinsic variance-covariance matrix, respectively. And (10) is its analogue for the second moments. This expression is not completely new: Except for the contribution from the bias vector (the 2nd term in (10)), such a decomposition appears in describing the relationship between state-1 and state-2 predictors in Zimmerman and Cressie [1992] where the authors analyze the impact of estimating covariance parameters on the MSE of prediction in the classical kriging context. Now, to further facilitate the analysis, we follow a similar approach as used in Kackar and Harville [1984] and Zimmerman and Cressie [1992], i.e., consider the first-order Taylor series expansion of \( \hat{Y}(x_0; \hat{\sigma})^o \) around \( \sigma \).

**Lemma 2** For the last term in (10), we have the following first-order approximation:

\[
\mathbb{E} \left[ (\hat{Y}(x_0; \hat{\sigma})^o - \hat{Y}(x_0; \sigma)^o)^2 \right] \approx \text{Tr} \left[ \left( \frac{\partial w(\sigma)}{\partial \sigma} \right)^T \left( \Sigma + F \beta \beta^T F^T \right) \frac{\partial w(\sigma)}{\partial \sigma} \right] \text{Cov}(\sigma)
\]

where \( \partial w(\sigma)/\partial \sigma \) is the differential of \( w(\sigma) \) such that its \( ij \)th entry is equal to \( \partial w_i(\sigma)/\partial \sigma_{jj} \) and \( \text{Tr}[A] \) denotes the trace of a matrix \( A \).
Finally, we arrive at the approximation to the mean squared error of predictor \( \hat{Y}(x_0; \hat{\sigma}) \):

\[
\text{MSE}\left( \hat{Y}(x_0; \hat{\sigma}) \right) \approx \text{MSE}\left( \hat{Y}(x_0; \sigma)^0 \right) + \mathbb{E}\left( \left( \frac{\partial w(\sigma)}{\partial \sigma} (\Sigma + F\beta F^\top) \frac{\partial w(\sigma)}{\partial \sigma} \right) \text{Cov}(\hat{\sigma}) \right).
\]

(11)

Notice that \( \hat{Y}(x_0; \sigma)^0 \) is unbiased for \( Y(x_0) \) and its MSE is given by \( \eta^\top (F^\top \Sigma^{-1} F)^{-1} \eta + (8) \). And the second term can be shown to be of order \( O(n_s^{-2\alpha}) \). As for the third term, thanks to Assumption 1 and the fact that simulation errors are independent across design points, the matrix \( \text{Cov}(\hat{\sigma}) \) is diagonal and \( (n-1)\hat{\sigma}^2_{i0}/\sigma^2_{i0} \) follows a chi-square distribution with \( n-1 \) degrees of freedom. The latter fact leads to

\[
\text{Cov}(\hat{\sigma}) = \frac{2}{n-1} \text{diag}\left\{ \sigma^4_{10}, \sigma^4_{20}, \ldots, \sigma^4_{k0} \right\} \approx \frac{2}{Nn_s} \text{diag}\left\{ A^2_1, A^2_2, \ldots, A^2_k \right\},
\]

where we used the specification that \( N = n \cdot n_s \).

To further facilitate our analysis, we focus on simplified two-design-point and \( k \)-design-point models with intercept only for the remainder of this subsection. For both models, we assume that the constant mean term is used, i.e., \( f(x)^\top \beta = \beta_0 \). Notice that the assumption on the constant mean actually has been widely used in the kriging literature for its simplicity and good performance in practice. In the two-design-point model, the other parameters are given by

\[
\Sigma_M = \tau^2 \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}, \quad \Sigma_M(x_0, \cdot) = \tau^2 \begin{pmatrix} r_1 \\ r_2 \end{pmatrix}, \quad \text{and} \quad \Sigma_e = \frac{1}{n} \begin{pmatrix} \sigma^2_{10} & 0 \\ 0 & \sigma^2_{20} \end{pmatrix}.
\]

In the \( k \)-design-point model, it is assumed that

\[
\Sigma_M = \tau^2 I_k, \quad \Sigma_M(x_0, \cdot) = \tau^2 (r_0, r_0, \ldots, r_0)^\top, \quad \text{and} \quad \Sigma_e = \frac{1}{n} \text{diag}\left\{ \sigma^2_{01}, \sigma^2_{02}, \ldots, \sigma^2_{0k} \right\},
\]

where \( I_k \) denotes the \( k \times k \) identity matrix. This represents a situation in which the extrinsic correlations among the design points are approximately 0 and the design points are equally correlated with the point we wish to predict, which might be (approximately) plausible if the design points are widely separated, say at the extremes of the region of interest, whereas \( x_0 \) is central.

Given these structures, the optimal budget allocation rule can be identified and it is presented in Proposition 2 below. We refer the reader to Appendix A.1 for the derivation for the two-design-point model. As the analysis of the \( k \)-design-point model leads to results in a similar spirit, for the sake of brevity, we choose not to present its details in the paper.

**Proposition 2** Consider the two-design-point and \( k \)-design-point models with assumed asymptotic bias and covariance structures. Suppose that the spatial parameters are given but \( \beta \) and \( \Sigma_e \) need to be estimated. Given Assumption 1, it is asymptotically MSE-optimal to set \( n^* = O\left(N^{2(\alpha-1)/(2\alpha+1)}\right) \) and \( n^*_s = O\left(N^{3/(2\alpha+1)}\right) \) as \( N \to \infty \).

To develop some intuition about the asymptotic optimal allocation rule given in Proposition 2, consider the special case of the two-design-point model with \( r_1 = r_2 = r, A_1 = A_2 = A, B_1 = B_2 = B \), in which
case we have identical bias and intrinsic variance at the two design points and they are equally correlated to the prediction point. In this simplified scenario, given a fixed computational budget $N$ the rule suggests to assign

$$n^* = \left( \frac{4\alpha\tau^2B^2}{A^2} \right)^{\frac{1}{2(\alpha+1)}} N \frac{1}{(2\alpha+1)}, \quad n^* = \left( \frac{A^2}{4\alpha\tau^2B^2} \right)^{\frac{1}{2(\alpha+1)}} N \frac{2(\alpha+1)}{(2\alpha+1)}. $$

It is now clear that a moderate number of replications $n$ is suitable in contrast with the one given in Proposition 1. This is precisely due to the fact that we need to estimate the intrinsic variance-covariance matrix $\Sigma_e$. The most outstanding implication is that a greater portion of the computational budget should be spent to increase the number of replications if the bias diminishing rate $\alpha$ is large. This is because the influence of the bias vanishes more quickly. The appropriate number for $n$ also seems to depend on the magnitudes of bias constants $B_i$, the intrinsic variance constants $A_i$ as well as the spatial variance $\tau^2$. If the bias constant $B$ is large, then it is better to allocate more to the sample size $n_s$. Meanwhile, if the variation ratio $\tau^2/A^2$ is considerably large, i.e., the spatial variance of the response surface far exceeds the square of the intrinsic variance constant, then it is better to use a larger $n_s$. Conversely, if this ratio is small, then one is better off having more replications to handle the influence from stochastic simulation errors.

Before we move on to the next part, recall that all the results in the analysis are obtained under the assumption that the spatial model parameters are known. To gain a more complete understanding, we relax this assumption and fully investigate the budget allocation rules when we estimate all the parameters in the next section.

### 3.3 Numerical experiments

#### 3.3.1 Quadratic loss of two assets and its risk measurements

In this subsection, we consider a two-dimensional problem borrowed from Hong and Liu [2009]. Our objective is to construct the response surfaces of two popular risk measures, namely, VaR and CVaR. Here, we assume that the random loss $L(\mu)$ is given by $L(\mu) = a_0 + a^\top \Delta S + \Delta S^\top H \Delta S$ where $\mu = (\mu_1, \mu_2)^\top$, and $a_0 = 0.3$, $a = (0.8, 1.5)^\top$, $H = \begin{pmatrix} 1.2 & 0.6 \\ 0.6 & 1.5 \end{pmatrix}$. Further, we assume that the risk factor $\Delta S$ is a bivariate normal random vector with mean $\mu$ and variance-covariance matrix $\Sigma_s = 0.02 \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$.

Writing VaR and CVaR at the level $p$ as $v_p(\mu)$ and $c_p(\mu)$ respectively, we have

$$c_p(\mu) = E \left[ L(\mu) \right| L(\mu) \geq v_p(\mu) ] = v_p(\mu) + \frac{1}{1-p} E \left[ L(\mu) - v_p(\mu) \right]^+. $$

Hong and Liu [2009] suggest using the following consistent estimators: given $n_s$ i.i.d. simulated random losses $L_h(\mu)$ for a given mean vector $\mu$,

$$ \hat{c}_p(\mu) = L_{[n_s\alpha]:n_s}(\mu), \quad \hat{c}_p(\mu) = \frac{1}{n_s(1-p)} \sum_{h=1}^{n_s} \left[ L_h(\mu) - \hat{v}_p(\mu) \right]^+, \quad (12) $$

where $L_{k:n_s}(\mu)$ is the $k$th order statistic of $\{L_h(\mu)\}_{h=1}^{n_s}$.
Experiments. The experimental design space is \( \Omega_{\mathbf{\mu}} = [0.001, 0.1]^2 \) from which we select \( k = 16 \) design points, i.e., choose a set of \( \mathbf{\mu} \) vectors, \( \{\mathbf{\mu}_i\}_{i=1}^k \). As for selecting the 16 design points, a “maxmin” Latin-hypercube sample of 12 design points from \( \Omega_{\mathbf{\mu}} \) plus its four corner points are used. At each design point we run \( n \) simulation replications and hence obtain \( n \) simulation estimates of VaR and CVaR. Notice that to obtain each such pair of estimates through (12), we need to simulate a sample of \( n_s \) random losses, \( \{ L_h(\mathbf{\mu}_i) \}_{h=1}^{n_s} \). One decision to be made by the experimenter is how to allocate the total computational budget \( N \) at each design point, between the number of replications \( n \) and the sample size \( n_s \) so that \( N = n \cdot n_s \). Different budget allocation rules are expected to lead to different performances of the standard stochastic kriging predictor.

We choose in total 1601 check-points for evaluating the performance of the stochastic kriging predictor under different budget allocation rules. In fact, the 1600 check-points are regularly spaced in \( \Omega_{\mathbf{\mu}} \) and we add the extra testing point from Hong and Liu [2009] for a sanity check. As the closed-form expressions of \( v_p(\mathbf{\mu}) \) and \( c_p(\mathbf{\mu}) \) are not available, the true values at each of the check-points are approximated by simulation with a sample size \( n_s = 10^6 \). Experiments for \( p \in \{0.95, 0.99\} \) are conducted, and since the results convey similar information only the ones for \( p = 0.99 \) are presented below. We use the estimated root mean squared error (ERMSE) over the 1601 check-points as the performance measure:

\[
\text{ERMSE}(\tilde{v}_p) = \sqrt{\frac{1}{1601} \sum_{i=1}^{1601} (v_p(\mathbf{\mu}_i) - \tilde{v}_p(\mathbf{\mu}_i))^2}, \quad \text{ERMSE}(\tilde{c}_p) = \sqrt{\frac{1}{1601} \sum_{i=1}^{1601} (c_p(\mathbf{\mu}_i) - \tilde{c}_p(\mathbf{\mu}_i))^2},
\]

where \( v_p(\mathbf{\mu}_i), c_p(\mathbf{\mu}_i) \) are true VaR, CVaR and \( \tilde{v}_p(\mathbf{\mu}_i), \tilde{c}_p(\mathbf{\mu}_i) \) are their predicted values at \( \mathbf{\mu}_i \) by stochastic kriging. We are interested in investigating the roles played by \( n_s \) and \( n \) on the performance of the stochastic kriging predictor given a fixed computational budget \( N \).

Results. The entire experiment is repeated for 100 macro-replications and the resulting ERMSEs for \( \tilde{c}_p(\mathbf{\mu}) \) are summarized in Figure 1. Results for \( \tilde{v}_p(\mathbf{\mu}) \) are similar in nature and hence are omitted. As explained in Hong and Liu [2009] and Trindade et al. [2007], the estimator \( \tilde{c}_p(\mathbf{\mu}) \) has an asymptotic variance of the form \( \text{Var} \left( \tilde{c}_p - E[\tilde{c}_p] \right) \approx A_p n_s^{-1} \) with \( A_p \) being some constant, which corresponds to the asymptotic variance of the simulation errors, say \( \varepsilon(\mathbf{\mu}_i; n_s) \) as in (7). Also, Trindade et al. [2007] proves that the asymptotic bias of \( \tilde{c}_p(\mathbf{\mu}) \) has an order of \( O(n_s^{-1}) \). In our setting, this means \( \alpha = 1 \). Hence, Proposition 2 implies that \( n = O(1) \) is optimal at least for the simplified two-design-point and \( k \)-design-point models analyzed.

Figure 1 presents the ERMSEs of the stochastic kriging predictor when the total simulation budget \( N \) at each design point increases from \( 10^4 \) to \( 10^5 \) in panels (a) to (c). Each panel contains the boxplots that summarize the 100 ERMSEs achieved under different budget allocation rules \( (n, n_s) \) with \( n_s = \lceil N/n \rceil \) as a function of \( n \). We observe that the average magnitude of ERMSEs decreases as the budget \( N \) increases, as expected. More importantly, the ERMSEs seem to get minimized around \( n = 25 \) regardless of the size of the budget \( N \). This agrees with the result implied by Proposition 2 even though we cannot determine the optimal value of \( n \) to use. At least, Proposition 2 shows that a moderate number of replications \( n \) works sufficiently well; there is no need to increase \( n \) even when a larger budget becomes available.

3.3.2 A steady-state simulation with initialization bias in an M/M/1 queue

Let \( y(x) \) represent the steady-state expected waiting time of customers with arrival rate 1 and service rate \( x \), \( x > 1 \). It is well known that the steady-state waiting time is exponential with mean \( 1/(x - 1) \) conditional
on it being positive. Hence it can be calculated that \( y(x) = 1/(x(x - 1)) \), which gives the surface we are trying to approximate through stochastic kriging. We use this example to evaluate the performance of stochastic kriging predictor as in Staum [2009]. While each simulation run is initialized in steady state in Staum [2009], we start with a completely empty and idle system and this is the source of bias in the estimates of steady-state expected waiting time. More specifically, each replication estimate of \( y(x) \) at a given \( x \) is achieved by simulating \( n_s \) customers from an empty system and computing the average of their simulated waiting times, which are known to be dependent.

**Experiments.** We consider the design space of service rate to be \( \Omega_x = [1.5, 2] \). Ten equispaced design points are selected from the design space, say \( \{x_i\}_{i=1}^k \) with \( k = 10 \). At a design point \( x \), a single simulation run is initiated with system empty and is terminated upon the departure of the \( n_s \)th customer. That is, obtaining each response estimate \( Y_j(x) \) involves generating a sample path composed of \( n_s \) customers. Notice that, unlike a typical steady-state simulation in which a single long sample path is simulated, to implement stochastic kriging technique we need to make \( n \) \((n \geq 2)\) simulation replications and the final output is the set of response estimates \( \{Y_j(x)\}_{j=1}^n \). Therefore, given a fixed total number of customers \( N \) to simulate at \( x \), one needs to decide how to allocate between \( n \) and \( n_s \) for a better performance of the stochastic kriging predictor. We study the impact of such allocation rules through comparing their corresponding ERMSEs. With 1000 regularly spaced check-points selected in \( \Omega_x \), the ERMSE is defined as

\[
\text{ERMSE}(\hat{y}) = \sqrt{\frac{1}{1000} \sum_{l=1}^{1000} (y(x_l) - \hat{y}(x_l))^2},
\]

Figure 1: Boxplots of ERMSEs for \( \hat{c}_p(\mu) \) over 100 macro-replications with \( k = 16 \) for \( p = 0.99 \).
where \( y(x_l) \) and \( \hat{y}(x_l) \) represent the true and the predicted steady-state expected waiting time respectively with service rate \( x_l \). Recall that the true response surface is known in closed form.

Results. The entire experiment is repeated for 100 macro-replications and the resulting ERMSEs are summarized in Figure 2. The four panels (a) to (d) present boxplots of ERMSEs with a simulation budget \( N \) ranges from \( 10^4 \) to \( 10^7 \). Each boxplot in a panel summarizes 100 ERMSEs obtained with a given number of simulation replications \( n \), subject to a fixed simulation budget \( N \). Hence the boxplots in each panel illustrate the effectiveness of different allocation rules \((n, n_s)\) with \( n_s = \lceil N/n \rceil \). As we expect, the average magnitude of the ERMSEs decreases as \( N \) increases. For a fixed \( N \), on the other hand, the minimal ERMSE seems to be achieved with a number of replications between \( n = 10 \) and \( n = 50 \) in all four panels. (In fact, it is difficult to distinguish performances when varying \( n \) from 10 up to 200 or 400 in panels (c) and (d).) The results demonstrate that a moderate number of replications \( n \) would work sufficiently for any size of \( N \) as long as the computational budget is sufficiently large.

Whitt [2006] shows the asymptotic bias of the estimator formed by averaging the \( n_s \) simulated waiting times is of order \( O(n_s^{-1}) \), that is, \( \alpha = 1 \) in our setting. Similarly, the asymptotic variance of simulation errors satisfies our assumption that \( \text{Var}(\varepsilon(x_i; n_s)) \approx A_i n_s^{-1} \). It follows from Proposition 2 that it is optimal to use \( n = O(1) \) at least for the simplified two-design-point and \( k \)-design-point models analyzed. This experiment confirms that even though that all the metamodel parameters need to be estimated the optimal allocation rule given by Proposition 2 still applies.

![Figure 2: Boxplots of ERMSEs for \( \hat{y} \) over 100 macro-replications with \( k = 10 \).](image-url)
Lastly, it is worthwhile mentioning that the results given in Proposition 2 are in a similar spirit to those given in Kelton [1986] who investigates the best way of splitting a fixed amount of simulation effort across replications for simulating stationary and certain non-stationary discrete-time stochastic processes, in the presence of positive autocorrelation. Kelton [1986] shows that shorter replications are preferable to reduce variances of the output estimators of interest, as long as the simulation budget is not tight. The author, however, further suggests to use a moderate number of replications, which may also help mitigate potential bias problems.

4 Extension to Two-Level Simulations

4.1 The extended model for two-level simulations

Our goal in this section is to do prediction using a stochastic kriging metamodel built on the simulation response estimates obtained in a two-level simulation experiment. Let us go back to the example of quantile estimation of a random variable \( L(x_0) \). However, this time obtaining one realization of a random variable \( L(x_i) \) at a design point \( x_i \) is more involved. Following Gordy and Juneja [2010], we give a brief description of a two-level simulation procedure below.

Suppose that we have a portfolio of some financial contracts at the current time 0 and those contracts expire at time \( T \). Let \( V(t; x) \) be the value of the portfolio at time \( t \) and set \( L(x_0) = V(T_0; x) \) with \( T_0 < T \). Here, \( x \) can be related to portfolio weights or market conditions, for instance. We are then concerned with the possible degradation of the portfolio value at the horizon \( T_0 \). However, oftentimes a portfolio value can only be simulated due to its complex structures or payoffs. In that case, we first generate a number of outer level simulated scenarios, say \( \omega_h \) for \( h = 1, \ldots, n_{\text{out}} \), to simulate market conditions at \( T_0 \), and then approximate a portfolio value \( V(T_0; x, \omega_h) \) by a number of inner level simulation runs up to time \( T \) for each \( h \), say \( \zeta_{hq} \) for \( q = 1, \ldots, n_{\text{in}} \). If we denote the zero-mean random error incurred on the \( q \)th inner level simulation run by \( \eta_{hq} \), then we can express the simulated estimate as

\[
\tilde{V}(T_0; x, \omega_h) = V(T_0; x, \omega_h) + \frac{1}{n_{\text{in}}} \sum_{q=1}^{n_{\text{in}}} \eta_{hq}.
\]

In summary, we need \( n_{\text{in}} \) inner level simulation runs to generate one realized (approximate) portfolio value; moreover, \( n_{\text{out}} \) such realized portfolio values are required to produce one simulation output \( Y_j(x_i) \) at design point \( x_i \) on the \( j \)th replication.

Notice that there exists bias in estimating quantiles of \( L(x_i) \) even if we can approximate \( V(T_0; x_i, \omega_h) \) without any bias. And it turns out that this bias is asymptotically of order \( O(n_{\text{in}}^{-1}) + O(n_{\text{out}}^{-1}) \) as shown in Gordy and Juneja [2010]. The variance of the simulation-based quantile estimators, on the other hand, is asymptotically of order \( O(n_{\text{out}}^{-1}) \). Therefore, we modify the metamodel given in (7) to the following form to describe a simulation response estimate obtained on a two-level simulation replication:

\[
Y_j(x_i) = f(x_i) + M(x_i) + \varepsilon_j(x_i; n_{\text{out}}) + \xi(x_i; n_{\text{in}}, n_{\text{out}}), \quad j = 1, 2, \ldots, n, \quad i = 1, 2, \ldots, k.
\]

In (13) above, we model the bias incurred at design point \( x_i \) as \( \xi(x_i; n_{\text{in}}, n_{\text{out}}) \) which is a deterministic quantity whose magnitude depends on the number of outer level scenarios \( n_{\text{out}} \) and the number of inner level simulation runs \( n_{\text{in}} \) used in a simulation replication.
As for the variance of the difference, recall that the k-design-point model. We refer the reader to Appendix A.2 for details. As the analysis of the intrinsic variance-covariance matrix is unknown is given by

$$\text{MSE} \left( \hat{Y}(x_0) \right) = \mathbb{E} \left[ (\hat{Y}(x_0) - Y(x_0))^2 \right] = \left( \mathbb{E} \left[ \hat{Y}(x_0) - Y(x_0) \right] \right)^2 + \text{Var} \left( \hat{Y}(x_0) - Y(x_0) \right).$$

The bias term follows as

$$\mathbb{E} \left[ \hat{Y}(x_0) - Y(x_0) \right] = \mathbf{w}^\top \xi(n_{in}, n_{out}) = \Sigma_M(x_0, \cdot)^\top \Sigma^{-1} \xi(n_{in}, n_{out}) + \eta^\top \mathbf{W} \xi(n_{in}, n_{out}),$$

where

$$\mathbf{w}^\top = \Sigma_M(x_0, \cdot)^\top \Sigma^{-1} + \eta^\top \mathbf{W}, \quad \eta = f(x_0) - \mathbf{F}^\top \Sigma^{-1} \Sigma_M(x_0, \cdot)$$

and

$$\mathbf{W} = (\mathbf{F}^\top \Sigma^{-1} \mathbf{F})^{-1} \mathbf{F}^\top \Sigma^{-1}.$$

As for the variance of the difference, recall that

$$\text{Var} \left( \hat{Y}(x_0) - Y(x_0) \right) = \Sigma_M(x_0, x_0) - \Sigma_M(x_0, \cdot)^\top \Sigma^{-1} \Sigma_M(x_0, \cdot) + \eta^\top (\mathbf{F}^\top \Sigma^{-1} \mathbf{F})^{-1} \eta.$$

We are now ready to analyze the two-design-point model and the k-design-point model as specified in Section 3.2 with known $\Sigma_x$. Some algebraic manipulations lead us to the following results for the two-design-point model. We refer the reader to Appendix A.2 for details. As the analysis of the k-design-point model

4.2 Analysis of the extended model for two-level simulations

4.2.1 Intrinsic variance-covariance matrix $\Sigma_x$ known

Recall from Section 3 that the standard stochastic kriging predictor $\hat{Y}(x_0)$ when $\beta$ is unknown is given by (4) and that the estimator of $\beta$ is given by (5). The associated MSE can be obtained by replacing $\xi(n_{in})$ in (9) with $\xi(n_{in}, n_{out})$. We decompose it into two parts and analyze them separately,

$$\text{MSE} \left( \hat{Y}(x_0) \right) = \mathbb{E} \left[ (\hat{Y}(x_0) - Y(x_0))^2 \right] = \left( \mathbb{E} \left[ \hat{Y}(x_0) - Y(x_0) \right] \right)^2 + \text{Var} \left( \hat{Y}(x_0) - Y(x_0) \right).$$

The bias term follows as

$$\mathbb{E} \left[ \hat{Y}(x_0) - Y(x_0) \right] = \mathbf{w}^\top \xi(n_{in}, n_{out}) = \Sigma_M(x_0, \cdot)^\top \Sigma^{-1} \xi(n_{in}, n_{out}) + \eta^\top \mathbf{W} \xi(n_{in}, n_{out}),$$

where

$$\mathbf{w}^\top = \Sigma_M(x_0, \cdot)^\top \Sigma^{-1} + \eta^\top \mathbf{W}, \quad \eta = f(x_0) - \mathbf{F}^\top \Sigma^{-1} \Sigma_M(x_0, \cdot)$$

and

$$\mathbf{W} = (\mathbf{F}^\top \Sigma^{-1} \mathbf{F})^{-1} \mathbf{F}^\top \Sigma^{-1}.$$

As for the variance of the difference, recall that

$$\text{Var} \left( \hat{Y}(x_0) - Y(x_0) \right) = \Sigma_M(x_0, x_0) - \Sigma_M(x_0, \cdot)^\top \Sigma^{-1} \Sigma_M(x_0, \cdot) + \eta^\top (\mathbf{F}^\top \Sigma^{-1} \mathbf{F})^{-1} \eta.$$

We are now ready to analyze the two-design-point model and the k-design-point model as specified in Section 3.2 with known $\Sigma_x$. Some algebraic manipulations lead us to the following results for the two-design-point model. We refer the reader to Appendix A.2 for details. As the analysis of the k-design-point
model leads to a similar conclusion, we omit its analysis for the sake of brevity.

\[
E \left[ \tilde{Y}(x_0) - Y(x_0) \right] = \frac{1}{n_{\text{in}}} \times \frac{\tau^2 \nu_1 + \nu_2/n_{\text{tot}}}{\Pi(n_{\text{tot}})} + \frac{1}{n_{\text{out}}} \times \frac{\tau^2 \nu_3 + \nu_4/n_{\text{tot}}}{\Pi(n_{\text{tot}})},
\]

(14)

\[
\text{Var} \left[ \tilde{Y}(x_0) - Y(x_0) \right] = \tau^2 - \tau^4 \times \frac{\tau^2 \nu_5 + \nu_6/n_{\text{tot}}}{\Gamma(n_{\text{tot}})} + \frac{(\tau^4 \nu_7 + \tau^2 \nu_8/n_{\text{tot}} + A_1 A_2/n_{\text{tot}}^2)^2}{\Gamma(n_{\text{tot}}) \Pi(n_{\text{tot}})},
\]

(15)

where

\[
\Pi(n_{\text{tot}}) = 2\tau^2(1 - r_{12}) + \frac{A_1 + A_2}{n_{\text{tot}}}, \quad \Gamma(n_{\text{tot}}) = \tau^4(1 - r_{12}) + \frac{\tau^2(A_1 + A_2)}{n_{\text{tot}}} + \frac{A_1 A_2}{n_{\text{tot}}^2},
\]

and \(\nu_i\)'s are some constants specified in Appendix A.2. The variance term involves \(n_{\text{tot}}\) only, while the bias term has both \(n_{\text{in}}\) and \(n_{\text{out}}\). Notice that \(n_{\text{in}} = N/n_{\text{tot}}\), therefore we see from (14) that as long as \(n_{\text{tot}}\) is fixed, the optimal \(n_{\text{out}}\) follows easily.

As the computational budget \(N\) increases, observe that the bias term converges to zero as long as \(n_{\text{in}}, n_{\text{out}}\) increase correspondingly. However, unlike the case where we obtain a point estimate through simulation, the variance term does not converge to zero but rather to some constant as the total number of outer level scenarios \(n_{\text{tot}}\) increases. The reason is that the predictor \(\tilde{Y}(x_0)\) is constructed based on simulation responses at \(k\) design points other than at the prediction point \(x_0\) itself. Let \(\text{MSE}_\infty(\tilde{Y}(x_0))\) denote the limiting MSE as \(n_{\text{tot}} \to \infty\), which is equivalent to the MSE that would have been obtained in the classical kriging context if bias and observation noise were both absent:

\[
\text{MSE}_\infty(\tilde{Y}(x_0)) = \tau^2 - \frac{\tau^2 \nu_5}{1 - r_{12}^2} + \frac{\tau^2(1 + r_{12} - r_1 - r_2)^2}{2(1 + r_{12})}.
\]

The next step is to find the leading terms for the bias and the variance, which are useful in deriving an optimal allocation rule.

**Lemma 3** For an allocation rule \((n, n_{\text{out}}, n_{\text{in}})\) to be effective, \(n_{\text{tot}}\) should be an increasing function of the simulation budget \(N\). In such a case, (14) and (15) can be approximated by

\[
\text{Bias} \approx \frac{1}{2(1 - r_{12})} \left( \frac{\nu_1}{n_{\text{in}}} + \frac{\nu_3}{n_{\text{out}}} \right),
\]

(16)

\[
\text{Variance} \approx \text{MSE}_\infty(\tilde{Y}(x_0)) + \frac{\nu_5(A_1 + A_2) - \nu_6(1 - r_{12}^2)}{(1 - r_{12}^2)^2 n_{\text{tot}}} + \frac{\nu_7}{4(1 - r_{12}^2)^2 n_{\text{tot}}} \left\{ 4(1 + r_{12}) \nu_8 - (A_1 + A_2)(1 + r_{12} - r_1 - r_2)(3 + r_{12}) \right\}.
\]

(17)

For the remainder of this subsection, we exclude the case where \(\nu_1 \nu_3 < 0\), that is, we only consider the scenario in which the bias moves in the same direction as \(n_{\text{in}}\) or \(n_{\text{out}}\) changes. This is further detailed in Remark 1 below. Then based on Lemma 3 we arrive at an asymptotically optimal budget allocation rule as follows.

**Proposition 3** Consider the two-design-point and \(k\)-design-point models with assumed asymptotic bias and covariance structures. Suppose that \(\Sigma_e\) is known and the spatial parameters are given, but \(\beta\) needs to be estimated. If \(\nu_1 \nu_3 > 0\), then it is MSE-optimal to make just one replication \((n = 1)\) and assign \(n_{\text{out}} = O(N^{2/3}), n_{\text{in}} = O(N^{1/3})\).
Before ending this subsection, we consider the special case of $r_1 = r_2 = r$, $A_1 = A_2 = A$, $B_1 = B_2 = B$, and $C_1 = C_2 = 0$ to develop some intuitive understanding of the results obtained above. That is, at the two design points the simulation response estimators have identical bias and variance, and the two design points share an equal spatial correlation with the prediction point. Furthermore, the bias only depends on $n_{\text{in}}$ but not $n_{\text{out}}$. One can readily verify that the asymptotic MSE of the stochastic kriging predictor minus the constant limit equals
\[
\text{MSE} \left( \tilde{Y}(x_0) \right) - \text{MSE}_\infty \left( \tilde{Y}(x_0) \right) = \frac{B^2 n_{\text{tot}}^2}{4N^2} + \frac{A}{2n_{\text{tot}}},
\]
and the resulting optimal allocation rule follows as
\[
n_{\text{tot}}^* = \left( \frac{A}{B^2} \right)^{\frac{1}{3}} N^{\frac{2}{3}}, \quad n_{\text{in}}^* = \left( \frac{B^2}{A} \right)^{\frac{1}{3}} N^{\frac{1}{3}}.
\]
The coefficients for optimal $n_{\text{tot}}^* = n_{\text{out}} \cdot n$ and $n_{\text{in}}^*$ suggest to increase $n_{\text{tot}}^*$ if the ratio of the intrinsic variance constant $A$ to the square of the bias constant $B$ is large. On the other hand, if this ratio is small then we should invest more on the inner level simulation runs to reduce the impact of bias. Lastly, we mention that the spatial variance $\tau^2$ does not play a role in determining the optimal allocation of the simulation budget $N$ at all in this case.

**Remark 1** As long as the spatial correlations $r_1, r_2$ are similar and $B_i, C_i$ are nonzero, we have positive $\nu_1, \nu_3$ and the result of the proposition applies. This is in fact the typical situation we have in mind for a two-design-point model. However, if $\nu_1 \nu_3 < 0$ as a result of somewhat unnatural combination of spatial correlations and $B_i, C_i$ values, then one can make the first order term in (16) vanish by setting $\nu_1 n_{\text{out}} = -\nu_3 n_{\text{in}}$ and get a bias of order $O(N^{-1})$. Indeed, one can then show that the optimal allocation rule in that case is to set $n = O(N)$, $n_{\text{out}} = O(1)$, and $n_{\text{in}} = O(1)$ for the convergence rate of the MSE equal to $O(N^{-1})$. However, we do not further delve into this rather pathological case.

### 4.2.2 Intrinsic variance-covariance matrix $\Sigma_e$ unknown

In this section, we extend the analysis conducted so far and discuss the impact of estimating the intrinsic variance-covariance matrix on the MSE of prediction achieved by $\tilde{Y}(x_0)$ when $\beta$ is unknown in the two-level simulation context. We focus on studying the two-design-point and $k$-design-point models but the analysis hereinafter is presented for the two-design-point model only. We impose the following assumption similar to Assumption 1 in Section 3.2.2.

**Assumption 3** At design point $x_i$, the simulation errors $\varepsilon_1(x_i), \varepsilon_2(x_i), \ldots$ are i.i.d. $N(0, A_i n_{\text{out}}^{-1}), i = 1, 2, \ldots, k$.

We notice that Lemmas 1 and 2 still apply. Therefore, we get the following approximation to the MSE of predictor $\tilde{Y}(x_0; \hat{\sigma})$:
\[
\text{MSE} \left( \tilde{Y}(x_0; \hat{\sigma}) \right) \approx \text{MSE} \left( \tilde{Y}(x_0; \sigma)^o \right) + \mathbb{E} \left[ \left( \mathbf{w}(\hat{\sigma})^\top \mathbf{\xi}(n_{\text{in}}, n_{\text{out}}) \right)^2 \right] \\
+ \text{Tr} \left[ \left( \frac{\partial \mathbf{w}(\sigma)}{\partial \sigma}^\top \left( \mathbf{\Sigma} + \mathbf{F} \beta \mathbf{F}^\top \right) \frac{\partial \mathbf{w}(\sigma)}{\partial \sigma} \right) \text{Cov}(\hat{\sigma}) \right].
\tag{18}
\]
This approximation explains how the MSE is affected by the bias and by the estimation of \( \Sigma_e \). The first term is the MSE of the unbiased stochastic kriging predictor, which can be further decomposed into the sum of the unbiased stochastic kriging predictor with known \( \beta \) and the MSE inflation due to estimating \( \beta \). Possible errors due to bias and parameter estimation naturally lead to further inflation of the MSE, and they constitute the second and the third terms in (18). It was the essence of Lemmas 1 and 2 that such effects are orthogonal in the mean square sense.

The computation of the MSE can be briefly described as follows. In Section 4.2.1, we decompose MSE \( \left( \hat{Y}(x_0) \right) \) into the bias effect and the variance of difference. Hence, without any bias, the MSE of the unbiased stochastic predictor with known \( \Sigma_e \) is just \( \text{Var} \left( \hat{Y}(x_0) - Y(x_0) \right) \). Therefore, the first term of (18) equals (17), which we denote by \( \Theta_1 n_{\text{tot}}^{-1} \). The third term does not involve \( n_{\text{in}} \) and therefore the arguments in Appendix A.1 lead to the same result except that we replace \( n_{\text{tot}}, N \) with \( n_{\text{out}}, n_{\text{tot}} \), respectively; we denote the approximation to the third term by \( \Theta_3 n_{\text{out}} n_{\text{tot}}^{-3} \). Lastly, the approximation to the second term follows as \( \Theta_2 (\nu_1 n_{\text{in}}^{-1} + \nu_3 n_{\text{out}}^{-1})^2 \). See Appendix A.3 for additional details. Again, for the remainder of this subsection we exclude the pathological case of \( \nu_1 \nu_3 < 0 \). In fact, it turns out to be beneficial to consider a more general model for bias, say
\[
\xi(x_i; n_{\text{in}}, n_{\text{out}}) = \frac{B_i}{n_{\text{in}}} + \frac{C_i}{n_{\text{out}}}
\]
for some constant \( \gamma \in [1, \infty) \). With this specific bias form, we present the main result of this subsection.

**Proposition 4** Consider the two-design-point and \( k \)-design-point models with assumed asymptotic bias and covariance structures. Suppose that the spatial parameters are given, but \( \beta \) and \( \Sigma_e \) need to be estimated. If \( \nu_1 \nu_3 > 0 \) and \( \gamma > 1.5 \), then it is MSE-optimal to assign \( n = O \left( N^{(2\gamma - 3)/(3\gamma + 3)} \right) \), \( n_{\text{out}} = O \left( N^{5/(3\gamma + 3)} \right) \), and \( n_{\text{in}} = O \left( N^{1/3} \right) \). If \( \nu_1 \nu_3 > 0 \) but \( \gamma \leq 1.5 \), then it is optimal to assign \( n = O(1) \), \( n_{\text{out}} = O \left( N^{2/3} \right) \), and \( n_{\text{in}} = O \left( N^{1/3} \right) \).

The reader is referred to Appendix A.3 for a proof. It is worth noting that one should set \( n^*_\text{in} = O \left( N^{1/3} \right) \) given a fixed simulation budget \( N \), even though the optimal allocation of \( n_{\text{tot}} \) to \( n_{\text{out}} \) and \( n \) is not as clear-cut from the results given by Proposition 4. Nevertheless, it becomes clear later that a moderate number of replications suffices to ensure good performance of the standard stochastic kriging predictor. We further notice that the effects of two different allocations for \( (n, n_{\text{out}}) \) in Proposition 4 are of lower order. That is, the asymptotic behavior of the MSEs for both allocation rules are dominated by the first term in (18), which is of order \( O(N^{-2/3}) \) as \( N \) becomes sufficiently large. The remaining two terms related to bias and estimation effect vanish faster, and they are affected by how we allocate \( n_{\text{tot}} \) between \( n \) and \( n_{\text{out}} \).

Lastly we note that the number of replications has to be greater than 1, because an actual implementation of stochastic kriging technique involves estimating \( \Sigma_e \) by the sample covariance matrix and hence a moderate number of simulation replications is needed. The results given in Proposition 4 should be interpreted as in some situations (e.g., when \( \gamma = 1 \)) we do not want to increase \( n \) even if a greater simulation budget \( N \) becomes available.

### 4.3 Numerical experiments: derivative portfolio on a single asset

The topic of optimal sampling rules for nested or two-level simulation has been explored by several researchers. See, for instance, Gordy and Juneja [2010] for pointwise estimation via simulation and Broadie
et al. [2011] through a regression approach. In this example, we assess the impact of budget allocation rules for \((n, n_{out}, n_{in})\) on the performance of the standard stochastic kriging predictor. A particular attention is paid to how the insights obtained from the previous subsections carry over to the case in which all model parameters are estimated.

Consider a portfolio consisting of three derivatives on a single asset with initial price \(S_0\):

- short one call option with payoff \(H_1(S) = -(S_T - K_c)^+\) with \(K_c = 101\).
- long one put option with payoff \(H_2(S) = (K_p - S_T)^+\) with \(K_p = 110\).
- short one Asian call option with payoff \(H_3(S) = -\left(M^{-1} \sum_{m=1}^{M} S_{t_m} - K_a\right)^+\) with \(0 \leq t_1 < \cdots < t_M \leq T\) with \(K_a = 110\).

Here, \(S_t\) denotes the stock price at time \(t\), \(T\) is the maturity of the three derivative contracts, and \(K\) denotes a strike of an option. All these are pre-specified in an option contract together with a set of time points \(\{t_m\}_{m=1}^{M}\) in an Asian option. We are interested in evaluating an extreme left quantile \(q_p\) of the value of this portfolio at a particular horizon \(t_\ast\) as a function of \(S_0\), say \(V(S_0)\); it gives the threshold below which \(V(S_0)\) falls with probability \(p\), for example, \(p = 0.01\). Mathematically, the portfolio value is equal to

\[
V(S_0) = \sum_{i=1}^{3} \mathbb{E}\left[ e^{-r(T-t_\ast)} H_i(S) \mid \mathcal{F}_{t_\ast} \right] = e^{-r(T-t_\ast)} \left\{ -\mathbb{E}\left[ (S_T - K_c)^+ \mid S_{t_\ast} \right] + \mathbb{E}\left[ (K_p - S_T)^+ \mid S_{t_\ast} \right] ight. \\
- \mathbb{E}\left[ \left( \frac{1}{M} \sum_{m=1}^{a} S_{t_m} + \frac{1}{M} \sum_{m=a+1}^{M} S_{t_m} - K_a \right)^+ \mid S_{t_1}, \ldots, S_{t_a}, S_{t_\ast} \right] \right\},
\]

(19)

where \(a\) is an integer such that \(t_1 < \cdots < t_a \leq t_\ast < t_{a+1} < \cdots < t_M \leq T\); \(r\) is the risk-free rate and \(\mathcal{F}_{t_\ast}\) is the information set available at time \(t_\ast\) that includes the history of the stock price up to time \(t_\ast\). The evaluation of quantiles of \(V_{t_\ast}(S_0)\) is done by simulating portfolio values at \(t_\ast\) for many different price paths \(\{S_t : 0 \leq t \leq t_\ast\}\).

We notice that the mathematical expectation in (19) is taken under the so called risk-neutral measure. On the other hand, price paths \(\{S_t : 0 \leq t \leq t_\ast\}\) need to be generated under the real world measure. In this example, the dynamics of the stock price under these two measures are modeled as geometric Brownian motion and given by

\[
S_t = S_0 \exp \left( (\mu - \sigma^2/2) t + \sigma W_t \right), \quad t \leq t_\ast, \\
S_t = S_{t_\ast} \exp \left( (r - \sigma^2/2) (t - t_\ast) + \sigma B_{t-t_\ast} \right), \quad t > t_\ast,
\]

where \(\mu\) is the real world drift, and \(W\) and \(B\) are two independent standard Brownian motions. The parameter values are given in Table 1. Lastly, the observation times for the Asian call option are specified as follows in years:

\[
\{t_m\}_{m=1}^{6} = \{0.0050, 0.0192, 0.0242, 0.0439, 0.0636, 0.0833\}.
\]
Experiment details. Let us describe the procedure for the two-level simulation in this example. First of all, the design space is $\Omega_{S_0} = [80, 120]$ from which $k = 13$ equispaced design points are selected, say $\{S_0^{(i)}\}_{i=1}^k$. Secondly, for the outer level simulation, we generate $n_{\text{out}}$ outer level scenarios up to the risk horizon $t_*$, say $\omega^1, \ldots, \omega^{n_{\text{out}}}$ which represent simulated Brownian paths of $W$. Conditional on each outer level scenario $\omega^h$, we generate $n_{\text{in}}$ independent inner level sample paths, $\zeta^{h1}, \ldots, \zeta^{hn_{\text{in}}}$, i.e., $n_{\text{in}}$ simulated sample paths of $B$; these inner level sample paths determine the cash flows of the portfolio from time $t_*$ to $T$. Given $\omega^h$ and $\{\zeta^{hq}\}_{q=1}^{n_{\text{in}}}$, we obtain the portfolio value estimate, say $V_h(S_0^{(i)})$, under outer level scenario $\omega^h$ at the design point $S_0^{(i)}$. A single estimate of $p$-quantile at $S_0^{(i)}$ can be found by looking at the $[p \cdot n_{\text{out}}]$th order statistic of $\{V_h(S_0^{(i)})\}_{h=1}^{n_{\text{out}}}$. We run the procedure above for $n$ simulation replications to obtain simulation response estimates $\hat{V}_j(S_0^{(i)})$ at design point $S_0^{(i)}$, $i = 1, 2, \ldots, k$, and subsequently build a stochastic kriging metamodel and do prediction.

It is not hard to tell that the aforementioned two-level simulation procedure can be time-consuming to implement, hence the task of devising an efficient budget allocation rule for $(n, n_{\text{out}}, n_{\text{in}})$ becomes very important given a fixed simulation budget $N = n \cdot n_{\text{out}} \cdot n_{\text{in}}$ to expend at each design point. To evaluate the performance of the standard stochastic kriging predictor under different budget allocation rules, we use ERMSE as defined in Section 3.3. Here the ERMSE is calculated based on the predicted results at 193 equispaced check-points in $\Omega_{S_0}$. Since a closed-form expression for quantiles of $V(S_0)$ is not available, the true values at the check-points are approximated using $n_{\text{out}} = 10^3$ and $n_{\text{in}} = 500$ with the value of $p$ set to 0.01.

Results. The entire experiment is repeated for 100 macro-replications and the resulting ERMSEs with varying computational budgets are summarized in Figure 3. For a fixed computational budget $N$, we compare the prediction results of the standard stochastic kriging predictor under different budget allocation rules and the estimation results from a naive simulation method. The budget allocation rules tested include those resulting from Proposition 4 by varying the number of replications $n$ as well as the 2/3 : 1/3 rule recommended by Gordy and Juneja [2010].

Specifically, given a fixed simulation budget $N$, an allocation rule of $(n_{\text{tot}}, n_{\text{in}}) = ([N^{2/3}], [N^{1/3}])$ is made following Proposition 4. Then given the resulting $n_{\text{tot}}$, we increase the number of simulation replications $n$ from 2 to 50 to explore the trade-off between $n$ and $n_{\text{out}}$. In contrast, Gordy and Juneja [2010] propose the 2/3 : 1/3 rule for two-level simulations that assigns $(n_{\text{out}}, n_{\text{in}}) = ([N^{2/3}], [N^{1/3}])$ to obtain a pointwise quantile estimate through simulation. We employ this rule for implementing stochastic kriging by first choosing a moderate number of replications $n$ (here $n = 10$ seems to give the best results) and further assigning the remaining budget to $n_{\text{out}}$ and $n_{\text{in}}$ according to the 2/3 : 1/3 rule, i.e., $(n_{\text{out}}, n_{\text{in}}) = ([((N/n)^{2/3}], [(N/n)^{1/3}])$. The naive simulation method is to simply apply simulations to estimate quantiles at the check-points directly. We divide the total budget $kN$ equally among all the check-points and obtain direct simulated quantile estimates at each one using Gordy and Juneja [2010]'s 2/3 : 1/3 rule. Notice that for this estimation method one replication is sufficient, for there does not involve model fitting or prediction. The results obtained by the naive method can be used as a benchmark for comparison.

Table 1: Parameter values for the derivative portfolio on a single asset

<table>
<thead>
<tr>
<th>$T$</th>
<th>$t_*$</th>
<th>$\mu$</th>
<th>$r$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0833 year</td>
<td>0.0192 year</td>
<td>8%</td>
<td>3%</td>
<td>20%</td>
</tr>
</tbody>
</table>
The first thing we notice from Figure 3 is the effectiveness of stochastic kriging for prediction with a fixed computational budget. In all panels, regardless of the number of replications \( n \) used, the standard stochastic kriging predictor with all the budget allocation rules outperforms the benchmark naive simulation method (naive). Secondly, we observe that except for two cases our proposed allocation rule \((n_{\text{tot}}, n_{\text{in}}) = ([N^{2/3}], [N^{1/3}])\) with a moderate number of replications \( n \ (2 \leq n \leq 50) \) yields better results compared to Gordy and Juneja [2010]'s 2/3 : 1/3 rule \((GJ_{n10})\). The minimum ERMSEs seem to be attained around \( n = 10 \) and \( n = 15 \) regardless of the size of \( N \) given that it is sufficiently large. This is along the same line as what Proposition 4 suggests because it is shown in Gordy and Juneja [2010] that the asymptotic bias and variance of the quantile estimators in two-level simulations satisfy Assumption 2.

![Figure 3: Boxplots of ERMSEs for \( p \)-quantile estimates over 100 macro-replications with \( k = 13 \) and \( p = 0.01 \).](attachment:image.png)

**5 Conclusions and Future Work**

In this paper, we investigate how to extend stochastic kriging metamodeling techniques to handle cases in which simulation response estimates are biased. Two well-known examples are presented to demonstrate the effectiveness of the modified stochastic kriging metamodels, respectively, estimation of the CVaR and VaR of a quadratic random loss distribution of two assets and prediction of the steady-state expected waiting time in an M/M/1 queue with initialization bias. One notable feature of these examples is that we need a sample of random outputs to obtain one simulated response at a single design point. The particular concern arises about experimental designs, namely, how to allocate available computational budget \( N \) between the number
of simulation replications $n$ and the sample size $n_s$ at each design point to minimize the mean squared error of the standard stochastic kriging predictor given the presence of bias. We find that the optimal allocation depends on the decreasing order of bias in $n_s$; and typically a moderate number of replications $n$ (e.g., $n \leq 50$) works sufficiently well given that the computational budget $N$ is not tight.

We further examine a similar problem in the two-level simulation context. The simulation procedure used to obtain a single simulation response estimate gets more involved because it necessitates outer level and inner level simulations. In this new setting, we study optimal allocation rules to spend the budget $N$ among the number of simulation replications $n$, the number of outer level simulated scenarios $n_{out}$, and the number of inner level simulation runs $n_{in}$. A collection of competing allocation rules for stochastic kriging are considered and their corresponding prediction performances are compared. The particular example used for demonstration here is the estimation of the extreme quantiles of a portfolio value distribution consisting of financial derivative contracts on a single asset. The results show that stochastic kriging with our proposed optimal budget allocation rule outperforms other competing rules and the results also exhibit features that are consistent with the analytical results.

Simulation experimental design is a burgeoning and active area of research in stochastic simulation, and this work is only one of first steps toward producing more efficient designs upon current practice for stochastic kriging metamodels. An in-depth analysis is provided for studying some special settings, and through empirical studies the analytical results are found to carry over into more general contexts. Nevertheless, a full theoretical treatment of the problem remains a future research topic. Other potential considerations include selection of design points, effects of estimating spatial parameters (in particular when simulated responses are biased) and metamodeling for steady-state simulations. Investigation of their joint effects including this paper under the stochastic kriging framework is expected to yield fruitful results.

Acknowledgement

The authors are grateful for many helpful and constructive comments from Prof. Barry L. Nelson. The work of K. Kim was supported by Basic Science Research Program through the NRF funded by the MEST (No. 2012-0003203).

References


A Appendix: Proofs

A.1 Proofs of Results when $\Sigma_e$ is Unknown

Proof of Lemma 1: We proceed as follows:

$$\text{MSE} \left( \hat{Y}(x_0; \hat{\sigma}) \right) = E \left[ \left( \hat{Y}(x_0; \hat{\sigma}) - \hat{Y}(x_0; \hat{\sigma})^o + \hat{Y}(x_0; \hat{\sigma})^o - Y(x_0) \right)^2 \right]$$

$$= E \left[ \left( w(\hat{\sigma})^T \xi(n_a) + \hat{Y}(x_0; \hat{\sigma})^o - Y(x_0) \right)^2 \right]$$

$$= E \left[ \left( w(\hat{\sigma})^T \xi(n_a) \right)^2 \right] + 2E \left[ w(\hat{\sigma})^T \xi(n_a) \left( \hat{Y}(x_0; \hat{\sigma})^o - Y(x_0) \right) \right]$$

$$+ E \left[ \left( \hat{Y}(x_0; \hat{\sigma})^o - Y(x_0) \right)^2 \right].$$

For the second term, we note that $\hat{\sigma}$ and $\hat{Y}^o$ are independent thanks to Assumption 1. In addition, $\hat{\sigma}$ and $Y(x_0)$ are independent because the former is related to simulation errors not to the extrinsic randomness. Therefore, we get

$$E \left[ \hat{Y}(x_0; \hat{\sigma})^o - Y(x_0) \right] = \left\{ f(x_0)^T \tilde{W} + \Sigma_M(x_0, \cdot)^T \Sigma^{-1} (I - F \tilde{W}) \right\} E \left[ \hat{Y}^o \right] - E \left[ Y(x_0) \right]$$

$$= f(x_0)^T \beta + \Sigma_M(x_0, \cdot)^T \Sigma^{-1} (F - F) \beta - f(x_0)^T \beta$$

$$= 0.$$ 

Hence, by the law of total expectation, we conclude that the second term is zero.

Next, we focus on the third term. Straightforward computations yield

$$E \left[ \left( \hat{Y}(x_0; \hat{\sigma})^o - Y(x_0) \right)^2 \right] = E \left[ \left( \hat{Y}(x_0; \hat{\sigma})^o - \hat{Y}(x_0; \sigma)^o + \hat{Y}(x_0; \sigma)^o - Y(x_0) \right)^2 \right]$$

$$= E \left[ \left( \hat{Y}(x_0; \hat{\sigma})^o - \hat{Y}(x_0; \sigma)^o \right)^2 \right] + \text{MSE} \left( \hat{Y}(x_0; \sigma)^o \right)$$

$$+ 2 \text{Cov} \left( \hat{Y}(x_0; \hat{\sigma})^o - \hat{Y}(x_0; \sigma)^o, \hat{Y}(x_0; \sigma)^o - Y(x_0) \right).$$

Then, the decomposition in the statement is immediate if the last covariance term is zero. To show that this is indeed the case, we first observe

$$\hat{Y}(x_0; \hat{\sigma})^o - \hat{Y}(x_0; \sigma)^o = (w(\hat{\sigma}) - w(\sigma))^T \hat{Y}^o,$$

and

$$\hat{Y}(x_0; \sigma)^o - Y(x_0) = w(\sigma)^T \hat{Y}^o - M(x_0) - f(x_0)^T \beta.$$ 

Now it is easy to see that

$$\text{Cov} \left( \hat{Y}(x_0; \hat{\sigma})^o - \hat{Y}(x_0; \sigma)^o, \hat{Y}(x_0; \sigma)^o - Y(x_0) \right)$$

$$= E \left[ w(\hat{\sigma})^T \hat{Y}^o (\hat{Y}^o)^T w(\sigma) \right] - E \left[ w(\sigma)^T \hat{Y}^o (\hat{Y}^o)^T w(\sigma) \right]$$

$$- E \left[ (w(\hat{\sigma}) - w(\sigma))^T \hat{Y}^o \left( M(x_0) + f(x_0)^T \beta \right) \right].$$
\[
E\left[w(\tilde{\sigma})^T \left(\Sigma + F\beta^T F^T\right) w(\sigma)\right] - w(\sigma)^T \left(\Sigma + F\beta^T F^T\right) w(\sigma)
\] 
\[-E\left[(w(\tilde{\sigma}) - w(\sigma))^T \left(\Sigma_M(x_0, \cdot) + F\beta f(x_0)^T \beta\right)\right].
\]

Using \(F^T w(\sigma) = F^T w(\tilde{\sigma}) = f(x_0)\) (which is straightforward to show), the above expression becomes
\[
= E\left[w(\tilde{\sigma})^T \left(\Sigma w(\sigma) + F\beta^T f(x_0)\right)\right] - w(\sigma)^T \left(\Sigma w(\sigma) + F\beta^T f(x_0)\right)
\] 
\[-E\left[(w(\tilde{\sigma}) - w(\sigma))^T \left(\Sigma_M(x_0, \cdot) + F\beta f(x_0)^T \beta\right)\right].
\]

Then direct computations show
\[
w(\tilde{\sigma})^T(\Sigma w(\sigma) - \Sigma_M(x_0, \cdot)) = f(x_0)^T \left(F^T \Sigma^{-1} F\right)^{-1} \left(f(x_0) - F^T \Sigma^{-1} \Sigma_M\right).
\]

It is true for the other term as well. Therefore, the covariance becomes zero. The proof is now complete.

**Proof of Lemma 2:** By definition, we have
\[
\check{Y}(x_0; \tilde{\sigma})^o - \check{Y}(x_0; \sigma)^o = (w(\tilde{\sigma}) - w(\sigma))^T \tilde{\gamma}^o.
\]

The first-order Taylor series approximation to \(w(\tilde{\sigma}) - w(\sigma)\) is given by
\[
w(\tilde{\sigma}) - w(\sigma) \approx \frac{\partial w(\sigma)}{\partial \sigma} (\tilde{\sigma} - \sigma).
\]

It follows that an approximation to the expectation in the statement can be given as follows:
\[
E\left[(\check{Y}(x_0; \tilde{\sigma})^o - \check{Y}(x_0; \sigma)^o)^2\right]
\] 
\[\approx E\left[(\tilde{\sigma} - \sigma)^T \frac{\partial w(\sigma)}{\partial \sigma}^T \tilde{\gamma}^o (\tilde{\gamma}^o)^T \frac{\partial w(\sigma)}{\partial \sigma} (\tilde{\sigma} - \sigma)\right]
\] 
\[= E\left[(\tilde{\sigma} - \sigma)^T \frac{\partial w(\sigma)}{\partial \sigma}^T (\Sigma + F\beta \beta^T F^T) \frac{\partial w(\sigma)}{\partial \sigma} (\tilde{\sigma} - \sigma)\right]
\] 
\[= \text{Tr}\left\{E\left[(\tilde{\sigma} - \sigma)^T \frac{\partial w(\sigma)}{\partial \sigma}^T (\Sigma + F\beta \beta^T F^T) \frac{\partial w(\sigma)}{\partial \sigma} (\tilde{\sigma} - \sigma)\right)\right\}
\] 
\[= \text{Tr}\left\{E\left[\frac{\partial w(\sigma)}{\partial \sigma}^T (\Sigma + F\beta \beta^T F^T) \frac{\partial w(\sigma)}{\partial \sigma} \text{Cov}(\tilde{\sigma})\right]\right\}.
\]
Proof of Proposition 2: Let us consider the first term in (11). The predictor \( \hat{Y}(x_0; \sigma)^o \) is unbiased. Hence, its MSE equals its variance, and we can obtain the result below by following exactly the same steps given in Appendix A.2.

\[
\text{MSE} \left( \hat{Y}(x_0; \sigma)^o \right) - \text{MSE}_\infty \left( \hat{Y}(x_0; \sigma)^o \right) \\
\approx \frac{\nu_5(A_1 + A_2) - \nu_6(1 - r_{12}^2)}{(1 - r_{12}^2)^2N} \\
+ \frac{\nu_7}{4(1 - r_{12}^2)^2N} \left\{ 4(1 + r_{12})\nu_8 - (A_1 + A_2)(1 + r_{12} - r_1 - r_2)(3 + r_{12}) \right\}.
\]

Here \( \text{MSE}_\infty \left( \hat{Y}(x_0; \sigma)^o \right) \) denotes the limit of MSE as \( N \rightarrow \infty \). Let us denote the term obtained above by \( \Theta_1 N^{-1} \).

Regarding the second term in (11), we notice that the arguments in Appendix A.2 can be applied by replacing \( \Gamma \) and \( \Sigma \) with \( \hat{\Gamma} \) and \( \hat{\Sigma} \) given below.

\[
\hat{\Sigma}^{-1} = \left( \begin{array}{cc} \tau^2 + \hat{\sigma}_{10}^2/n & \tau^2 r_{12}/n \\ \tau^2 r_{12}/n & \tau^2 + \hat{\sigma}_{20}^2/n \end{array} \right)^{-1} = \frac{1}{\hat{\Gamma}} \left( \begin{array}{c} \hat{a} \\ \hat{b} \\ \hat{c} \end{array} \right),
\]

where \( \hat{\Gamma} = a\hat{c} - \hat{b}^2 \). Then it is easy to see that with \( \hat{\Pi} = \hat{\alpha} + \hat{c} + 2\hat{b} \),

\[
\mathbf{w}(\hat{\sigma})^\top \mathbf{\xi}(n_s) = \frac{1}{n_s^2 \hat{\Pi}} \left\{ \tau^2(1 - r_2) + \hat{a} + \hat{b} \tau^2(r_2 - r_1) + \hat{b} + \hat{c} \right\} \left( \begin{array}{c} B_1 \\ B_2 \end{array} \right)
\]

\[
= \frac{1}{n_s^2 \hat{\Pi}} \left\{ \tau^2 \nu_1 + \frac{B_2 \hat{\sigma}_{10}^2 + B_1 \hat{\sigma}_{20}^2}{n} \right\},
\]

where \( \nu_1 \) is defined in Appendix A.2. Now, we observe that

\[
\text{E} \left[ (\mathbf{w}(\hat{\sigma})^\top \mathbf{\xi}(n_s))^2 \right] = \text{E} \left[ \frac{1}{n_s^{2\alpha}} \left( \frac{\tau^2 \nu_1 + \frac{B_2 \hat{\sigma}_{10}^2 + B_1 \hat{\sigma}_{20}^2}{n}}{2\tau^2(1 - r_{12})n + \hat{\sigma}_{10}^2 + \hat{\sigma}_{20}^2} \right)^2 \right]
\]

\[
\approx \frac{\nu_1^2}{4(1 - r_{12})^2n_s^{2\alpha}}.
\]

Let us denote this term by \( \Theta_2 n_s^{-2\alpha} \).

Lastly, we take a look at the third term. We note that

\[
\mathbf{w}(\sigma) = \frac{1}{\Pi} \left( \begin{array}{cc} \tau^2(r_1 - r_2) + a + b & \tau^2(r_2 - r_1) + b + c \end{array} \right)
\]

for \( \Sigma^{-1} = \Gamma^{-1} \left( \begin{array}{cc} a & b \\ b & c \end{array} \right) \) and \( \Gamma = ac - b^2 \), \( \Pi = a + c + 2b \). Also, it can be shown that

\[
\mathbf{W} = \mathbf{F}^\top \Sigma^{-1} \mathbf{F} = \frac{1}{\Pi} \left( \begin{array}{cc} a + b & b + c \end{array} \right) \text{ and thus } \mathbf{Q} := \Sigma^{-1} (\mathbf{I} - \mathbf{F} \mathbf{W}) = \frac{1}{\Pi} \left( \begin{array}{cc} 1 & -1 \\ -1 & 1 \end{array} \right).
\]

27
Recall that \( w(\sigma)^\top = f(x_0)^\top W + \Sigma_M(x_0, \cdot)^\top \Sigma^{-1} (I - FW) \). To compute \( \partial w(\sigma)/\partial \sigma \), note that the derivative of a matrix inverse \( A^{-1} \) is given by \(-A^{-1}(\partial A/\partial \sigma_{i0})A^{-1}\). Thus, we have (with \( X = F^\top \Sigma^{-1} F \) and \( W = X^{-1} F^\top \Sigma^{-1} \))
\[
\frac{\partial X}{\partial \sigma_{i0}} = -F^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \sigma_{i0}^2} \Sigma^{-1} F,
\]
and consequently,
\[
\frac{\partial W}{\partial \sigma_{i0}^2} = -X^{-1} \frac{\partial X}{\partial \sigma_{i0}^2} X^{-1} F^\top \Sigma^{-1} - X^{-1} F^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \sigma_{i0}^2} \Sigma^{-1} = W \frac{\partial \Sigma}{\partial \sigma_{i0}^2} W^\top F^\top \Sigma^{-1} - W \frac{\partial \Sigma}{\partial \sigma_{i0}^2} \Sigma^{-1} = W \frac{\partial \Sigma}{\partial \sigma_{i0}^2} \left( W^\top F^\top - I \right) \Sigma^{-1} = -W \frac{\partial \Sigma}{\partial \sigma_{i0}^2} Q.
\]
Next, we compute \((\partial w(\sigma)/\partial \sigma_{i0}^2)^\top\) as follows:
\[
\frac{\partial w(\sigma)}{\partial \sigma_{i0}^2} = f(x_0)^\top \frac{\partial W}{\partial \sigma_{i0}^2} + \Sigma_M(x_0, \cdot)^\top \frac{\partial \Sigma^{-1}}{\partial \sigma_{i0}^2} (I - FW) - \Sigma_M(x_0, \cdot)^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \sigma_{i0}^2} F \frac{\partial W}{\partial \sigma_{i0}^2} = -f(x_0)^\top W \frac{\partial \Sigma}{\partial \sigma_{i0}^2} Q - \Sigma_M(x_0, \cdot)^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \sigma_{i0}^2} Q + \Sigma_M(x_0, \cdot)^\top \Sigma^{-1} FW \frac{\partial \Sigma}{\partial \sigma_{i0}^2} Q = -f(x_0)^\top W \frac{\partial \Sigma}{\partial \sigma_{i0}^2} Q - \Sigma_M(x_0, \cdot)^\top \Sigma^{-1} (I - FW) \frac{\partial \Sigma}{\partial \sigma_{i0}^2} Q = -w^\top \frac{\partial \Sigma}{\partial \sigma_{i0}^2} Q.
\]
On the other hand,
\[
\frac{\partial \Sigma}{\partial \sigma_{i0}^2} = \frac{1}{n} \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ \end{pmatrix}, \quad \frac{\partial \Sigma}{\partial \sigma_{20}^2} = \frac{1}{n} \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ \end{pmatrix}.
\]
This leads us to
\[
\frac{\partial w(\sigma)}{\partial \sigma} = -\frac{1}{n\Pi} \begin{pmatrix} \tau^2(r_1 - r_2) + a + b & 0 \\ 0 & \tau^2(r_2 - r_1) + b + c \\ \end{pmatrix} Q =: -\frac{1}{n\Pi} PQ.
\]
Since \( QF = 0 \),
\[
\frac{\partial w(\sigma)}{\partial \sigma} \left( \Sigma + F\beta\beta^\top F^\top \right) \frac{\partial w(\sigma)}{\partial \sigma} = \frac{1}{n^2\Pi^2} P \left( Q\Sigma Q + QF\beta\beta^\top F^\top Q \right) P^\top = \frac{1}{n^2\Pi^2} PQP^\top = \frac{1}{n^2\Pi^2} \begin{pmatrix} P_{11}^2 & -P_{11}P_{22} \\ -P_{11}P_{22} & P_{22}^2 \\ \end{pmatrix}.
\]
Now we are ready to compute the last term in (11),
\[
\text{Tr} \left[ \left( \frac{\partial w(\sigma)}{\partial \sigma} \left( \Sigma + F\beta\beta^\top F^\top \right) \frac{\partial w(\sigma)}{\partial \sigma} \right) \text{Cov}(\hat{\sigma}) \right]
\]

28
where constants $w$'s are known. Consequently, we get
\[
\begin{align*}
\mathbf{w}^\top &= \mathbf{M}(\mathbf{x}_0, \cdot)^\top \mathbf{M}^{-1} + \eta^\top \mathbf{W} \\
&= \frac{1}{\Gamma} \mathbf{M}(\mathbf{x}_0, \cdot)^\top \left\{ \left( \begin{array}{cc} a & b \\ b & c \end{array} \right) - \frac{1}{\Pi} \left( \begin{array}{cc} a + b \\ b + c \end{array} \right) \left( \begin{array}{cc} a + b \\ b + c \end{array} \right) \right\} + \mathbf{W} \\
&= \frac{1}{\Pi} \left( \tau^2 (r_1 - r_2) + a + b \tau^2 (r_2 - r_1) + b + c \right).
\end{align*}
\]
Since the bias is $\mathbf{w}^\top \xi(n_{\text{in}}, n_{\text{out}})$ with $\xi(n_{\text{in}}, n_{\text{out}}) = \mathbf{n}^{-1}_{\text{in}}(\mathbf{B}_1, \mathbf{B}_2)^\top + \mathbf{n}^{-1}_{\text{out}}(\mathbf{C}_1, \mathbf{C}_2)^\top$, we conclude that
\[
\text{Bias} = \frac{1}{n_{\text{in}}} \times \frac{\tau^2 \nu_1 + \nu_2/n_{\text{tot}}}{\Pi(n_{\text{tot}})} + \frac{1}{n_{\text{out}}} \times \frac{\tau^2 \nu_3 + \nu_4/n_{\text{tot}}}{\Pi(n_{\text{tot}})},
\]
where constants $\nu_i$'s are
\[
\nu_1 = (r_1 - r_2 + 1 - r_{12}) B_1 + (r_2 - r_1 + 1 - r_{12}) B_2,
\]
\[
\nu_2 = (r_1 - r_2 + 1 - r_{12}) B_1 - (r_2 - r_1 + 1 - r_{12}) B_2,
\]
\[
\nu_3 = (r_1 - r_2 + 1 - r_{12}) B_1 - (r_2 - r_1 + 1 - r_{12}) B_2,
\]
\[
\nu_4 = (r_1 - r_2 + 1 - r_{12}) B_1 + (r_2 - r_1 + 1 - r_{12}) B_2.
\]
Proof of Proposition 3:
Regarding the variance part, we know that the limit is \( \text{MSE} \). Suppose that

\[
\nu_2 = A_2B_1 + A_1B_2, \\
\nu_3 = (r_1 - r_2 + 1 - r_{12})C_1 + (r_2 - r_1 + 1 - r_{12})C_2, \\
\nu_4 = A_2C_1 + A_1C_2. 
\]

The variance part can be computed as follows,

\[
\text{Var} \left( \hat{Y}(x_0) - Y(x_0) \right) = \mathbf{\Sigma}_M(x_0, x_0) - \mathbf{\Sigma}_M(x_0, \cdot)^\top \mathbf{\Sigma}^{-1} \mathbf{\Sigma}_M(x_0, \cdot) + \frac{\eta^2}{F^\top \mathbf{\Sigma}^{-1} F},
\]

with \( \nu_5, \nu_6 \) given by

\[
\nu_5 = r_1^2 + r_2^2 - 2r_1r_2r_{12}, \\
\nu_6 = r_2^2A_2 + r_1^2A_1.
\]

Notice that \( \nu_5 \) is nonnegative because \( 0 \leq |r_{12}| \leq 1 \) and thus \( \nu_5 \geq r_1^2 + r_2^2 - 2r_1 \cdot r_2 = (r_1 - r_2)^2 \geq 0 \). Therefore the final expression is readily obtained by writing \( \eta \) as

\[
\eta = \frac{1}{\Gamma(n_{tot})} \left\{ \frac{2}{n_{tot}} \frac{\nu_7 + \nu_8}{n_{tot}} + \frac{A_1A_2}{n_{tot}} \right\},
\]

in which

\[
\nu_7 = (1 - r_{12})(1 + r_{12} - r_1 - r_2), \\
\nu_8 = A_1 + A_2 - A_1r_2 - A_2r_1.
\]

Proof of Lemma 3:
Direct computations yield

\[
\text{Bias} = \frac{1}{\Pi(n_{tot})n_{tot}} \left\{ \frac{\tau^2 \nu_1n_{tot}^2 + \nu_2n_{tot}}{N} + \frac{\tau^2 \nu_3n_{tot} + \nu_4}{n_{out}} \right\}.
\]

Suppose that \( n_{tot} \) is not an increasing function of \( N \). This implies that \( n = O(1) \), \( n_{out} = O(1) \), and \( n_{in} = O(N) \). Then it is clear from the above expression that the bias becomes of constant order \( O(1) \). It is easy to see that the variance term has the same magnitude.

Suppose now that \( n_{tot} \) is increasing in \( N \). Then the bias term can be approximated by

\[
\text{Bias} \approx \frac{1}{2\tau^2(1 - r_{12})n_{tot}} \times \frac{\tau^2 \nu_1n_{tot}^2 + \nu_2n_{tot}n_{in}}{N} = \frac{1}{2(1 - r_{12})} \times \frac{\nu_1n_{tot} + \nu_3n_{in}}{N}.
\]

Regarding the variance part, we know that the limit is \( \text{MSE}_\infty(\hat{Y}(x_0)) \). The asymptotic variance minus the limiting MSE can be shown equal to the sum of the second and third terms in (17). It then becomes clear that we can make the sum converge to zero by choosing \( n_{in} \) and \( n_{out} \) that increase to infinity as \( N \to \infty \). Therefore, in any effective allocation rule \( n_{tot} \) should increase in \( N \).

Proof of Proposition 3: Since \( n_{tot} \) should be increasing in \( N \) according to Lemma 3, some algebraic manipulations yield the following approximation:

\[
\text{MSE}(\hat{Y}(x_0)) - \text{MSE}_\infty(\hat{Y}(x_0)) \approx \Lambda_1 \left( \frac{\nu_1}{n_{in}} + \frac{\nu_3}{n_{out}} \right)^2 + \frac{\Lambda_2}{n_{tot}}.
\]

30
for some positive constants $\Lambda_1$ and $\Lambda_2$. It is then obvious that the quantity above is minimized at $n = 1$ for each fixed $n_{\text{tot}}$ if $\nu_1 \nu_3 > 0$.

Let us define a function $f(n_{\text{tot}}) = \Lambda_1(n_{\text{tot}}/N + \nu_3/n_{\text{tot}})^2 + \Lambda_2/n_{\text{tot}}$. The first-order optimality condition reads

$$2\Lambda_1 \left(\nu_1^2 n_{\text{tot}}^2 - \nu_3^2 N^2\right) = \Lambda_2 N^2 n_{\text{tot}}.$$ 

Since $n_{\text{tot}}$ is increasing in $N$, we obtain $n^*_\text{tot} = O(N^{2/3})$ by matching the leading order terms. To get a more precise description, let us set $n^*_\text{tot} = cN^{2/3} + o(N^{2/3})$. Some calculations lead to the following optimal allocation rule

$$n^* = 1, \quad n^*_\text{out} = \left(\frac{\Lambda_2}{2\Lambda_1 \nu_1^2}\right)^{\frac{1}{3}} N^{\frac{2}{3}}, \quad n^*_\text{in} = \left(\frac{2\Lambda_1 \nu_1^2}{\Lambda_2}\right)^{\frac{1}{3}} N^{\frac{1}{3}}.$$

A.3 Proofs for the Two-design-point Model in the Two-level Simulation Context when $\Sigma_c$ is Unknown

We assume that the bias in the two-level simulation under consideration is given by

$$\xi(x_i; n_{\text{in}}, n_{\text{out}}) = \frac{B_i}{n_{\text{in}}} + \frac{C_i}{n_{\text{out}}}.$$

As explained in the main body of the paper, the first term in (18) minus $\text{MSE}_\infty(\hat{Y}(x_0))$ is denoted by $\Theta_1 n_{\text{tot}}^{-1}$ which is equal to (17). For the second term in (18), we borrow the arguments in Appendix A.1. By modifying the notation, we arrive at

$$E \left[ \left( w(\hat{\sigma})^\top \xi(n_{\text{in}}, n_{\text{out}}) \right)^2 \right] \approx \frac{1}{4(1-r_{12})^2} \left( \frac{\nu_1}{n_{\text{in}}} + \frac{\nu_3}{n_{\text{out}}} \right)^2 \frac{n_{\text{out}}}{n_{\text{tot}}}.$$ 

We denote it by $\Theta_2 = 4(1-r_{12})^{-2}$. As for the third and the last terms, we note that the role of $n_s$ in the proof of Proposition 2 is now replaced by $n_{\text{out}}$. Hence the final result is unchanged and we have

$$\text{The third term} = \frac{(r_1 - r_2 + 1 - r_{12})^2 A_1^2 + (r_2 - r_1 + 1 - r_{12})^2 A_2^2}{4r^2(1-r_{12})^3} \times \frac{n_{\text{out}}}{n_{\text{tot}}} n_{\text{tot}},$$

which we write as $\Theta_3 n_{\text{out}} n_{\text{tot}}^{-3}$.

Proof of Proposition 4: Using $n_{\text{in}} = N/n_{\text{out}}$ and $n = n_{\text{tot}}/n_{\text{out}}$, an approximation to the objective function that we intend to minimize can be expressed as follows

$$\text{MSE} \left( \hat{Y}(x_0; \hat{\sigma}) \right) - \text{MSE}_\infty \left( \hat{Y}(x_0; \sigma^0) \right) \approx \frac{\Theta_1}{n_{\text{tot}}} + \Theta_2 \left( \frac{\nu_1 n_{\text{tot}}}{N} + \frac{\nu_3 n_{\text{tot}}}{n_{\text{tot}}} \right)^2 \frac{\Theta_3}{n_{\text{tot}} n_{\text{out}}} = (20).$$
where \( \gamma \geq 1 \) and \( 1 \leq n \leq n_{\text{tot}} \leq N \). It is clear that choosing \( n_{\text{tot}} = O(1) \) cannot be optimal. Let us assume that \( n_{\text{tot}} \) is some increasing function of \( N \). Then the first-order optimality condition yields

\[
2\nu_3\gamma \Theta_2 \left( \nu_1 n_{\text{tot}}^{\gamma+1} + \nu_3 n^\gamma N \right) n^{\gamma+1} = \Theta_3 n_{\text{tot}}^{2\gamma-2} N.
\]

Matching leading order terms shows that an optimal \( n \) is of form \( cn_{\text{tot}}^q + o(n_{\text{tot}}^q) \) for some \( q \in [0, 1] \). To see this, let us consider two possibilities: (1) \( n_{\text{tot}}^{\gamma+1} = O(n^\gamma N) \) and (2) \( n^\gamma N = o(n_{\text{tot}}^{\gamma+1}) \). Case (1) leads us to \( n = O \left( n_{\text{tot}}^{(2\gamma-2)/(2\gamma+1)} \right) \). Case (2) first yields \( n = O \left( n_{\text{tot}}^{(\gamma-3)(\gamma+1)/N} \right) \); however, by plugging this term back into (20) and matching the leading order terms through tedious algebraic manipulations, we see that \( n_{\text{tot}} = O(N^p) \) for some \( p \in (0, 1) \).

For more concrete expressions, we solve two scenarios \( q = 0 \) and \( q \in (0, 1] \) separately. Consider \( q = 0 \) and let \( n = O(1) \). Then the first-order optimality condition reads

\[
\frac{d\text{MSE}}{dn_{\text{tot}}} = -\frac{\Theta_1}{n_{\text{tot}}} + 2\Theta_2 \left( \frac{\nu_1 n_{\text{tot}}}{N} + \frac{\nu_3 n^\gamma}{n_{\text{tot}}^{\gamma+1}} \right) \left( \frac{\nu_1}{N} - \frac{\gamma \nu_3 n^\gamma}{n_{\text{tot}}^{\gamma+1}} \right) - \frac{2\Theta_3}{n_{\text{tot}}^2 N} + \text{(lower order terms)} = 0.
\]

Rearranging terms, we obtain

\[
2\Theta_2 \left( \nu_1 n_{\text{tot}}^{\gamma+1} + \nu_3 n^\gamma N \right) \left( \nu_1 n_{\text{tot}}^{\gamma+1} - \gamma \nu_3 n^\gamma N \right) = \Theta_1 N^2 n_{\text{tot}}^{2\gamma-1} + 2\Theta_3 N^2 n_{\text{tot}}^{2\gamma-2} + \text{(lower order terms)}.
\]

It follows that either \( n_{\text{tot}}^{2\gamma+2} = O \left( N^2 n_{\text{tot}}^{2\gamma-1} \right) \) or \( n_{\text{tot}}^{\gamma+1} = O \left( N^2 n_{\text{tot}}^{2\gamma-1} \right) \). The former yields \( n_{\text{tot}} = O \left( N^{2/3} \right) \) while the latter gives \( n_{\text{tot}}^{2\gamma} = O(N) \) which is only possible when \( \gamma = 1 \) and \( n_{\text{tot}} = O(N) \). However, it can be shown that this allocation rule produces a slower convergence rate of MSE as compared to \( n_{\text{tot}} = O \left( N^{2/3} \right) \). Therefore, the scenario of \( q = 0 \) results in the following allocation rule

\[
n^* = O(1), \quad n_{\text{out}}^* = O \left( N^{\frac{2}{3}} \right), \quad n_{\text{in}}^* = O \left( N^{\frac{1}{3}} \right) .
\]

Now let \( n = cn_{\text{tot}}^q + o(n_{\text{tot}}^q) \) for some \( q \in (0, 1] \). Also recall that \( n_{\text{tot}} = O(N^p) \) for some \( p \in (0, 1) \). In this case, it is useful to adopt a different parametrization, namely, \( x = n_{\text{out}} = \alpha N^r + o(N^r) \) and \( y = n_{\text{in}} = \beta N^s + o(N^s) \) for some \( r, s \geq 0 \). The objective function now can be written as a function of \( x \) and \( y \) below

\[
f(x, y) = \frac{\Theta_1}{N} y + \Theta_2 \left( \frac{\nu_1}{y} + \frac{\nu_3}{x^\gamma} \right)^2 + \frac{\Theta_3}{N^3} x^2 y^3 .
\]

The first-order optimality conditions yield

\[
\frac{\partial f}{\partial x} = -2\Theta_2 \left( \frac{\nu_1}{y} + \frac{\nu_3}{x^\gamma} \right) \frac{\gamma \nu_3}{x^{\gamma+1}} + \frac{\Theta_3}{N^3} y^3 = 0 ,
\]

\[
\frac{\partial f}{\partial y} = \frac{\Theta_1}{N} - 2\Theta_2 \left( \frac{\nu_1}{y} + \frac{\nu_3}{x^\gamma} \right) \frac{\nu_1}{y^2} + \frac{3\Theta_3}{N^3} x^2 y^2 = 0 .
\]

We rewrite these equations as follows,

\[
2\gamma \Theta_2 \nu_3 (\nu_1 x^\gamma + \nu_3 y) = \Theta_3 x^{\gamma+1} y^4 N^{-3} ,
\]

\[
2\Theta_2 \nu_1 (\nu_1 x^\gamma + \nu_3 y) = \Theta_1 x^\gamma y^3 N^{-1} + 3\Theta_3 x^{\gamma+1} y^5 N^{-3} .
\]
Since $x = \alpha N^r + o(N^r)$ and $y = \beta N^s + o(N^s)$, it follows that

\[ 2\gamma \mathcal{M}_2 \nu_3 (\nu_1 \alpha^3 N^{r\gamma} + \nu_3 \beta N^s) + \text{(lower orders)} = \Theta_3 \alpha^{2\gamma + 1} \beta^4 N^{r(2\gamma + 1) + 4s - 3}, \]

\[ 2\mathcal{M}_2 \nu_1 (\nu_1 \alpha^3 N^{r\gamma} + \nu_3 \beta N^s) + \text{(lower orders)} = \Theta_4 \alpha^{\gamma} \beta^3 N^{r\gamma + 3s - 1} + 3\Theta_3 \alpha^{\gamma + 1} \beta^5 N^{r(\gamma + 1) + 5s - 3}. \]

Depending on whether $r \gamma \geq s$ or not, two solutions result from matching leading order terms. If $r \gamma \geq s$, we have

\[
\begin{align*}
\frac{r\gamma}{r} = r(2\gamma + 1) + 4s - 3 & \quad \text{and} \quad \frac{r\gamma}{s} = r\gamma + 3s - 1 \\
\frac{r\gamma}{s} = r(2\gamma + 1) + 4s - 3 & \quad \text{and} \quad \frac{r\gamma}{s} = r(\gamma + 1) + 5s - 3.
\end{align*}
\]

The system of equations on the right-hand side yields $r = 3/(5\gamma + 1)$ and $s = 3\gamma/(5\gamma + 1)$. But it is easy to see that this choice violates the optimality equations. The system of equations on the left-hand side yields $r = 5/(3\gamma + 3)$ and $s = 1/3$. However, we find that this allocation rule yields $n = O(n_{\text{tot}}^q)$ for $q > 0$ only when $r < 2/3$, i.e., $2\gamma > 3$. If this is the case, then we get

\[
n^* = O\left( N^{(\frac{2\gamma - 3}{5\gamma + 3})} \right), \quad n_{\text{out}}^* = O\left( N^{\frac{5}{(3\gamma + 3)}} \right), \quad n_{\text{in}}^* = O\left( N^{\frac{1}{3}} \right). \tag{22}
\]

When $r \gamma < s$, we also have two systems of equations. One can check that neither one leads to a valid solution. Therefore, if $\gamma \leq 1.5$, then the allocation rule given in (21) is optimal; and if $\gamma > 1.5$, then we have two possible choices, namely, (21) and (22).

We next determine which one of the two allocation rules is better when $\gamma > 1.5$. For (21), we see that the objective function has

\[
\Theta_1 \frac{1}{n_{\text{tot}}} + \Theta_2 \frac{\nu_1^2 n_{\text{tot}}^2}{N^2} + O\left( N^{-\frac{2\gamma + 1}{3}} \right) + O\left( N^{-\frac{4\gamma}{3}} \right) + O\left( N^{-\frac{4}{3}} \right),
\]

where the dominating term is $O(N^{-4/3})$ excluding the first two terms. For (22), it yields

\[
\Theta_1 \frac{1}{n_{\text{tot}}} + \Theta_2 \frac{\nu_1^2 n_{\text{tot}}^2}{N^2} + O\left( N^{-\frac{6\gamma + 1}{3(\gamma + 1)}} \right) + O\left( N^{-\frac{10\gamma}{3(\gamma + 1)}} \right) + O\left( N^{-\frac{6\gamma + 1}{3(\gamma + 1)}} \right),
\]

in which the dominating term is $O\left( N^{-\frac{6\gamma + 1}{3(\gamma + 1)}} \right)$ excluding the first two terms. Since the latter one decreases faster than $O(N^{-4/3})$, the allocation rule given in (22) is a better choice when $\gamma > 1.5$. Nevertheless, we should notice that the dominating terms are the first two terms for both cases.