Reducing Parameter Uncertainty for Stochastic Systems*

Szu Hui NG† Stephen E. CHICK‡

November 21, 2005

Abstract

The design of many production and service systems is informed by stochastic model analysis. But the parameters of statistical distributions of stochastic models are rarely known with certainty, and are often estimated from field data. Even if the mean system performance is a known function of the model's parameters, there may still be uncertainty about the mean performance because the parameters are not known precisely. Several methods have been proposed to quantify this uncertainty, but data sampling plans have not yet been provided to reduce parameter uncertainty in a way that effectively reduces uncertainty about mean performance. The optimal solution is challenging, so we use asymptotic approximations to obtain closed-form results for sampling plans. The results apply to a wide class of stochastic models, including situations where the mean performance is unknown but estimated with simulation. Analytical and empirical results for the *M/M/1* queue, a quadratic response-surface model, and a simulated critical care facility illustrate the ideas.

Category: G.3: Probability and Statistics, Probabilistic algorithms (including Monte Carlo) experimental design

Category: I.6: Simulation and Modeling, Simulation output analysis

Terms: Experimentation, Performance

Keywords: Stochastic simulation, uncertainty analysis, parameter estimation, Bayesian statistics

1 Introduction

Stochastic modeling and simulation analysis are useful approaches to evaluate the performance of production and service systems as a function of design and statistical parameters [e.g., Buzacott and Shanthikumar 1993; Law and Kelton 2000]. Models of existing systems, or variations of existing systems, can help inform system design and improvement decisions. Data may be available to help estimate statistical parameters, but estimators are subject to random variation because they are functions of random phenomena. Parameter uncertainty means that there is a risk that a planned system will not perform as expected [Cheng and Holland 1997; Chick 1997, 2001; Barton and Schruben 2001], even if the relationship between the parameters and system performance is

†S.-H. Ng, Dept. Industrial and Systems Engineering, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260, isensh@nus.edu.sg.

‡S. E. Chick, Technology and Operations Management Area, INSEAD, Boulevard de Constance, 77305 Fontainebleau CEDEX, France, stephen.chick@insead.edu.

completely known. For example, the stationary mean occupancy of a stable *M/M/1* queue is a known function of the arrival and service rates, but the mean occupancy is unknown if those rates are not precisely known. When the system performance is an unknown function of the parameters, and is estimated by simulation, the problem is aggravated. The common practice of inputting point estimates of parameters into simulations can dramatically reduce the coverage of a supposedly $100(1-\alpha)\%$ confidence interval (CI) for the mean performance (Barton and Schruben 2001).

Bootstrap methods (Cheng and Holland 1997; Barton and Schruben 2001), asymptotic normality approximations (Cheng and Holland 1997; Ng and Chick 2001), and Bayesian model averaging [Draper 1995; Chick 2001; Zouaoui and Wilson 2003, 2004] can quantify the effect of input parameter uncertainty on output uncertainty.

This paper goes beyond quantifying uncertainty by suggesting how to reduce parameter uncertainty in a way that effectively reduces uncertainty about the mean performance of a system. We do so for a broad class of stochastic systems by using asymptotic variance approximations for the unknown mean system performance. Section 2 formalizes the stochastic model assumptions along with examples. Section 3 introduces the asymptotic variance approximations for input parameter uncertainty. Section 4 describes how input parameter uncertainty affects output uncertainty and formulates an optimization problem to reduce output uncertainty. When the system response is a known function of the inputs, the optimal sampling plan turns out to be very similar to known results for stratified sampling. When the system response must be estimated with simulation, a different sampling allocation is obtained. The allocation shows how to balance the cost of running additional simulations (to reduce uncertainty about the mean response and its gradient) and data collection (to reduce uncertainty about the inputs to the system). Under some general conditions, the optimal sampling allocations are shown to have the attractive property of being invariant to the coordinate systems used to represent the parameter of each input distribution. Section 5 applies the results to several examples. This paper extends preliminary work (Ng and Chick 2001) by generalizing results to a broader class of distributions, and by providing results for unknown response functions. An appendix also describes how to reduce the mean-squared error (MSE) rather than the variance. The approach taken is Bayesian, so parameter uncertainty is quantified with random variables.

2 Model Formulation and Examples

Consider a system with multiple statistical parameters $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k)$ that influence the system performance $g(\boldsymbol{\Theta})$. Simulated performance realizations Y_l ("output") with input parameter $\boldsymbol{\Theta}_l$ for replication $l = 1, 2, \ldots,$

$$Y_l = g(\boldsymbol{\Theta}_l) + \sigma(\boldsymbol{\Theta}_l) Z_l,$$

may be observed if $g(\boldsymbol{\Theta})$ is unknown, where σ^2 is the output variance, and the Z_l are independent zero-mean random variables with unit variance.

The system is driven by k scalar (real-valued) input processes that are mutually independent sources of randomness. Given the parameter vector $\boldsymbol{\theta}_i$ for the ith input process, the corresponding observations $\{x_{i\ell} : \ell = 1, 2, \ldots\}$ are independent and identically distributed with conditional probability density function $p_i(x_{i\ell} \mid \boldsymbol{\theta}_i)$. Uncertainty about the parameter $\boldsymbol{\theta}_i$ is initially described with a probability model $\pi_i(\boldsymbol{\theta}_i)$. In Bayesian statistics, $\pi_i(\boldsymbol{\theta}_i)$ is called a prior probability distribution. Here we presume joint independence across sources of randomness, $\pi(\boldsymbol{\Theta}) = \prod_{i=1}^{k} \pi_i(\boldsymbol{\theta}_i)$.

Each $\boldsymbol{\theta}_i$ is assumed to be inferrable from data. The data $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in_i})$ that has been observed so far for the ith input process can be used to get a more precise idea about the value of

the unknown parameter $\boldsymbol{\theta}_i$, whose uncertainty is characterized by the posterior distribution

$$f_{in_i}(\boldsymbol{\theta}_i \mid \boldsymbol{x}_i) \propto \pi_i(\boldsymbol{\theta}_i) \prod_{\ell=1}^{n_i} p_i(x_{i\ell} \mid \boldsymbol{\theta}_i), \tag{1}$$

for $i = 1, 2, \ldots, k$. We will use those distributions to figure out how to collect additional data, if needed, to reduce uncertainty about input parameters. Let $\boldsymbol{n} = (n_1, \ldots, n_k)$ specify the number of data points observed so far for each source of randomness. Probability statements below are all conditional on all data collected so far, $\mathcal{E}_{\boldsymbol{n}} = (\boldsymbol{x}_1^{\mathrm{T}}, \ldots, \boldsymbol{x}_k^{\mathrm{T}})$, unless otherwise specified.

A tilde denotes the maximum *a posteriori* (MAP) estimate of any parameter, e.g. $\tilde{\boldsymbol{\theta}}_{in_i}$ maximizes f_{in_i}. Maximum likelihood estimates (MLEs) are denoted with a hat, $\hat{\boldsymbol{\theta}}_{in_i}$. Given the data $\boldsymbol{x}_i$, the MLE maximizes $\prod_{\ell=1}^{n_i} p_i(x_{i\ell} \mid \boldsymbol{\theta}_i)$.

Parameters may be multivariate, such as $\boldsymbol{\theta}_1 = (\vartheta_1, \vartheta_2)$. It will be convenient at times to refer to the individual components of the parameter vector $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k)$ encompassing all k sources of randomness; and for this purpose we write $\boldsymbol{\Theta} = (\vartheta_1, \vartheta_2, \ldots, \vartheta_K)$, where $K > k$ when at least one of the $\boldsymbol{\theta}_i$ are multidimensional. Decision variables and system control parameters are considered to be fixed and implicit in the definition of g, so that we focus upon the effects of input parameter uncertainty. Table 1 summarizes this notation, as well as much of the notation used in the rest of the paper.

The *M/M/1* queue can be described with this notation as a system with $k = 2$ sources of randomness (arrival and service times), with arrival rate $\boldsymbol{\theta}_1 = \lambda$ and service rate $\boldsymbol{\theta}_2 = \mu$. There are a total of $K = 2$ components of the overall parameter vector $\boldsymbol{\Theta}$, so $\boldsymbol{\Theta} = (\lambda, \mu)$. The stationary mean occupancy is known in closed form if the queue is stable, $g(\boldsymbol{\Theta}) = \lambda/(\mu - \lambda)$.

A less trivial example for which the system response is unknown is the critical care facility depicted in Figure 1. Schruben and Margolin (1978) studied that system to determine the expected number of patients per month that are denied a bed, as a function of the number of beds in the intensive care unit (ICU), coronary care unit (CCU), and intermediate care units. Patients arrive according to a Poisson process and are routed through the system depending upon their specific health condition. Their analysis presumed fixed point estimates (MLE) for the parameters of the $k = 6$ sources of randomness: the patient arrival process (Poisson arrivals, mean $\hat{\boldsymbol{\theta}}_{1n} = 3.3$/day), ICU stay duration (lognormal $\boldsymbol{\theta}_2 = (\vartheta_2, \vartheta_3)$, with mean $\hat{\vartheta}_{2n} = 3.4$ and standard deviation $\hat{\vartheta}_{3n} = 3.5$ days), three more bivariate input parameters $\boldsymbol{\theta}_3, \boldsymbol{\theta}_4, \boldsymbol{\theta}_5$ for the lognormal service times at the intermediate ICU, intermediate CCU, and CCU processes, and a sixth parameter for patient routing (multinomial, $\hat{\boldsymbol{\theta}}_{6n} = (\hat{p_1}, \hat{p_3}, \hat{p_4}) = (.2, .2, .05)$, note that $p_2 = 1 - p_1 - p_3 - p_4$). There are therefore a total of $K = 12$ components of the overall parameter vector $\boldsymbol{\Theta}$. The function $g(\cdot)$ is not known, and is to be estimated from simulation near the point $\hat{\boldsymbol{\Theta}}_n$.

This paper will use approximations that assume a constant variance, $\sigma(\boldsymbol{\Theta}) = \sigma$ (homoscedastic). If the variance actually depends upon $\boldsymbol{\Theta}$ (heteroscedastic), then we will estimate σ by the MAP estimator $\tilde{\sigma}(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}})$, a Bayesian analog of the usual MLE for the standard deviation, $\hat{\sigma}(\hat{\boldsymbol{\Theta}}_n)$. This approximation is a standard assumption in the design of experiments literature (Box, Hunter, and Hunter 1978; Box and Draper 1987; Myers, Khuri, and Carter 1989), and is asymptotically valid under certain conditions (Mendoza 1994, p. 176) if $\sigma(\boldsymbol{\Theta})$ is a "nice" function of $\boldsymbol{\Theta}$ near $\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}}$.

The approximations made below in a Bayesian context are analogous to frequentist approximations by Cheng and Holland (1997). The modification of the analysis is with a straightforward application of results of Bernardo and Smith (1994), and result in a decoupling of stochastic uncertainty from parameter uncertainty when $g(\cdot)$ is a known function. Zouaoui and Wilson (2001) used a similar decoupling to obtain a variance reduction for estimates of $\mathrm{E}[g(\boldsymbol{\Theta}) \mid \mathcal{E}_{\boldsymbol{n}}]$ with the Bayesian Model Average (BMA) when multiple candidate distributions are proposed for a single

Table 1: Summary Of Notation[1]

k	number of independent sources of randomness, each consisting of i.i.d. scalar (real-valued) random variables
$\boldsymbol{\theta}_i$	statistical input parameter for ith source of randomness (can be multivariate with dimension $d_i \geq 1$), $i = 1, 2, \ldots, k$
K	total number of components of all k parameter vectors, $K = \sum_{i=1}^{k} d_i$
$\boldsymbol{\Theta}$	vector of all statistical parameters, $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k)$
ϑ_j	the jth individual component of $\boldsymbol{\Theta}$ for $j = 1, \ldots, K$, so that we have $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k) = (\vartheta_1, \vartheta_2, \ldots, \vartheta_K)$
$g(\boldsymbol{\Theta})$	the mean response of the model, as a function of the statistical parameters
r	number of replications a simulated model
σ	standard deviation of output of a simulated model
n_i	number of observations taken from the ith source of randomness and used to estimate the associated parameter vector $\boldsymbol{\theta}_i$, for $i = 1, \ldots, k$
$\boldsymbol{n}$	vector of number of observations, $\boldsymbol{n} = (n_1, n_2, \ldots, n_k)$
$x_{i\ell}$	ℓth observation for ith statistical parameter, $i = 1, \ldots, k$; $\ell = 1, 2, \ldots, n_i$
$\boldsymbol{x}_i$	vector of observations $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in_i})$ for ith statistical parameter
$\mathcal{E}_{\boldsymbol{n}}$	all data observed so far, $\mathcal{E}_{\boldsymbol{n}} = (\boldsymbol{x}_1^{\mathrm{T}}, \boldsymbol{x}_2^{\mathrm{T}}, \ldots, \boldsymbol{x}_k^{\mathrm{T}})$
$\pi_i(\boldsymbol{\theta}_i)$	prior probability density of parameter $\boldsymbol{\theta}_i$, $i = 1, \ldots, k$
$p_i(x \mid \boldsymbol{\theta}_i)$	density function for data x given parameter $\boldsymbol{\theta}_i$, $i = 1, \ldots, k$
$f_{in_i}(\boldsymbol{\theta}_i)$	posterior probability density of parameter $\boldsymbol{\theta}_i$ given data $\boldsymbol{x}_i$ (n_i observations)
$\tilde{\boldsymbol{\theta}}_i$	MAP estimate for ith statistical parameter, $i = 1, \ldots, k$
$\hat{\boldsymbol{\theta}}_i$	MLE for ith statistical parameter, $i = 1, \ldots, k$
$\boldsymbol{G}_i = \boldsymbol{\Sigma}_i^{-1}$	observed information matrix associated with the parameter vector $\boldsymbol{\theta}_i$, given $\boldsymbol{x}_i$ for $i = 1, \ldots, k$. Also sometimes written $\boldsymbol{G}_{in_i} = \boldsymbol{\Sigma}_{in_i}^{-1}$ to emphasize the dependence on the sample size n_i
$\boldsymbol{\Sigma}$	Block diagonal matrix $\boldsymbol{\Sigma} = \mathrm{diag}[\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k] = \mathrm{diag}[\boldsymbol{G}_1^{-1}, \ldots, \boldsymbol{G}_k^{-1}]$
$\boldsymbol{H}_i$	Expected information matrix of one observation for parameter i, $i = 1, \ldots, k$
β_j	Derivative of $g(\boldsymbol{\Theta})$ with respect to jth component of $\boldsymbol{\Theta}$, evaluated at $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$, for $j = 1, \ldots, K$
$\boldsymbol{\beta_i}$	$\nabla_{\boldsymbol{\theta}_i} g(\boldsymbol{\Theta}) \mid_{\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}}$, the gradient vector of the function $g(\boldsymbol{\Theta})$ with respect to the parameter vector $\boldsymbol{\theta}_i$ for the ith source of randomness, evaluated at $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$
$\boldsymbol{\beta}$	$\nabla_{\boldsymbol{\Theta}} g(\boldsymbol{\Theta}) \mid_{\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}}$, the gradient vector of the function $g(\boldsymbol{\Theta})$ with respect to the parameter vector $\boldsymbol{\Theta}$, evaluated at $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$
m_i	number of additional data points to be collected for ith statistical parameter
$\boldsymbol{m}$	vector of number of additional points, $\boldsymbol{m} = (m_1, m_2, \ldots, m_k)$
m_0	number of additional simulation replications to be run
$\mathcal{D}$	all additional data to be collected (m_i points for ith source, $i = 0, 1, \ldots, k$)

[1] A subscript i may be dropped to denote a generic source of randomness to simplify notation. The addition of (a) a $\hat{}$ denotes maximum likelihood estimate (MLE), (b) a $\tilde{}$ denotes a MAP estimator, (c) an extra subscript n, n_i or $\boldsymbol{n}$ emphasizes the number of data points that determine the estimate, (d) an extra subscript $\boldsymbol{\Theta}$, $\boldsymbol{\psi}$ may be added to emphasize the coordinate system for the parameters, as in $\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\psi}}$.

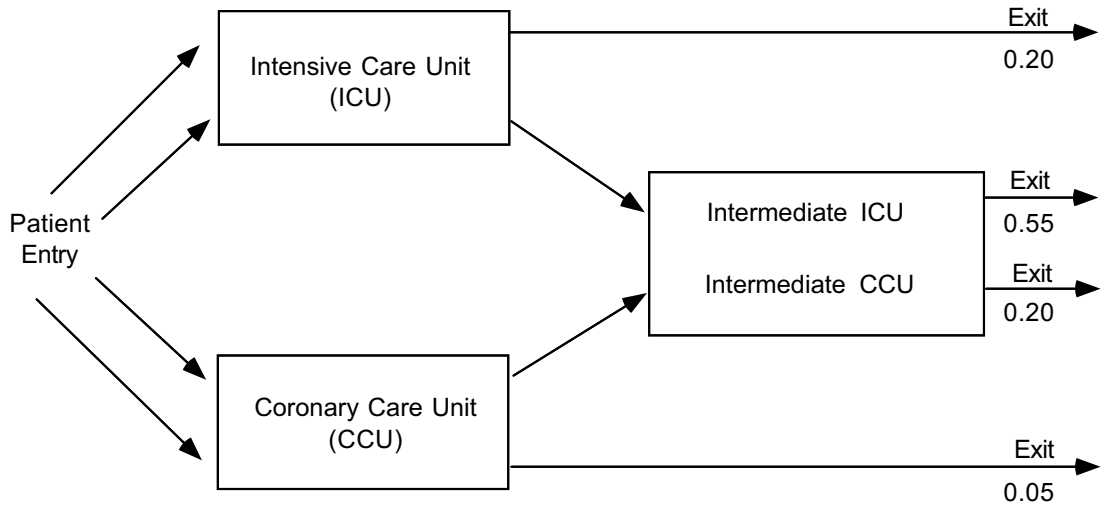

Figure 1: Fraction of patients routed through different units of a critical care facility.

source of randomness, and the response is homoscedastic. Their result appears to be extendable to the heteroscedastic case ($n \to \infty$) by approximating σ with $\tilde{\sigma}(\tilde{\boldsymbol{\Theta}}_n)$. The idea of decoupling stochastic and parameter uncertainty is applied to the case of an unknown $g(\boldsymbol{\Theta})$ with an unknown gradient in Section 4.2.

3 Approximations for Parameter Uncertainty

Uncertainty about parameters, and the effect of data collection variability, is approximated by the variance. The mean squared error (MSE) of estimates is treated in Appendix A. The analysis is motivated by analogy with the following well-known result for confidence intervals based on asymptotic normality. Suppose that the simulation outputs $y_1, \ldots, y_r$ are observed, with sample mean $\bar{y}_r$, sample variance S_r^2, and $100(1-\alpha)\%$ CI for the mean $\bar{y}_r \pm t_{r-1,1-\alpha/2}(S_r^2/r)^{\frac{1}{2}}$. An *approximate* expression for the number of additional observations m^* required to obtain an absolute error of ϵ is $m^* = \min\left\{m \geq 0 : t_{r+m-1,1-\alpha/2}\left[S_r^2/(r+m)\right]^{1/2} \leq \epsilon\right\}$. For large r, $t_{r+m-1,1-\alpha/2}$ does not change much as a function of m. The CI width is essentially scaled by shrinking the variance for the unknown mean from S_r^2/r to

$$S_r^2/(r+m). \tag{2}$$

Analogous results exist for asymptotic normal approximations to the posterior distributions of input parameters.

For the rest of this section, we focus on *one* parameter $\boldsymbol{\theta} = (\vartheta_1, \ldots, \vartheta_d)$ for a single source of randomness (dropping the subscript $_i$ for notational simplicity).

Proposition 3.1. *For each n, let $f_n(\cdot|\boldsymbol{x}_n)$ denote the posterior p.d.f. of the d-dimensional parameter vector $\boldsymbol{\theta}_n$; let $\tilde{\boldsymbol{\theta}}_n$ be the associated MAP; and define the Bayesian observed information $\boldsymbol{G}_n = \boldsymbol{\Sigma}_n^{-1}$ by*

$$[\boldsymbol{G}_n]_{jl} = \left[\boldsymbol{\Sigma}_n^{-1}\right]_{jl} = -\left.\frac{\partial^2 \log f_n(\boldsymbol{\theta} \mid \boldsymbol{x}_n)}{\partial\vartheta_j \partial\vartheta_l}\right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_n} \text{ for } j,l = 1,\ldots,d. \tag{3}$$

Then the random variable $\boldsymbol{\Sigma}_n^{-1/2}(\boldsymbol{\theta}_n - \tilde{\boldsymbol{\theta}}_n)$ converges in distribution to a standard (multivariate) normal random variable as $n \to \infty$ if certain regularity conditions hold.

Proof. See Bernardo and Smith (1994, Prop 5.14). The regularity conditions presume a nonzero prior probability density near the "true" $\boldsymbol{\theta}$, and the "steepness", "smoothness", and "concentration" regularity conditions imply that the data is informative for all components of $\boldsymbol{\theta}_n$, so the posterior becomes peaked near $\tilde{\boldsymbol{\theta}}_n$. □

Colloquially, $\boldsymbol{\Sigma}_n$ is proportional to $1/n$, so the posterior variance shrinks as in Equation (2) as additional samples are observed. Gelman, Carlin, Stern, and Rubin (1995, Sec. 4.3) describe when the regularity conditions may not hold, including underspecified models, nonidentified parameters, cases in which the number of parameters grows with the sample size, unbounded likelihood functions, improper posterior distributions, or convergence to a boundary of the parameter space.

The expected information matrix in *one* observation is

$$[\boldsymbol{H}(\tilde{\boldsymbol{\theta}}_n)]_{jl} = \mathrm{E}_{X_\ell}\left[-\frac{\partial^2 \log p(X_\ell \mid \boldsymbol{\theta})}{\partial\vartheta_j \partial\vartheta_l}\right]\Bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_n} \text{ for } j,l = 1,\ldots,d.$$

so the analog to the asymptotic approximation in Equation (2) for the variance of $\boldsymbol{\theta}$, given that m additional data points are to be collected, is

$$(\boldsymbol{\Sigma}_n^{-1} + m\boldsymbol{H}(\tilde{\boldsymbol{\theta}}_n))^{-1}. \tag{4}$$

Equation (4) simplifies under some special conditions. A sufficient condition is that the distribution for $\boldsymbol{\theta}$ be in the regular exponential family (which includes the exponential, gamma, normal, binomial, multinomial, lognormal, Poisson, and many other distributions) and that a conjugate prior distribution be used to describe initial parameter uncertainty (see Appendix B). Use of a conjugate prior distribution implies that the posterior distribution will have the same functional form as the prior distribution.

Appendix B shows that if the prior distribution before observing any data is a conjugate prior, then $\boldsymbol{\Sigma}_n^{-1} = (n_0 + n)\boldsymbol{H}(\tilde{\boldsymbol{\theta}}_n)$ for some n_0 and Equation (4) simplifies:

$$(\boldsymbol{\Sigma}_n^{-1} + m\boldsymbol{H}(\tilde{\boldsymbol{\theta}}_n))^{-1} = \frac{\boldsymbol{H}^{-1}(\tilde{\boldsymbol{\theta}}_n)}{n_0 + n + m}. \tag{5}$$

Equation (4) also simplifies if the MLE $\hat{\boldsymbol{\theta}}_n$ replaces $\tilde{\boldsymbol{\theta}}_n$ and the expected information $\hat{\boldsymbol{\Sigma}}_n^{-1} = n\boldsymbol{H}(\hat{\boldsymbol{\theta}}_n)$ replaces $\boldsymbol{\Sigma}_n^{-1}$.

All of the distributions for the *M/M/1* queue and the critical care facility are in the regular exponential family and satisfy the regularity conditions of Proposition 3.1. The specific numbers of service time and routing decision observations for the critical care facility were not published. For the analysis here, we presumed standard noninformative prior distributions for each unknown parameter (Bernardo and Smith 1994). We further assumed that the published point estimates were MLEs based on 100 patients, so that 20 patients were routed to the ICU alone, 5 were routed to the CCU alone, and so forth. This assumption uniquely determines the sufficient statistics of the data and the proper posterior distributions for the parameters. It also uniquely determines the observed information matrices $\boldsymbol{\Sigma}_n^{-1}$ for each parameter. Ng and Chick (2001) provided details for the information matrix and input uncertainty of the Poisson and lognormal distributions. For the multinomial routing probabilities, the Dirichlet distribution is the conjugate prior, $\pi(p_1, p_3, p_4) \propto p_1^{\alpha_{01}-1} p_3^{\alpha_{03}-1} p_4^{\alpha_{04}-1}(1 - p_1 - p_3 - p_4)^{\alpha_{02}-1}$, with noninformative distribution $\alpha_{0j} = 1/2$ for $j = 1,\ldots,4$ (Bernardo and Smith 1994). Reparametrizing this into the canonical conjugate form as in Appendix B, we have $\boldsymbol{\psi} = (\log p_1, \log p_3, \log p_4)$. Since this mapping is bijective, the results from Appendix B also hold in the (p_1, p_3, p_4) coordinates. If s_j patients are routed to location j, then the

posterior distribution is Dirichlet with parameters $\alpha_j = \alpha_{0j} + s_j$. If $\alpha_j > 1$ for all j, then the MAP is $\tilde{p}_j = (\alpha_j - 1)/[\sum_{j=1}^{4}(\alpha_j - 1)]$, and $\boldsymbol{\Sigma}_n^{-1} = [\sum_{j=1}^{4}(\alpha_j - 1)]\boldsymbol{H}(\tilde{\boldsymbol{\theta}}_n) = [\frac{4}{2} + n - 4]\boldsymbol{H}(\tilde{\boldsymbol{\theta}}_n) = (n_0 + n)\boldsymbol{H}(\tilde{\boldsymbol{\theta}}_n)$, where $n_0 = -2$.

4 Parameter and Performance Uncertainty

Section 4.1 derives approximations for performance uncertainty, measured by variance, as a function of parameter uncertainty when $g(\cdot)$ is a known function. It formulates and solves an optimization problem that allocates resources for further data collection to minimize an asymptotic approximation to performance uncertainty. Section 4.2 identifies sampling allocations when $g(\cdot)$ is unknown. It shows whether it is more important to run more simulation replications (to improve the gradient estimate) or to collect more field data (to reduce parameter uncertainty).

4.1 Known Response Function

Denote gradient information of the known response surface, $g(\boldsymbol{\Theta})$, evaluated at the MAP estimator of $\boldsymbol{\Theta}$, by

$$\begin{aligned} \beta_j &= \left.\frac{\partial g(\boldsymbol{\Theta})}{\partial \vartheta_j}\right|_{\boldsymbol{\Theta}=\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}}} \text{ for } j = 1, \ldots, K \\ \boldsymbol{\beta_i} &= \left.\nabla_{\boldsymbol{\theta}_i} g(\boldsymbol{\Theta})\right|_{\boldsymbol{\Theta}=\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}}} \text{ for } i = 1, \ldots, k \\ \boldsymbol{\beta} &= (\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \ldots, \boldsymbol{\beta_k}). \end{aligned} \tag{6}$$

Suppose that the parameters for each source of randomness are estimated with n observations each, and that each satisfies the asymptotic normality properties of Proposition 3.1. Proposition 4.1 states that $g(\boldsymbol{\Theta})$ is also asymptotically normal as more observations are collected from each of the sources of randomness.

Proposition 4.1. *Use the assumptions, notation and technical conditions of Proposition 3.1, with the modification that $\boldsymbol{\Theta}_n = (\boldsymbol{\theta}_{1n}, \ldots, \boldsymbol{\theta}_{kn})$ is the nth vector of parameters, so that $f_{in}(\cdot)$ is the posterior density function of the random variable $\boldsymbol{\theta}_{in}$, the MAP estimator of $\boldsymbol{\Theta}$ is $\tilde{\boldsymbol{\Theta}}_n = (\tilde{\boldsymbol{\theta}}_{1n}, \ldots, \tilde{\boldsymbol{\theta}}_{kn})$, and $\boldsymbol{\Sigma}_n = diag[\boldsymbol{\Sigma}_{1n} \ldots \boldsymbol{\Sigma}_{kn}]$ is the nth block diagonal matrix with appropriate submatrices for parameters $i = 1, \ldots, k$, assuming n observations are available from each source of randomness. Suppose that $\boldsymbol{\Theta}$ has an asymptotic Normal$\left(\tilde{\boldsymbol{\Theta}}_n, \boldsymbol{\Sigma}_n\right)$ distribution, as in the conclusion of Proposition 3.1. Suppose that g is continuously differentiable near $\boldsymbol{\Theta}_0$, that $\bar{\sigma}_n^2 \to 0$ and $\tilde{\boldsymbol{\Theta}}_n \to \boldsymbol{\Theta}_0$ in probability, and $\bar{\sigma}_n^2 = O(\underline{\sigma}_n^2)$, where $\bar{\sigma}_n^2$ and $\underline{\sigma}_n^2$ respectively denote the largest and smallest eigenvalues of $\boldsymbol{\Sigma}_n$.*

The random variable $(\boldsymbol{\beta}_n \boldsymbol{\Sigma}_n \boldsymbol{\beta}_n^T)^{-1/2}[g(\boldsymbol{\Theta}) - g(\tilde{\boldsymbol{\Theta}}_n)]$ converges in distribution to the standard normal distribution. Colloquially, $g(\boldsymbol{\Theta})$ is asymptotically distributed

$$Normal\left(g(\tilde{\boldsymbol{\Theta}}_n), \boldsymbol{\beta}_n \boldsymbol{\Sigma}_n \boldsymbol{\beta}_n^T\right).$$

Proof. Mendoza (1994) and Bernardo and Smith (1994, Prop. 5.17) provide an analogous result for bijective functions $g(\boldsymbol{\Theta})$ when $k = 1$. The result holds for univariate $g(\boldsymbol{\Theta})$ because the marginal distribution of a single dimension of a multivariate joint Gaussian is also Gaussian. That result generalizes directly to $k \geq 1$, as required here, with straightforward algebra. This is a Bayesian analog for classical results from Serfling (1980, Sec. 3.3). □

Block diagonality reflects an assumption that the parameters for different sources of randomness are *a priori* independent, and that samples are conditionally independent, given the value of $\boldsymbol{\Theta}$.

The plug-in estimator $g(\tilde{\boldsymbol{\Theta}}_n)$ converges to $g(\boldsymbol{\Theta}_0)$ but is biased for nonlinear functions. Appendix A discusses bias and the mean squared error of estimates, using a first-order bias approximation. For now we focus on the variance of the estimators.

The number of observations for each of the k sources of randomness may differ, as may the number of additional data points to collect. Let n_i be the number of data points for source i, suppose that m_i additional points are to be collected for each source of randomness i, and set $\boldsymbol{m} = (m_1, \ldots, m_k)$. Using the approximation of Equation (4), we see that the variance in the unknown performance after collecting additional information is approximately

$$V_{\text{par}}(\boldsymbol{m}) = \sum_{i=1}^{k} \boldsymbol{\beta}_{\boldsymbol{in_i}} (\boldsymbol{\Sigma}_{in_i}^{-1} + m_i \boldsymbol{H}_i(\tilde{\boldsymbol{\theta}}_{in_i}))^{-1} \boldsymbol{\beta}_{\boldsymbol{in_i}}{}^T \tag{7}$$

If each observed information matrix is proportional to the corresponding expected information matrix, as when conjugate priors are used, then Equation (7) simplifies:

$$V_{\text{par}}(\boldsymbol{m}) = \sum_{i=1}^{k} \frac{\xi_i}{n_{0i} + n_i + m_i}, \text{ where } \xi_i = \boldsymbol{\beta}_{\boldsymbol{in_i}} \boldsymbol{H}_i^{-1}(\tilde{\boldsymbol{\theta}}_{in_i}) \boldsymbol{\beta}_{\boldsymbol{in_i}}{}^T. \tag{8}$$

The optimal sampling plan for reducing this asymptotic approximation to the variance in system performance, assuming that sampling costs for source i is linear in the number of samples, is therefore the solution to Problem (9).

$$\begin{array}{lll} \min & V_{\text{par}}(\boldsymbol{m}) & \\ \text{s.t.} & m_i \geq 0 & \text{for } i = 1, \ldots, k \\ & \sum c_i m_i = B & \end{array} \tag{9}$$

Cost differences may arise if data collection can be automated for some but not all sources of randomness, if suppliers differ in willingness to share data, and so forth.

Proposition 4.2. *If $V_{par}(\boldsymbol{m})$ simplifies to Equation (8), the integer restriction is relaxed (continuous m_i), and B is large, then the solution to Problem (9) is:*

$$m_i^* = \frac{B + \sum_{\ell=1}^{k} (n_{0\ell} + n_\ell) c_\ell}{\sum_{j=1}^{k} \left(\frac{\xi_j c_i c_j}{\xi_i} \right)^{1/2}} - (n_{0i} + n_i) \text{ for } i = 1, \ldots, k \tag{10}$$

Proof. See Appendix C. □

For small B, the nonnegativity constraints for the m_i need consideration. Special features of the data collection process can be handled by adding constraints, e.g. set a subset of m_i's to be equal for simultaneously reported data.

Example: The mean occupancy of an *M/M/1* queue is $g(\lambda, \mu) = \lambda/(\mu - \lambda)$ when $\mu > \lambda$. The gamma distribution is the conjugate prior for an unknown rate of samples with an exponential distribution. Suppose that a gamma prior distribution Gamma (α, ν) models uncertainty for the unknown arrival rate, with mean α/ν and density $f_\Lambda(w) = \nu^\alpha w^{\alpha-1} \exp(-w\nu)/\Gamma(\alpha)$ for all $w \geq 0$. If n_a arrivals are observed, the posterior distribution is also Gamma with parameters $\alpha_\lambda = \alpha + n_a, \nu_\lambda = \nu + \sum_{l=1}^{n_a} x_{il}$. The MAP estimate is $\tilde{\lambda} = (\alpha_\lambda - 1)/\nu_\lambda$, and $\boldsymbol{H}_1(\tilde{\lambda}) = 1/\tilde{\lambda}^2 = \nu_\lambda^2/(\alpha_\lambda - 1)^2$.

With calculus, $\boldsymbol{\Sigma}_{1n_a}^{-1} = (\alpha_\lambda - 1)/\tilde{\lambda}^2 = \nu_\lambda^2/(\alpha_\lambda - 1)$. Similarly assume the service rate has a conjugate gamma prior, that n_s service times are observed, resulting in a Gamma (α_μ, ν_μ) posterior distribution. If $\tilde{\mu} > \tilde{\lambda}$ (otherwise, we would consider a multiserver system), then Equation (7) implies

$$V_{\text{par}}(\boldsymbol{m}) \;=\; \frac{\tilde{\lambda}^2\tilde{\mu}^2}{(\tilde{\mu}-\tilde{\lambda})^4}\left(\frac{1}{\alpha_\lambda - 1 + m_a} + \frac{1}{\alpha_\mu - 1 + m_s}\right).$$

As conjugate priors are used, $\boldsymbol{\Sigma}_{1n_a}^{-1} = (\alpha_\lambda - 1)\boldsymbol{H}_1(\tilde{\lambda}_{n_a})$ and $\boldsymbol{\Sigma}_{2n_s}^{-1} = (\alpha_\mu - 1)\boldsymbol{H}_2(\tilde{\mu}_{n_s})$. This reduces to Equation (8) where $n_{0a} + n_a = \alpha_\lambda - 1$ and $n_{0s} + n_s = \alpha_\mu - 1$.

If m_a interarrival times can be collected at cost c_a each, and m_s service times can be collected at cost c_s each, then by Proposition 4.2, it is optimal to sample with $m_s^* = (B - c_a m_a^*)/c_s$ and

$$m_a^* \;=\; \frac{B + (\alpha_\lambda - 1)c_a + (\alpha_\mu - 1)c_s}{c_a + (c_a c_s)^{1/2}} - (\alpha_\lambda - 1)$$

for large B. If $c_a = c_s$, this choice minimizes the difference between $\alpha_\lambda + m_a$ and $\alpha_\mu + m_s$ (the same effective number of data points is desired for both parameters).

If the arrival and service rates are believed to have a gamma distribution, then the queue is unstable with positive probability. In fact, the expectation of the average queue length does not exist even after conditioning on stability, $\mathrm{E}_{\lambda,\mu}[\lambda/(\mu - \lambda) \mid \lambda < \mu] = \infty$ (Chick 2001). This means that using Bayesian models for unknown statistical input parameters may lead to simulation output processes whose expectation is not defined. At least two alternative approaches can be taken. First, a similar analysis can be done for a model that not only has finite moments but is also more realistic (say, by considering a finite waiting capacity or by modeling a finite time horizon rather than an infinite horizon time average). Second, one can use the asymptotic normality approximation in spite of the lack of moments as a starting point for reducing input uncertainty. Section 5.2 illustrates by example when this second alternative might be expected to work well or not.

The optimal allocation for Problem (9) does not depend on the parametrization $\boldsymbol{\Theta}$ under certain conditions. Let $\boldsymbol{T} : \mathbb{R}^K \mapsto \mathbb{R}^K$ be a one-to-one continuously differentiable transformation $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_K) = \boldsymbol{T}(\boldsymbol{\Theta}) = [T_1(\boldsymbol{\Theta}), \ldots, T_K(\boldsymbol{\Theta})]$ with differentiable inverse $\boldsymbol{\Theta} = \boldsymbol{T}^{-1}(\boldsymbol{\psi})$, and nonsingular matrix of derivatives $\boldsymbol{J}_{\boldsymbol{T}}(\boldsymbol{\Theta})$, where $[\boldsymbol{J}_{\boldsymbol{T}}(\boldsymbol{\Theta})]_{j\ell} = \partial T_j/\partial\vartheta_\ell$. We write $\boldsymbol{J}_{\boldsymbol{T}^{-1}}(\boldsymbol{\psi})$, where $[\boldsymbol{J}_{\boldsymbol{T}^{-1}}(\boldsymbol{\psi})]_{j\ell} = \partial T_j^{-1}/\partial\psi_\ell$. Let $\tilde{\boldsymbol{\psi}}_{\boldsymbol{n}}$ be the MAP estimator for $\boldsymbol{\psi}$ given $\mathcal{E}_{\boldsymbol{n}}$. Denote the response gradient with respect to $\boldsymbol{\psi}$ by $\boldsymbol{\beta}_{\boldsymbol{\psi}}$ and the gradient in Equation (6) with respect to $\boldsymbol{\Theta}$ by $\boldsymbol{\beta}_{\boldsymbol{\Theta}}$. Similarly denote the observed information matrix and expected information in one observation (measured in $\boldsymbol{\psi}$ coordinates) by $\boldsymbol{\Sigma}_{\boldsymbol{\psi},\boldsymbol{n}}$ and $\boldsymbol{H}_{\boldsymbol{\psi}}$ to emphasize the coordinate system. The expression $V_{\text{par}}(\boldsymbol{m})$ in Equation (8) explicitly depends upon $\boldsymbol{\Theta}$. We write $V_{\text{par}}(\boldsymbol{m};\boldsymbol{\Theta})$ to denote that dependence explicitly. Let $V_{\text{par}}(\boldsymbol{m};\boldsymbol{\psi})$ denote the analogous approximation approximation when computed in $\boldsymbol{\psi}$ coordinates.

Proposition 4.3. *Suppose the setup of Proposition 4.1, that the distribution of $\boldsymbol{\Theta}$ is in the regular exponential family, and that a conjugate prior is adopted. Let $\boldsymbol{T} : \mathbb{R}^K \mapsto \mathbb{R}^K$ be a one-to-one continuously differentiable transformation $\boldsymbol{\psi} = \boldsymbol{T}(\boldsymbol{\Theta}) = [T_1(\boldsymbol{\Theta}), \ldots, T_K(\boldsymbol{\Theta})]$ with differentiable inverse $\boldsymbol{\Theta} = \boldsymbol{T}^{-1}(\boldsymbol{\psi})$, and nonsingular Jacobian $\boldsymbol{J}_{\boldsymbol{T}}(\boldsymbol{\Theta})$. If conditions (a) for each i, $\psi_i = T_i(\boldsymbol{\Theta})$ depends upon only one $\boldsymbol{\theta}_j$, and (b) $\tilde{\boldsymbol{\psi}}_{\boldsymbol{n}} = \boldsymbol{T}(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}})$ are true, then $V_{par}(\boldsymbol{m};\boldsymbol{\Theta}) = V_{par}(\boldsymbol{m};\boldsymbol{\psi})$.*

Proof. Presume the setup of Proposition 4.1, define $\boldsymbol{T}$ and the other notation as in the problem statement, and suppose conditions (a) and (b) hold. Condition (a) means that $\boldsymbol{J}_{\boldsymbol{T}}(\boldsymbol{\Theta})$ has the same block diagonal structure as $\boldsymbol{\Sigma}$, and that no ψ_i is a function of parameter vectors for multiple sources of randomness.

Suppose first that there is only $k = 1$ source of randomness with parameter $\boldsymbol{\theta}$ and drop the subscript i for notational simplicity. In this notation, ξ_i can be rewritten $\xi = \boldsymbol{\beta}_{\boldsymbol{\Theta}} \boldsymbol{H}_{\boldsymbol{\Theta}}^{-1}(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}}) \boldsymbol{\beta}_{\boldsymbol{\Theta}}{}^{\mathrm{T}}$ in $\boldsymbol{\Theta}$ coordinates. By the chain rule for partial differentiation, and the assumption $\tilde{\boldsymbol{\psi}}_{\boldsymbol{n}} = \boldsymbol{T}(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}})$,

$$\begin{aligned} \beta_{\boldsymbol{\psi}} &= \nabla_{\boldsymbol{\psi}} g(\boldsymbol{\psi})\big|_{\boldsymbol{\psi}=\tilde{\boldsymbol{\psi}}_{\boldsymbol{n}}} \\ &= \left(\nabla_{\boldsymbol{\Theta}} g(\boldsymbol{\Theta})\big|_{\boldsymbol{\Theta}=\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}}}\right)\left(\boldsymbol{J}_{\boldsymbol{T}^{-1}}(\tilde{\boldsymbol{\psi}}_{\boldsymbol{n}})\right). \end{aligned}$$

By Schervish (1995, p. 115), $\boldsymbol{H}_{\boldsymbol{\psi}}(\boldsymbol{\psi}) = \boldsymbol{J}_{\boldsymbol{T}^{-1}}^{\mathrm{T}}(\boldsymbol{\psi}) \boldsymbol{H}_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}) \boldsymbol{J}_{\boldsymbol{T}^{-1}}(\boldsymbol{\psi})$, so that with additional calculus $\boldsymbol{H}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\psi}) = (\boldsymbol{J}_{\boldsymbol{T}^{-1}}^{\mathrm{T}}(\boldsymbol{\psi}) \boldsymbol{H}_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}) \boldsymbol{J}_{\boldsymbol{T}^{-1}}(\boldsymbol{\psi}))^{-1} = \boldsymbol{J}_{\boldsymbol{T}}(\boldsymbol{\Theta}) \boldsymbol{H}_{\boldsymbol{\Theta}}^{-1}(\boldsymbol{\Theta}) \boldsymbol{J}_{\boldsymbol{T}}^{\mathrm{T}}(\boldsymbol{\Theta})$. Therefore, in $\boldsymbol{\psi}$ coordinates,

$$\begin{aligned} \xi &= \boldsymbol{\beta}_{\boldsymbol{\psi}} \boldsymbol{H}_{\boldsymbol{\psi}}^{-1}(\tilde{\boldsymbol{\psi}}_{\boldsymbol{n}}) \boldsymbol{\beta}_{\boldsymbol{\psi}}{}^{\mathrm{T}} \\ &= [\nabla_{\boldsymbol{\Theta}} g(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}}) \boldsymbol{J}_{\boldsymbol{T}^{-1}}(\tilde{\boldsymbol{\psi}}_{\boldsymbol{n}})][\boldsymbol{J}_{\boldsymbol{T}}(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}}) \boldsymbol{H}_{\boldsymbol{\Theta}}^{-1}(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}}) \boldsymbol{J}_{\boldsymbol{T}}^{\mathrm{T}}(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}})][\boldsymbol{J}_{\boldsymbol{T}^{-1}}^{\mathrm{T}}(\tilde{\boldsymbol{\psi}}_{\boldsymbol{n}}) \nabla_{\boldsymbol{\Theta}}^{\mathrm{T}} g(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}})] \\ &= \nabla_{\boldsymbol{\Theta}} g(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}}) \boldsymbol{H}_{\boldsymbol{\Theta}}^{-1}(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}}) \nabla_{\boldsymbol{\Theta}}^{\mathrm{T}} g(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}}) \\ &= \boldsymbol{\beta}_{\boldsymbol{\Theta}} \boldsymbol{H}_{\boldsymbol{\Theta}}^{-1}(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}}) \boldsymbol{\beta}_{\boldsymbol{\Theta}}{}^{\mathrm{T}}. \end{aligned}$$

The third equality follows because $\boldsymbol{J}_{\boldsymbol{T}}(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}}) \boldsymbol{J}_{\boldsymbol{T}^{-1}}(\tilde{\boldsymbol{\psi}}_{\boldsymbol{n}}) = \boldsymbol{J}_{\boldsymbol{T}^{-1}}(\tilde{\boldsymbol{\psi}}_{\boldsymbol{n}}) \boldsymbol{J}_{\boldsymbol{T}}(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}})$ is the $K \times K$ identity matrix, when $\tilde{\boldsymbol{\psi}}_{\boldsymbol{n}} = \boldsymbol{T}(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}})$ holds.

For $k > 1$, the block diagonality assumption of Condition (a) assures that the ξ_i decouple, one for each of the k independent sources of randomness. □

The MAP estimator is not invariant to coordinate changes. That is, if $\tilde{\boldsymbol{\Theta}}_n$ is the MAP in $\boldsymbol{\Theta}$ coordinates, and $\tilde{\boldsymbol{\psi}}_n$ is the MAP in $\boldsymbol{\psi}$ coordinates, then it may not be true that $\tilde{\boldsymbol{\psi}}_n = \boldsymbol{T}(\tilde{\boldsymbol{\Theta}}_n)$. For example, if the unknown rate ψ of an exponential distribution has a Gamma (α_n, ν_n) distribution, then $\tilde{\psi}_n = (\alpha_n - 1)/\nu_n$ and the mean $\theta = 1/\psi$ has an inverted gamma distribution, with MAP $\tilde{\theta}_n = \nu_n/(\alpha_n + 1) \neq 1/\tilde{\psi}_n$. The multivariate generalization of Proposition 4.1 implies that $\tilde{\boldsymbol{\psi}}_n$ and $\boldsymbol{T}(\tilde{\boldsymbol{\Theta}}_n)$ are asymptotically equal. For a Bayesian, then, the allocations are asymptotically invariant to parametrization, in this sense.

For a frequentist, the allocations are invariant for finite n, as the MLE is invariant, given mild regularity conditions (Edwards 1984), and expected information matrices behave as required for a development parallel to Proposition 4.3.

4.2 Unknown Response Function

If the response function $g(\boldsymbol{\Theta})$ is unknown, an approximation is needed to estimate the mean of some future output Y_{r+1}. The asymptotic approximations above "hit their limit" in some sense with linear approximations, so we approximate $g(\boldsymbol{\Theta})$ near $\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}}$ with a first-order regression model that may depend on past simulation output.

$$\mathrm{E}[Y_{r+1} \mid \boldsymbol{\Theta}] \approx h(\boldsymbol{\Theta}, \boldsymbol{\beta}) = \beta_0 + \sum_{j=1}^{K} \beta_j (\vartheta_j - \tilde{\vartheta}_j) \tag{11}$$

The notation $h(\boldsymbol{\Theta}, \boldsymbol{\beta})$ replaces the notation $g(\boldsymbol{\Theta})$ to make it explicit that the estimated response depends upon both $\boldsymbol{\Theta}$ and $\boldsymbol{\beta}$. Note that β_0 takes the role of $g(\tilde{\boldsymbol{\Theta}})$ and β_j takes the role of $\partial g(\boldsymbol{\Theta})/\partial \vartheta_j$. If the derivatives $\beta_j = \partial h/\partial \vartheta_j$ are estimated from simulations, for $j = 1, 2, \ldots, K$, then uncertainty about both $\boldsymbol{\Theta}$ and $\boldsymbol{\beta}$ must be considered. We take a Bayesian approach and presume $\boldsymbol{\beta}$ to be a random vector, whose posterior distribution is updated when simulation replications are run.

Suppose that r replications have been run, with $r \times (K+1)$ design matrix $\mathbf{M} = [(\boldsymbol{\Theta}_1 - \tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}}), (\boldsymbol{\Theta}_2 - \tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}}), \ldots, (\boldsymbol{\Theta}_r - \tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}})]^T$, resulting in outputs $\boldsymbol{y} = [y_1, \ldots, y_r]^T$ with sample mean $\bar{y} = r^{-1} \sum_{i=1}^{r} y_i$. The r runs may have come from q replications at each of t design points selected with classical design of experiments (DOE) techniques, with $r = qt$. Or the $\boldsymbol{\Theta}_i$ may have come from using a BMA to sample t values of $\boldsymbol{\Theta}_i$ independently from $f(\boldsymbol{\Theta} \mid \boldsymbol{x})$, with q replications for each $\boldsymbol{\Theta}_i$ (Chick (2001) used $q = 1$, Zouaoui and Wilson (2003) extended to $q > 1$ to assess stochastic uncertainty, also see Zouaoui and Wilson (2004)). In either case, the posterior distribution of $\boldsymbol{\beta}$, assuming the standard linear model with a noninformative prior distribution $p(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$, and that $r > (K+1)$ (Gelman, Carlin, Stern, and Rubin 1995, Sec. 8.3), is

$$\begin{aligned} p(\boldsymbol{\beta} \mid \boldsymbol{y}, \mathbf{M}, \sigma^2) &\sim \text{Normal}\left(\hat{\boldsymbol{\beta}}_r, (\mathbf{M}^T\mathbf{M})^{-1}\sigma^2)\right), \\ p(\sigma^2 \mid \boldsymbol{y}, \mathbf{M}) &\sim \text{Inverse-Chi}^2\left(r - (K+1), s^2\right), \\ \hat{\boldsymbol{\beta}}_r &= (\hat{\beta}_0, \hat{\boldsymbol{\beta}}_{\mathbf{1}}, \hat{\boldsymbol{\beta}}_{\mathbf{2}}, \ldots, \hat{\boldsymbol{\beta}}_{\boldsymbol{k}}) = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\boldsymbol{y} \\ s^2 &= \frac{1}{r - (K+1)}(\boldsymbol{y} - \mathbf{M}\hat{\boldsymbol{\beta}}_r)^T(\boldsymbol{y} - \mathbf{M}\hat{\boldsymbol{\beta}}_r). \end{aligned} \tag{12}$$

The standard classical estimate for $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}_r$, and for σ^2 is s^2.

The posterior marginal distribution of $\boldsymbol{\beta}$ is actually multivariate t, not normal. We approximate this distribution by estimating σ^2 by s^2 and presuming the variance is known, as we did in Equation (2), to arrive at a normal distribution approximation for the unknown response vector that has mean $\hat{\boldsymbol{\beta}}_r$ and variance of the form $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}/(r_0 + r)$. Similarly, we presume that the posterior variance of $\boldsymbol{\beta}$ is of the form $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}/(r_0 + r + m_0)$ when m_0 additional replications are run (cf. Equation (8)). Amongst other scenarios, this approximation makes sense when $m_0 = q't$, with q' additional replications are run at each of the t originally selected $\boldsymbol{\Theta}_i$ as above.

Then the approximation $V_{\text{tot}} = \text{Var}(h(\boldsymbol{\Theta}, \boldsymbol{\beta}) \mid \mathcal{E}_{\boldsymbol{n}}, \boldsymbol{y}, \mathbf{M}, \sigma^2 = s^2)$ for overall performance uncertainty $\text{Var}(\text{E}[Y_{r+1}] \mid \mathcal{E}_{\boldsymbol{n}}, \boldsymbol{y}, \mathbf{M})$ asymptotically consists of components that reflect parameter uncertainty (V_{par}), response surface/gradient uncertainty (V_{resp}), and stochastic noise ($\sigma^2 = s^2$). The following expectations, through Equation (13), implicitly condition on all observations so far, $\mathcal{E}_{\boldsymbol{n}}, \boldsymbol{y}, \mathbf{M}$.

$$\begin{aligned} V_{\text{tot}} &\approx \frac{s^2}{r_0 + r} + \sum_{i=1}^{K}\sum_{j=1}^{K} \text{E}\left\{\left[\beta_i(\vartheta_i - \tilde{\vartheta}_i)(\vartheta_j - \tilde{\vartheta}_j)^T\beta_j\right]\right. \\ &\qquad \left. - \text{E}\left[\beta_i(\vartheta_i - \tilde{\vartheta}_i)\right]\text{E}\left[\beta_j(\vartheta_j - \tilde{\vartheta}_j)\right]\right\} \\ &= \frac{s^2}{r_0 + r} + \sum_{i=1}^{K}\sum_{j=1}^{K} \text{E}\left[\beta_i\beta_j\right]\text{Cov}(\vartheta_i, \vartheta_j) \\ &= \frac{s^2}{r_0 + r} + \sum_{i=1}^{K}\sum_{j=1}^{K} \text{Cov}(\vartheta_i, \vartheta_j)\left[\text{E}(\beta_i)\text{E}(\beta_j) + \text{Cov}(\beta_i, \beta_j)\right] \\ &\approx \frac{s^2}{r} + \sum_{i=1}^{k} \nabla_{\boldsymbol{\theta}_i} h(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}}, \hat{\boldsymbol{\beta}}_r)\boldsymbol{\Sigma}_{in_i}\nabla_{\boldsymbol{\theta}_i} h(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}}, \hat{\boldsymbol{\beta}}_r)^{\text{T}} \\ &\qquad + \sum_{i=1}^{K}\sum_{j=1}^{K} \frac{[\boldsymbol{\Sigma}_{\boldsymbol{\beta}}]_{ij}}{r}\left[\boldsymbol{\Sigma}_{\boldsymbol{n}}\right]_{ij} \\ &\stackrel{def}{=} V_{\text{stoch}} + V_{\text{par}} + V_{\text{resp}} \end{aligned} \tag{13}$$

The last equality defines V_{resp}. The preceding equalities depend upon independence and asymptotic normality assumptions. The term $V_{\text{stoch}} = s^2/r$ in the second approximation (with $r_0 = 0$) follows directly from plugging in s^2 for σ^2 and examining the posterior conditional variance of β_0 in Equation (12). Cheng and Holland (2004) also found an estimator of V_{stoch} that is proportional to $1/r$.

Cheng and Holland (2004) describe several methods to estimate the gradients of a stochastic simulation model with an unknown response model, and use those results to approximate V_{stoch} and V_{par}. A full study of the merits of the many ways to estimate the gradient $\boldsymbol{\beta}$, and their influence upon V_{tot}, is an area for further study.

Since $\text{Cov}(\beta_i, \beta_j)$ is inversely proportional to the number of runs, the asymptotic approximation for all uncertainty if m_0 more replications and m_i data samples for each input are collected can be approximated (cf. Equation (8)) by

$$V_{\text{tot}}(\boldsymbol{m}, m_0) = \frac{s^2}{r + m_0} + \sum_{i=1}^{k} \frac{\xi_i}{n_{0i} + n_i + m_i} + \frac{a_i}{(r + m_0)(n_{0i} + n_i + m_i)}, \tag{14}$$

where a_i represents the sum of the subset of the terms in $\sum_{l=1}^{K} \sum_{j=1}^{K} [\boldsymbol{\Sigma}_{\boldsymbol{\beta}}]_{lj} \left[\boldsymbol{H}_{\boldsymbol{n}}^{-1}\right]_{lj}$ that have both l and j indexing components of the overall parameter vector that deal with $\boldsymbol{\theta}_i$, for $i = 1, \ldots, k$. The double sum decouples into one term per input parameter vector due to the block diagonality of $\boldsymbol{\Sigma}_{\boldsymbol{n}}$. Derivatives confirm that $V_{\text{tot}}(\boldsymbol{m}, m_0)$ is convex and decreasing in m_0 and each m_i.

Suppose there is a sampling constraint for more field data and simulation runs,

$$B = c_0 m_0 + \sum_{i=1}^{k} c_i m_i.$$

By Proposition 4.2, the $\boldsymbol{m}$ that minimizes $V_{\text{tot}}(\boldsymbol{m}, m_0)$ for a fixed $m_0 < B/c_0$, subject to the sampling constraint, can be found by substituting B with $B - c_0 m_0$, and replacing ξ_i with $\xi_i + a_i/(r + m_0)$ for $i = 1, 2, \ldots, k$ in Equation (10). The optimal $(\boldsymbol{m}, m_0)$ can then be found by univariate search for the optimal m_0. In practice, one might run more replications if $V_{\text{stoch}} + V_{\text{resp}}$ is large relative to $V_{\text{par}}(\boldsymbol{m})$, and do more data collection if $V_{\text{par}}(\boldsymbol{m})$ is unacceptably large.

5 Experimental Analysis

This section presents numerical studies to assess how close the asymptotic approximations are to the actual variance. It also assesses the variance, bias, and root mean squared error (RMSE) of estimators of the unknown mean performance as a function of both the number of initially collected samples of data and the total number of additional samples to be collected. We evaluate the influence of response surface uncertainty for a quadratic response in Section 5.1, consider a known nonlinear function with unknown parameters in Section 5.2, and apply the ideas to the ICU simulation in Section 5.3. In all three examples, we assume noninformative prior distributions for the unknown parameters, and initial data collection is complemented with a follow-up data collection plan based upon Equation (7) or Equation (14).

5.1 Unknown Nonlinear Response

The first numerical experiment assesses the reduction in uncertainty about the mean performance assuming that the response surface is unknown. Assume that the unknown parameters are the mean μ_i and precision λ_i of $k = 3$ sources of randomness for which data $x_{i\ell}$ has been observed

$(i = 1, 2, 3;\ \ell = 1, 2, \ldots, n_i)$. The response function is quadratic in $\boldsymbol{\Theta} = (\mu_1, \lambda_1, \mu_2, \lambda_2, \mu_3, \lambda_3)$, with known functional form but unknown response parameters.

Figure 2 summarizes the experiment. Based on initial data collection, the input parameters of the k normal distributions are estimated. The "true" response is

$$Y = g(\boldsymbol{\Theta}) + \sigma Z = \beta_0 + \left(\sum_{i=1}^{k} \beta_{2i-1}\mu_i + \beta_{2i}\lambda_i \right) + \beta_7 \mu_1 \mu_2 + \beta_8 \mu_1^2 + \sigma Z$$

where $\boldsymbol{\beta} = (0, -2, 1, -2, 1, 1, 1, 1, 1)$, $\boldsymbol{\mu} = (2, 10, 2)$, $\boldsymbol{\lambda} = (1, 2, 1)$, $\sigma^2 = 1/2$. The vector $(\boldsymbol{\beta}, \sigma^2)$ is estimated from independent model output tested with $q = 1$ replication at $t = 29$ design points from a central composite design (CCD, Box and Draper 1987, with a 2^{6-2} design for the factorial portion, 2×6 star points with $\alpha = 1$, and one center point, for 16+12+1=29 points total). Design points were taken to be the MAP and endpoints of a 95% credible set for the six parameters (using a normal distribution approximation). Follow-up data collection and simulation run allocations from Equation (14) are used to reduce input and response parameter uncertainty.

For the inference process, we used a noninformative prior distribution $\pi_i(\mu_i, \lambda_i) \propto \lambda_i^{-1}$ for $i = 1, 2, 3$. After n observations for a given source of randomness, and dropping the i momentarily for notational clarity, this results in normal gamma posterior pdf $f_n(\mu, \lambda \mid \mathcal{E}_{\boldsymbol{n}}) = \frac{(n\lambda)^{\frac{1}{2}}}{\sqrt{2\pi}} e^{-\frac{n\lambda}{2}(\mu - \tilde{\mu}_n)^2} \frac{\nu_n^{\alpha_n}}{\Gamma(\alpha_n)} \lambda^{\alpha_n - 1} e^{-\nu_n \lambda}$, where $\tilde{\mu}_n = \sum_{\ell=1}^{n} x_\ell / n$, $\alpha_n = (n-1)/2$, and $\nu_n = \sum_{\ell=1}^{n} (x_\ell - \tilde{\mu}_n)^2/2$ (Bernardo and Smith 1994). This implies that the posterior means of an unknown mean μ is the sample average $\tilde{\mu}_n$, and the posterior marginal distribution of λ is a Gamma (α_n, ν_n) distribution, with MAP $\tilde{\lambda} = (\alpha_n - 1)/\nu_n = (n-3)/(2\nu_n)$ and posterior variance $\mathrm{Var}[\Lambda \mid \mathcal{E}_{\boldsymbol{n}}] = \alpha_n / \nu_n^2$.

We estimated the mean of the variance approximation, $\mathrm{E}_{\mathcal{E}_{\boldsymbol{n}}}[V_{\mathrm{par}}(\boldsymbol{m}^*)]$, the expectation taken over the sampling distribution of the initial data samples, and with $\boldsymbol{m}^*$ computed for $B = 0, 10, \ldots, 100$; and the variance $\mathrm{Var}_{\mathcal{E}_{\boldsymbol{n}}, \mathcal{D}}[h(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}+\boldsymbol{m}^*}, \tilde{\boldsymbol{\beta}}_{\boldsymbol{n}+\boldsymbol{m}^*})]$ and bias $\kappa = \mathrm{E}_{\mathcal{E}_{\boldsymbol{n}}, \mathcal{D}}[h(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}+\boldsymbol{m}^*}, \tilde{\boldsymbol{\beta}}_{\boldsymbol{n}+\boldsymbol{m}^*})] - g(\boldsymbol{\Theta})$ of the overall estimator of the mean, determined with respect to the sampling distributions of both rounds of sampling. Those samples determine the estimates $\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}+\boldsymbol{m}^*}$, $\tilde{\boldsymbol{\beta}}_{\boldsymbol{n}+\boldsymbol{m}^*}$ and best local linear approximation $h(\boldsymbol{\Theta}, \tilde{\boldsymbol{\beta}}_{\boldsymbol{n}+\boldsymbol{m}^*})$ to $g(\boldsymbol{\Theta})$ near $\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}+\boldsymbol{m}^*}$. Estimates are based on 1000 macroreplications of data collection experiments for each combination of the number of initial samples ($n_i = n = 10, 20, \ldots, 100$ for $i = 1, \ldots, k$).

Figure 3 shows that $\mathrm{E}_{\mathcal{E}_{\boldsymbol{n}}}[V_{\mathrm{par}}(\boldsymbol{m}^*)]$ and $\mathrm{Var}_{\mathcal{E}_{\boldsymbol{n}}, \mathcal{D}}[h(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}+\boldsymbol{m}^*}, \tilde{\boldsymbol{\beta}}_{\boldsymbol{n}+\boldsymbol{m}^*})]$ track each other reasonably well, which is expected since the response is a low order polynomial. The allocations tended to allocate more samples from inputs with a larger variance and/or larger influence on the output (intuitive, so data not shown). In spite of the relatively close match of variance, Figure 4 indicates the fraction m_1^*/B of samples allocated to learning about the major quadratic term (μ_1, λ_1) varies as a function of B and n more significantly when there is less information about the response parameters (small $r = qt$). This indicates that variability in the regression estimates may affect the optimal data collection allocations.

Figure 5 illustrates how significantly deleterious ignoring parameter uncertainty can be for the coverage probability and the mean half-width size of confidence intervals. For this example, constructing 95% CI with stochastic uncertainty alone (using $z_{97.5}\sqrt{V_{\mathrm{stoch}}}$) results in an extremely poor coverage probability (about 0 to 2% rather than 95%) because input parameters are poorly identified. The wider CI half-widths based on $z_{97.5}\sqrt{V_{\mathrm{tot}}}$ result in much better coverage (empirical coverage of 93-95%, very close to the nominal 95% level, when $n_i = n = 20, 40, 60, 80$ data points were initially collected and when $B = 0$). For a fixed B, the coverage is worst when the least amount of initial data is available (smaller n). For a fixed n, the coverage declined somewhat with increasing B, dropping to 87-89% when $B = 1000$ and down further to 77-82% when $B = 4000$. For

1. For $i = 1, 2, \ldots, 1000$ macroreplications

   (a) Sample n conditionally independent, normally distributed samples ($k = 3$ sources of randomness), $x_{j\ell} \sim \text{Normal}(\mu_j, \lambda_j^{-1})$, for $j = 1, \ldots, k$; $\ell = 1, 2, \ldots, n$.

   (b) Determine MAP estimators $\tilde{\mu}_i = \sum_{\ell=1}^{n_i} x_{i\ell}/n_i$, $\tilde{\lambda}_i$ given the samples from Step (a).

   (c) Generate q independent samples of $y_{(l-1)q+j} = g(\boldsymbol{\Theta}_{(l-1)q+j})$ for $j = 1, \ldots, q$, at each design point $\boldsymbol{\Theta}_l$, for $l = 1, 2, \ldots, t$. The design points were on the lattice generated by the MAP and MAP$\pm z_{0.95}$SE for each of the $2k$ components of the parameter vector (the unknown μ_i, λ_i).

   (d) Estimate the response surface parameters $\boldsymbol{\beta}, \sigma^2$, using Equation (12).

   (e) Compute the optimal sampling allocations $\boldsymbol{m}_{ji}^*, m_{0i}^*$ for the ith macroreplication to minimize the approximation $V_{\text{tot}}(\boldsymbol{m}, m_0)$ in Equation (14) (given sampling budget B with costs $c_0 = c_j = 1$, for $j = 1, \ldots, k$).

   (f) Generate m_{0i}^* more additional runs at the simulation design points and reestimate the response surface (update $\tilde{\boldsymbol{\beta}}, \tilde{\sigma}$).

   (g) Observe $\boldsymbol{m}_i^*$ additional data points then update the estimates $\tilde{\mu}_j, \tilde{\lambda}_j$ for $j = 1, 2, 3$, and $\tilde{y}_i = h(\tilde{\boldsymbol{\Theta}}, \tilde{\boldsymbol{\beta}})$ (i.e. the estimate of the response function $g(\boldsymbol{\Theta})$).

2. Compute the sample mean $\bar{\tilde{y}}_n = \sum_{i=1}^{1000} \tilde{y}_i/1000$, sample bias $\bar{\kappa}_n$, sample variance S_n^2 of the $\tilde{y}_i$, and average number of additional samples $\bar{\boldsymbol{m}}$.

Figure 2: Algorithm to assess variance, bias, and RMSE for the response surface example in Section 5.1, assuming $\boldsymbol{\Theta}_r = (\mu_1, \lambda_1, \ldots, \mu_k, \lambda_k)$ and $Y_r = g(\boldsymbol{\Theta}_r) + \sigma Z_r$.

fixed n, we hypothesize that the coverage degrades somewhat as B increases because the allocations and response surface estimates may be influenced by errors in the gradient estimators. In spite of the degradation in coverage for large B, the coverage is still orders of magnitude better than the coverage that considers stochastic uncertainty alone. Note that $z_{97.5}\sqrt{V_{\text{tot}}}$ does not depend strongly upon n for large B in this example because V_{resp} remains a factor in the uncertainty. That term could be reduced by running additional replications at each design point.

Similar results were obtained for the first-order bias corrected estimators, for a different values of the response gradients, and for the 90% confidence level.

5.2 *M/M/1* Queue

We assessed the influence of input uncertainty about the arrival rate λ and service rate μ on the stationary mean $g(\lambda, \mu) = \lambda/(\mu - \lambda)$ for an *M/M/1* queue. Noninformative prior distributions are used for the unknown arrival and service parameters so that $\pi_\Lambda(\lambda) \propto \lambda^{-1}$ and $\pi_\mu(\mu) \propto \mu^{-1}$.

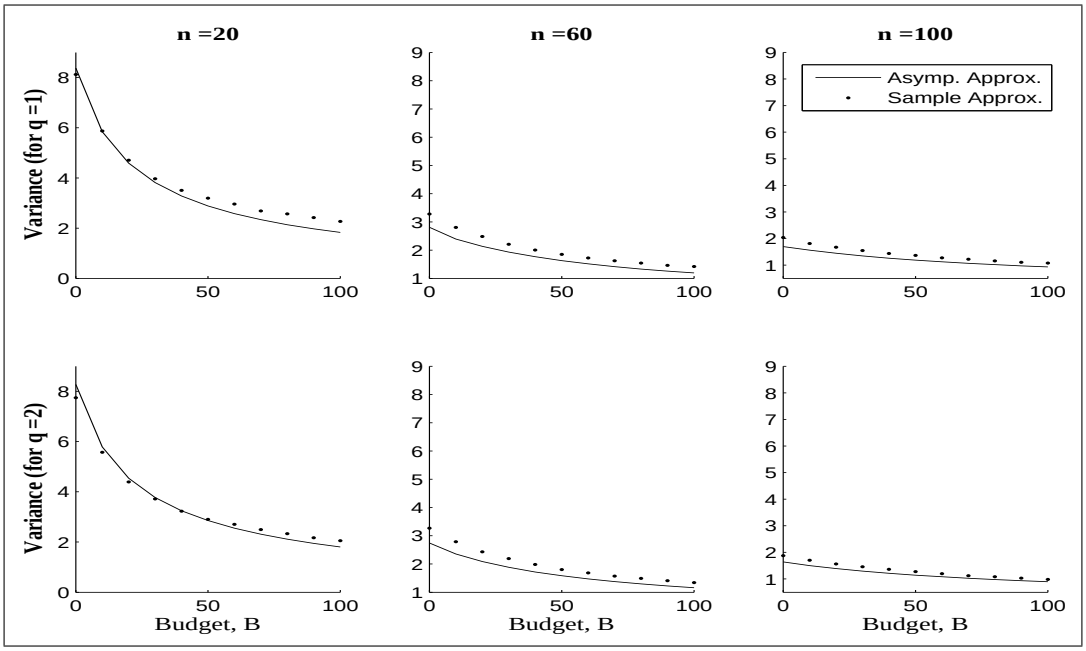

Figure 3: Comparison of mean asymptotic variance approximation $\mathrm{E}_{\mathcal{E}_{\boldsymbol{n}}}[V_{\mathrm{par}}(\boldsymbol{m}^*)]$ (labeled "Asymp. Approx.") with realized sample variance estimate of $\mathrm{Var}_{\mathcal{E}_{\boldsymbol{n}},\mathcal{D}}[h(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}+\boldsymbol{m}^*}, \tilde{\boldsymbol{\beta}}_{\boldsymbol{n}+\boldsymbol{m}^*})]$ (labeled "Sample Approx."), if q iterations are taken from each of $t = 29$ design points in a CCD.

After observing n simulated arrival times and service times, $\mathcal{E}_{\boldsymbol{n}} = (a_1, a_2, \ldots, a_n, s_1, s_2, \ldots, s_n)$, the modeler's posterior distribution for Λ is Gamma $(n, \sum_{\ell=1}^{n} a_\ell)$ (Bernardo and Smith 1994). A similar posterior distribution holds for μ. Interarrival times and service times were simulated from the "true" exponential(λ) and exponential(μ) distributions respectively.

Additional data $\mathcal{D}$ collected from $\boldsymbol{m}^* = (m_a^*, m_s^*)$ optimally-allocated follow-up samples were then simulated. We compared the mean of the asymptotic variance approximation $\mathrm{E}_{\mathcal{E}_{\boldsymbol{n}}}[V_{\mathrm{par}}(\boldsymbol{m}^*)]$ predicted after the initial samples were collected but before the follow-up samples were collected (averaged over realizations of the initial n data observations) with the variance of the estimated mean $\mathrm{Var}_{\mathcal{E}_{\boldsymbol{n}},\mathcal{D}}[g(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}+\boldsymbol{m}^*})]$ (expectation with respect to both stages of sampling that estimate $\boldsymbol{\Theta} = (\lambda, \mu)$) and computed the overall bias $\kappa = \mathrm{E}_{\mathcal{E}_{\boldsymbol{n}},\mathcal{D}}[g(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}+\boldsymbol{m}^*})] - g(\boldsymbol{\Theta})$ after both rounds of sampling. Because expectations may not exist, we conditioned on the utilization being less than 0.995. Estimates are based on 1000 macroreplications of data collection experiments as in Figure 6, for each combination of the number of initial samples ($n = 20, 40, \ldots, 100$), follow-up samples ($B = m_a^* + m_s^* = 0, 10, \ldots, 100, 200, \ldots, 1000$), and "true" utilizations ($\lambda/\mu = 0.2, 0.4, 0.6, 0.8$ with $\mu = 1$).

Several qualitative points that can be made. One, the mean asymptotic variance approximation $\mathrm{E}_{\mathcal{E}_{\boldsymbol{n}}}[V_{\mathrm{par}}(\boldsymbol{m}^*)]$ better matches the sample variance estimate of $\mathrm{Var}_{\mathcal{E}_{\boldsymbol{n}},\mathcal{D}}[g(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}+\boldsymbol{m}^*})]$ in areas where g is more linear (low utilization) and/or the number of initial samples n is larger (Figure 7). The approximation can be considered very useful when g is relatively linear in the area where most of the posterior probability of the parameters is found. For higher utilization levels and fewer data points, the asymptotic approximation for variance and for bias was higher than the empirically observed variance and bias.

Two, the variance, bias, and MSE of the plug-in estimator $g(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}+\boldsymbol{m}^*})$ was still reduced even when the approximation was poor. For example, the solid line in Figure 8 predicts how much first-order bias is expected, given that n observations of arrival and service times have been ob-

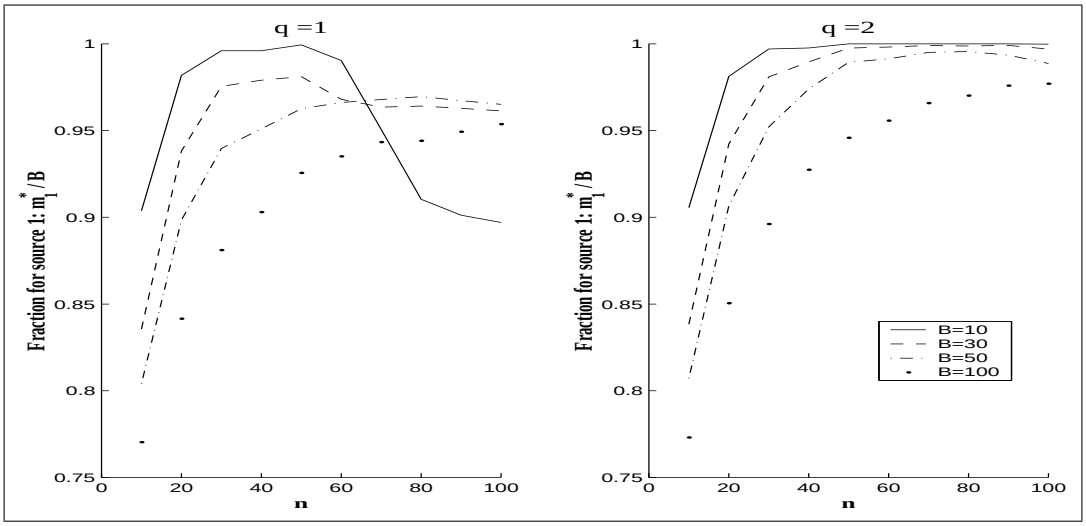

Figure 4: Fraction of budget for sampling from source of randomness 1, if q iterations are taken from each of $t = 29$ design points in a CCD.

served, and another m_a^*, m_s^* observations are expected (based on the variance-reducing solution in Proposition 4.2). The first-order bias estimate is given in Appendix A. The dotted line represents the empirically observed bias after those additional observations were collected, and the updated plug-in estimator is used. The bias is low and the predicted bias is proportional to the estimated bias when the response surface is linear (note the scale in the row for $\rho = 0.2$). The approximation is worse where the response surface is nonlinear, but an improvement in bias is still observed with more data.

Three, the optimal allocation for variance reduction samples both service times and arrival times equally, but the optimal MSE allocation tends to allocate more samples for service times, particularly with a small budget for additional samples or a high utilization (Figure 9).

Four, the MSE allocation does slightly better than the variance allocation in terms of bias and MSE, but not significantly so. Including the bias term in the optimal sampling allocation therefore added to the computing time to determine the optimal allocation, but did not significantly improve the statistics of any estimates (data not shown, graphs are so similar to the variance allocation graphs).

Five, the function g is particularly nonlinear for larger values of $\hat{\rho}$, so the estimator $g(\hat{\lambda}, \hat{\mu})$ is biased. We also checked the variance, bias, and MSE of the first-order bias-corrected estimator (the plug-in estimator $g(\hat{\lambda}, \hat{\mu})$ minus the bias correction from Equation (15) in Section A) and found that the bias correction actually hurt the variance, bias and MSE of $g(\hat{\lambda}, \hat{\mu})$ because the first-order bias correction fared particularly poorly when $\hat{\rho} > 0.92$. We therefore recommend the plug-in estimator and asymptotic variance sampling allocation for the *M/M/1* queue, and by extension to other functions where the first-order bias correction is poor. We do not replicate the poor confidence interval coverage for the *M/M/1* queue already demonstrated (Barton and Schruben 2001) in related work.

5.3 Critical Care Facility

The critical care facility described in Section 2 can be used to illustrate the above analysis on a practical problem when the response surface is unknown. Suppose initial field data is collected,

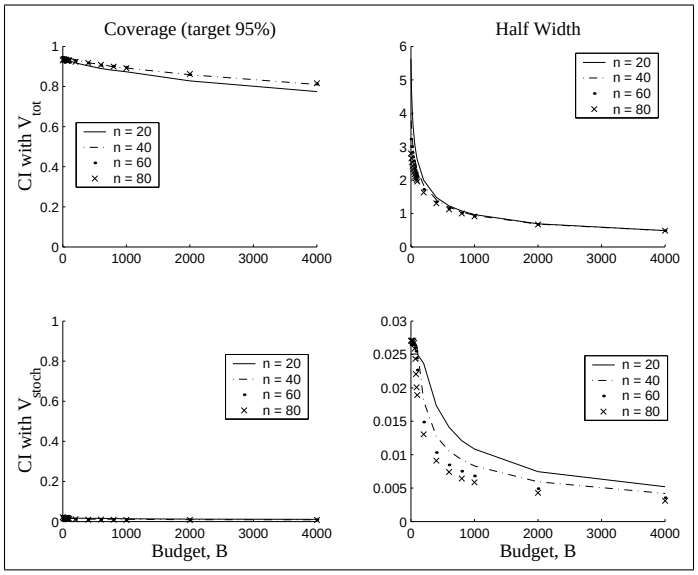

Figure 5: Empirical coverage and half-width of CIs constructed with V_{tot} (accounts for input and output uncertainty) and V_{stoch} (output uncertainty only).

resulting in asymptotically normal approximations $\boldsymbol{\theta}_i \sim \text{Normal}\left(\hat{\boldsymbol{\theta}}_{in_i}, \boldsymbol{\Sigma}_{in_i}\right)$ for each of the $k = 6$ input parameters. For the unknown arrival rate λ for a Poisson number of arrivals per day, we used a noninformative prior distribution, $\pi(\lambda) \propto \lambda^{-1/2}$, resulting in a Gamma $(1/2 + \sum_{\ell=1}^{n} x_\ell, n)$ posterior distribution after observing n daily counts of arrivals (Bernardo and Smith 1994). So $\tilde{\lambda} = (-1/2 + \sum_{\ell=1}^{n} x_\ell)/n$. The prior distribution for the unknown routing probabilities is given in Section 3. The prior distributions for the lognormal service time parameters were chosen identical to those for the unknown normal distribution parameters in Section 5.1.

We used a BMA to sample 32 independent sets of input parameters for the critical care simulation. For each set of input parameters, we ran $q = 4$ independent replications of the critical care unit, for a total of $r = 128$ replications. Each replication simulated 50 months of operation (after a 10 month warm-up period).

Assuming the linear model in Equation (11), the simulation output implies that $\tilde{\boldsymbol{\beta}} = (28.9, 1.52, -0.29, 29.4, -0.57$ and $\tilde{\sigma}^2 = 1.078$. This results in $V_{\text{par}} \approx 98.1$ and approximations for the variance in the performance due to parameter uncertainty of $V_{\text{resp}} \approx 0.99$. The uncertainty due to unknown parameters greatly outweighs the response parameter uncertainty, which in turn outweighs the stochastic uncertainty from the random output $V_{\text{stoch}} \approx \tilde{\sigma}^2/128 = 0.0084$.

The mean allocations for further data collection suggested by minimizing Equation (13) subject to nonnegative sampling constraints for a variety of sampling budgets are shown in Table 2. The table presumes that a total of B data points can be collected, that the collection cost for each area is the same ($c_i = 1$ for $i = 1, \ldots, k$), and that each "real" data sample is 4 times as expensive as running one more replication at each of the original 32 design points ($c_0 = 1/(4 \times 32)$). The most important area for additional data collection, based on this analysis, is the arrival rate. Uncertainty about routing probabilities and intermediate ICU service times are of secondary importance. Simulation replications may help reduce performance uncertainty due to response surface uncertainty. Uncertainty about the other input parameters play a much less important role for output uncertainty.

The specific mechanism for estimating the gradient in this example did not significantly affect

1. For $i = 1, 2, \ldots, 1000$ macroreplications

   (a) Sample n independent exponential(λ) inter-arrival times and n independent exponential(μ) service times to obtain MAP estimates $\tilde{\mu}_i$, $\tilde{\lambda}_i$ (resampling until $\tilde{\lambda}_i/\tilde{\mu}_i < .995$), assuming a noninformative prior distribution $p(\lambda, \mu) \propto \lambda^{-1}\mu^{-1}$.

   (b) Given a sampling budget B with costs $c_i = 1$, compute the optimal sampling allocations $m^*_{a,\mathrm{v}}, m^*_{s,\mathrm{v}}$ to minimize the variance approximation $V_{\mathrm{par}}(\boldsymbol{m})$ and optimal allocations $m^*_{a,\mathrm{mse}}, m^*_{s,\mathrm{mse}}$ to minimize the MSE approximation (see Appendix A).

   (c) Collect additional input samples to obtain updated estimates $\tilde{\lambda}_{i,\mathrm{v}}, \tilde{\mu}_{i,\mathrm{v}}$, and $y_{i,\mathrm{v}} = g(\tilde{\lambda}_{i,\mathrm{v}}, \tilde{\mu}_{i,\mathrm{v}})$ for the variance allocation, and $\tilde{\lambda}_{i,\mathrm{mse}}, \tilde{\mu}_{i,\mathrm{mse}}$, and $y_{i,\mathrm{mse}} = g(\tilde{\lambda}_{i,\mathrm{mse}}, \tilde{\mu}_{i,\mathrm{mse}})$ for the MSE allocation.

2. Compute the sample mean $\bar{y}_{n,\mathrm{v}}$ to estimate $g(\lambda, \mu)$, the sample bias $\bar{\kappa}_{n,\mathrm{v}}$, and the sample variance $S^2_{n,\mathrm{v}}$ of the output, as well as the average number of additional samples $\bar{m}^*_{a,\mathrm{v}}, \bar{m}^*_{s,\mathrm{v}}$ for the variance-based allocation (subscript $_{\mathrm{v}}$), and analogous statistics $\bar{y}_{n,\mathrm{mse}}$, $\bar{\kappa}_{n,\mathrm{mse}}$, $S^2_{n,\mathrm{mse}}$, $\bar{m}^*_{a,\mathrm{mse}}, \bar{m}^*_{s,\mathrm{mse}}$ for the MSE-based allocation (subscript $_{\mathrm{mse}}$).

Figure 6: Algorithm to the assess variance, bias, and RMSE for the *M/M/1* queue example in Section 5.2 for a given λ, μ and sampling parameters n, B.

the data collection allocation that was ultimately suggested. Kleijnen (2001) notes that spurious regressors in parametric regression can deteriorate the predictor, so a parsimonious metamodel is desirable. We repeated the experiment with another technique for response surface modeling due to Raftery, Madigan, and Hoeting (1997) that identifies the most important response parameters (effectively screening out β_j that are relatively close to 0). That analysis determined essentially the same sampling plans in Table 2. We did not check the CI coverage for this example since the true performance is unknown.

6 Comments and Conclusions

Stochastic models are useful system design and analysis tools. The mean system performance depends on statistical parameters that describe the model. There are several ways to handle parameter uncertainty. One is to design a system to handle "worst case" scenarios, but this may be cost ineffective. Another way is to develop robust designs that perform well over a range of parameter values. This begs the question of what range to consider. A reduction of parameter uncertainty can reduce the range of parameters for which the design must perform well. When possible, a third way is to learn about parameters as the system operates and adapt the operation of the system to the extent possible as parameters are inferred, given the existing design of a system (e.g. as did Scarf (1959) for the news vendor problem).

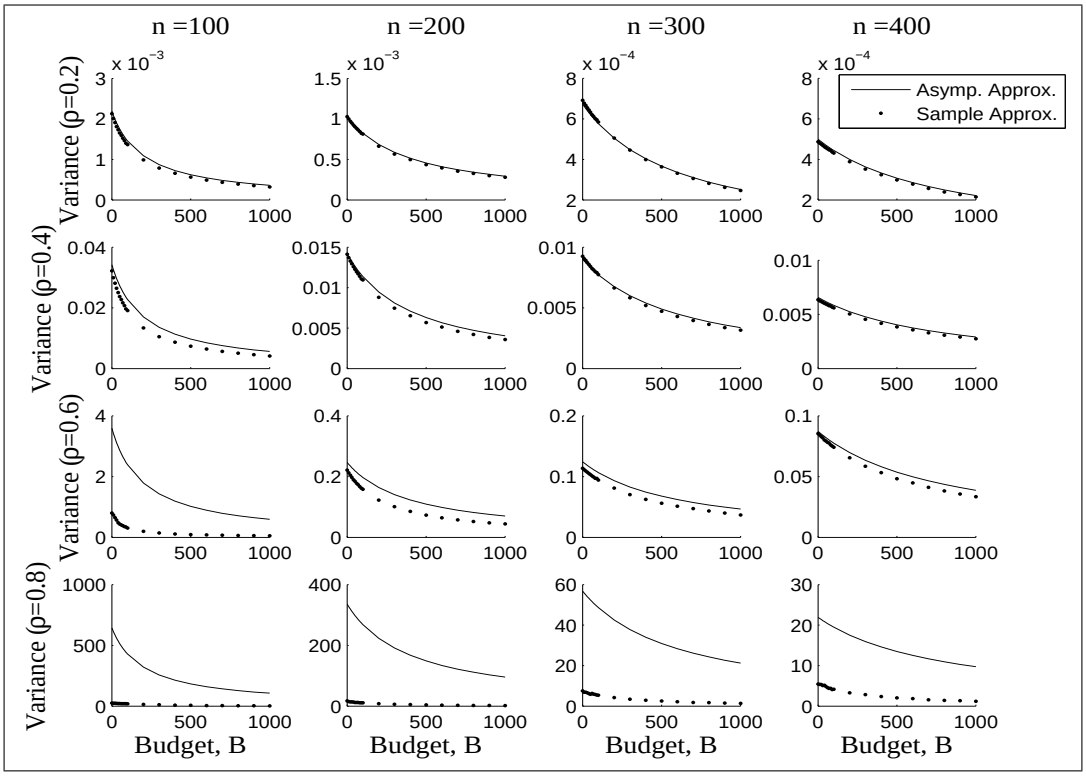

Figure 7: Comparison of mean asymptotic variance approximation $\mathrm{E}_{\mathcal{E}_{\boldsymbol{n}}}[V_{\mathrm{par}}(\boldsymbol{m}^*)]$ ("Asymp. Approx.") with realized sample variance estimate of $\mathrm{Var}_{\mathcal{E}_{\boldsymbol{n}},\mathcal{D}}[g(\tilde{\boldsymbol{\Theta}}_{\boldsymbol{n}+\boldsymbol{m}^*})]$ ("Sample Approx.").

Another approach, taken by this paper, assumes that a number of parameters may need estimation, and that data sampling is possible to better infer their values before a final design is chosen. Asymptotic approximations simplify the problem, and closed form solutions can be found for a broad class of input distribution models. One advantage of this approach is its generality for input models, and the ability to use a variety of gradient estimation tools, including but not limited to regression with the BMA, a natural tool for exploring input parameter uncertainty. Our approach was Bayesian, but some frequentist results emerge as corollaries.

Several issues warrant further exploration. The asymptotic variance approximation is poor for some systems, including the *M/M/1* queue where even the posterior mean of the system performance may not exist. The nonexistence of moments might be avoided by making the model more realistic (assume capacitated queues, run transient rather than steady state simulations), but nonnormal approximations may be of use. Also, some Bayesian representations of parameter uncertainty may have no obvious sampling mechanism for further inference. Nonetheless, the formulas in this paper present an implementable mechanism for guiding data collection plans to reduce parameter uncertainty in a way that in some sense effectively reduces performance uncertainty.

APPENDIX

A Reducing MSE with Bias Estimates

The paper principally discusses sampling plans to reduce an approximation to the variance of the unknown performance, based on parameter variance and gradient information. The linear,

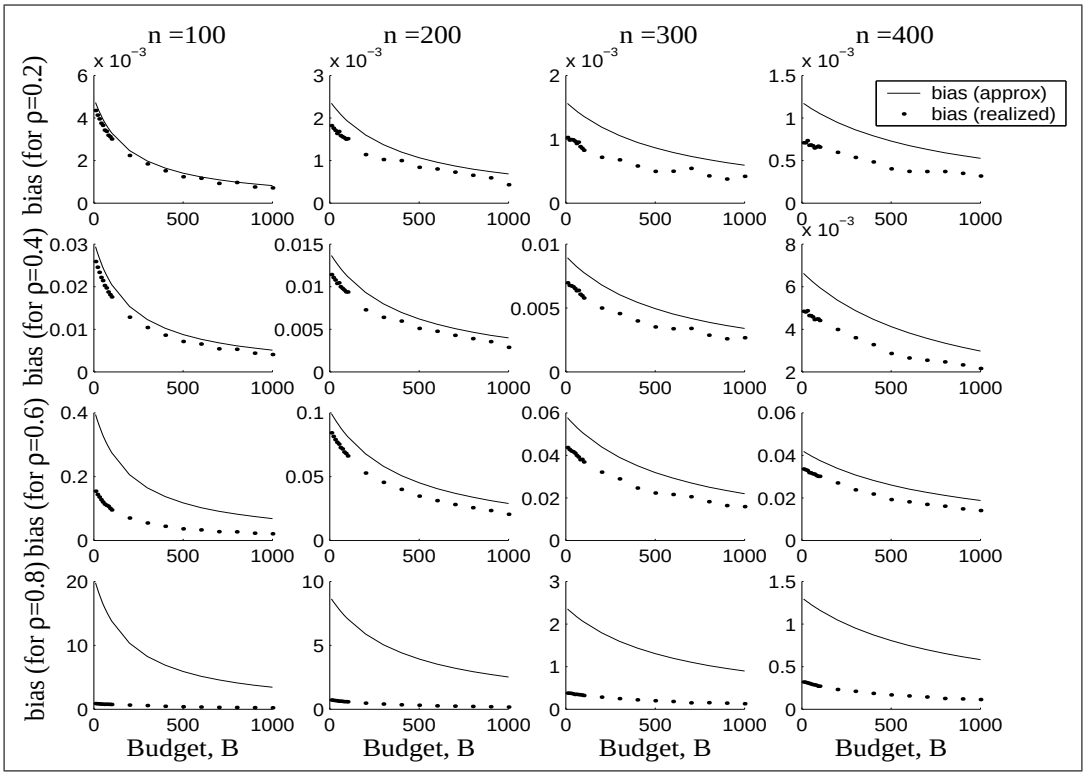

Figure 8: Bias comparison using the optimal allocation for variance, measuring the predicted bias and the empirical bias.

gradient-based approximation of the response function ignores bias introduced by naive estimates of the mean due to any nonlinearity of $g(\cdot)$. Bias increases the mean squared error. In particular, if $\boldsymbol{\theta} = (\vartheta_1, \ldots, \vartheta_d)$ is a single parameter for which n data points have been observed, and $\boldsymbol{\Sigma}_n$ is as in Equation (3), then the bias $\mathrm{E}[g(\tilde{\boldsymbol{\theta}}_n)] - g(\boldsymbol{\theta})$ is approximately

$$\gamma_n \;=\; \frac{1}{2}\sum_{j=1}^{d}\sum_{l=1}^{d}\frac{\partial^2 g(\tilde{\boldsymbol{\theta}}_n)}{\partial\vartheta_j\partial\vartheta_l}\mathrm{Cov}(\vartheta_j,\vartheta_l) = \frac{1}{2}\,\mathrm{tr}\left(\nabla^2_{\boldsymbol{\theta}} g(\tilde{\boldsymbol{\theta}}_n)\boldsymbol{\Sigma}_n\right), \tag{15}$$

a well-known result obtained with a Taylor series expansion with linear and quadratic terms. The MSE is (asymptotically) approximately $V_{\mathrm{par}} + \gamma_n^2$. Informally, as n increases, V_{par} decreases like $1/n$ and the bias-squared term decreases like $1/n^2$.

If $V_{\mathrm{par}}(\boldsymbol{m})$ simplifies to Equation (8) with $k \geq 1$ input parameters then the asymptotic estimate for MSE after $\boldsymbol{m}$ samples are collected simplifies.

$$\begin{aligned}
\mathrm{MSE}_n(\boldsymbol{m}) &= V_{\mathrm{par}}(\boldsymbol{m}) + \gamma_n^2(\boldsymbol{m}) \\
\gamma_n(\boldsymbol{m}) &= \frac{1}{2}\left(\mathrm{tr}\sum_{i=1}^{k}\nabla^2_{\boldsymbol{\theta}_i} g(\tilde{\boldsymbol{\Theta}}_n)\left(\boldsymbol{\Sigma}^{-1}_{in_i} + (m_i \boldsymbol{H}_i(\tilde{\boldsymbol{\theta}}_{in_i})\right)^{-1}\right) \\
&= \sum_{i=1}^{k}\frac{\zeta_i}{n_{0i}+n_i+m_i}
\end{aligned} \tag{16}$$

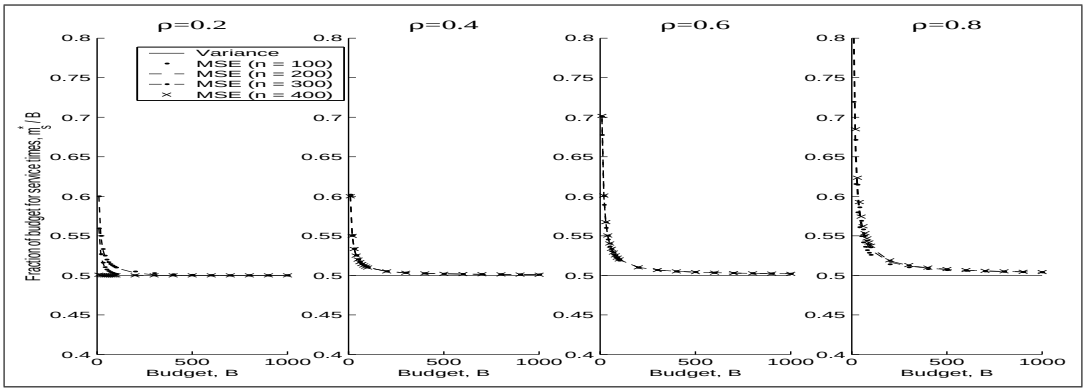

Figure 9: Average fraction of samples $\bar{m}_s^*/B$ allocated for service times for the variance allocation (always 50%) and the MSE optimal allocation for various n).

Table 2: Predicted Input Uncertainty $V_{\text{tot}}(\boldsymbol{m}, m_0)$ as a Function of Sampling Budget B and Resulting Optimal Sampling Plan (m_1=Days of Arrival Data, m_2=# ICU Service Times, m_3=# Intermediate ICU Service Times, m_4=# Intermediate CCU Service Times, m_5=# CCU Service Times, m_6=# Routing Decisions, $q' = m_0/32$=# Replications per Design Point)

B	m_1	m_2	m_3	m_4	m_5	m_6	q'	$V_{\text{par}}(\boldsymbol{m})$	$V_{\text{tot}}(\boldsymbol{m}, r)$
0	0	0	0	0	0	0	0	98.1	99.1
100	99	0	0	0	0	0	1	53.8	54.3
200	198	0	6	0	0	0	8	38.8	39.2
500	440	0	25	0	0	30	20	23.2	23.3
1000	794	0	78	6	0	115	28	14.0	14.1

where $\zeta_i = \text{tr}[\nabla^2_{\boldsymbol{\theta}_i} g(\tilde{\boldsymbol{\Theta}}_n) \boldsymbol{H}_i^{-1}(\tilde{\boldsymbol{\theta}}_{in_i})]/2$. For the *M/M/1* queue in Section 4.1,

$$\gamma_n(\boldsymbol{m}) = \frac{\tilde{\lambda}}{(\tilde{\mu}-\tilde{\lambda})^3} \frac{\tilde{\mu}^2}{(\alpha_\mu - 1 + m_s)} + \frac{\tilde{\mu}}{(\tilde{\mu}-\tilde{\lambda})^3} \frac{\tilde{\lambda}^2}{(\alpha_\lambda - 1 + m_a)}.$$

B Canonical conjugate prior distribution.

This section justifies the claimed simplification of Equation (4) to Equation (5) when (i) the likelihood is in the regular exponential family and has a finite dimensional sufficient statistic, (ii) the prior distribution for the unknown parameter is a canonical conjugate distribution, and (iii) some differentiability conditions hold. This section fixes typographical errors in Bernardo and Smith (1994, Prop. 5.16).

If X has a likelihood model in the regular exponential family, with d-dimensional parameter $\boldsymbol{\theta} = (\vartheta_1, \ldots, \vartheta_d)$, the likelihood function can be written

$$p\left(x \mid \boldsymbol{\theta}\right) = a_1(x) g(\boldsymbol{\theta}) \exp\left[\sum_{j=1}^{d} c_j \phi_j(\boldsymbol{\theta}) h_j(x)\right] \tag{17}$$

for some $a_1(\cdot)$, $g(\cdot)$, c_j, $\phi_j(\cdot)$, and $h_j(\cdot)$. The conjugate prior distribution for $\boldsymbol{\theta}$ with parameter

$n_0, \boldsymbol{t} = (t_1, \ldots, t_d)$ is

$$p\left(\boldsymbol{\theta} \mid n_0, \boldsymbol{t}\right) \;=\; [K(n_0, \boldsymbol{t})]^{-1} g(\boldsymbol{\theta})^{n_0} \exp\left[\sum_{j=1}^{d} c_j \phi_j(\boldsymbol{\theta}) t_j\right], \tag{18}$$

where $K(n_0, \boldsymbol{t}) = \int g(\boldsymbol{\theta})^{n_0} \exp[\sum_{j=1}^{d} c_j \phi_j(\boldsymbol{\theta}) t_j] d\boldsymbol{\theta} < \infty$.

Suppose that the data $x_1, \ldots, x_n$ are observed. Set $v_{\ell j} = h_j(x_\ell)$ and $\boldsymbol{v}_\ell = (v_{\ell 1}, \ldots, v_{\ell d})$ with sample average $\bar{\boldsymbol{v}}_n = \sum_{\ell=1}^{n} \boldsymbol{v}_\ell / n$. Then $n\bar{\boldsymbol{v}}_n$ is a sufficient statistic for $\boldsymbol{\theta}$ and the posterior distribution of $\boldsymbol{\theta}$ is $p\left(\boldsymbol{\theta} \mid n_0 + n, \boldsymbol{t} + n\bar{\boldsymbol{v}}_n\right)$.

A final reparametrization is useful. Set $\psi_j = c_j \phi_j(\boldsymbol{\theta})$ and $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_d)$. Equation (17) can be rewritten in what is known as canonical conjugate form (Bernardo and Smith 1994, Definition 4.12),

$$p\left(\boldsymbol{v}_\ell \mid \boldsymbol{\psi}\right) = a_2(\boldsymbol{v}) \exp\left[\boldsymbol{v}_\ell \boldsymbol{\psi}^{\mathrm{T}} - b(\boldsymbol{\psi})\right]$$

for some real-valued $a_2(\boldsymbol{v})$, $b(\boldsymbol{\psi})$. The canonical conjugate prior distribution is

$$p\left(\boldsymbol{\psi} \mid n_0, \boldsymbol{t}\right) \;=\; c(n_0, \boldsymbol{t}) \exp\left[n_0 \boldsymbol{t} \boldsymbol{\psi}^{\mathrm{T}} - n_0 b(\boldsymbol{\psi})\right],$$

where $n_0, \boldsymbol{t} = (t_1, \ldots, t_d)$ are parameters of the prior distribution.

Bernardo and Smith (1994, Prop. 5.16) note that the posterior distribution has the same functional form as the prior distribution,

$$p\left(\boldsymbol{\psi} \mid n_0, \boldsymbol{t}, \boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\right) \;=\; p\left(\boldsymbol{\psi} \mid n_0 + n, n_0 \boldsymbol{t} + n\bar{\boldsymbol{v}}_n\right).$$

Further, if $\tilde{\boldsymbol{\psi}}$ is the MAP estimator of $\boldsymbol{\psi}$, then $\left(\boldsymbol{\Sigma}_n^{-1}\right)_{jl} = (n_0 + n) \frac{\partial^2 b(\boldsymbol{\psi})}{\partial \psi_j \partial \psi_l}\big|_{\boldsymbol{\psi} = \tilde{\boldsymbol{\psi}}}$. The claim is proven by noting that the expected information in one observation is

$$\left(\boldsymbol{H}(\boldsymbol{\psi})\right)_{jl} \;=\; \mathrm{E}\left[\frac{\partial^2 b(\boldsymbol{\psi})}{\partial \psi_j \partial \psi_l}\right]\Bigg|_{\tilde{\boldsymbol{\psi}}} = \frac{\partial^2 b(\boldsymbol{\psi})}{\partial \psi_j \partial \psi_l}\Bigg|_{\tilde{\boldsymbol{\psi}}}.$$

The relationship also holds in the $\boldsymbol{\theta}$ coordinate system if the map from $\boldsymbol{\psi}$ to $\boldsymbol{\theta}$ is bijective in a neighborhood of $\tilde{\boldsymbol{\psi}}$.

C Proof of Proposition 4.2.

Using the method of Lagrange multipliers, we obtain the Lagrangian function

$$\begin{aligned} L(\boldsymbol{m}, \tau) \;&=\; V_{\mathrm{par}}(\boldsymbol{m}) - \sum_{i=1}^{k} \tau\left(c_i m_i - B\right) \\ &=\; \sum_{i=1}^{k} \frac{\xi_i}{n_{0i} + n_i + m_i} - \sum_{i=1}^{k} \tau\left(c_i m_i - B\right), \end{aligned}$$

where τ is the Lagrange multiplier. Set the $(k+1)$ partial derivatives to 0 to get

$$-\frac{\xi_i}{(n_{0i} + n_i + m_i)^2} - \tau c_i \;=\; 0, \text{ for } i = 1, \ldots, k, \text{ and} \tag{19}$$

$$\sum_{i=1}^{k} c_i m_i - B \;=\; 0. \tag{20}$$

Equation (19) implies $\frac{\xi_i/c_i}{(n_{0i}+n_i+m_i)^2} = \frac{\xi_j/c_j}{(n_{0j}+n_j+m_j)^2}$. After some algebra we get

$$c_j(n_{0j} + n_j + m_j) = \left(\frac{\xi_j c_i c_j}{\xi_i}\right)^{1/2} (n_{0i} + n_i + m_i). \tag{21}$$

Equation (20) can be rewritten $\sum_{j=1}^{k} c_j(n_{0j} + n_j + m_j) = B + \sum_{\ell=1}^{k} c_\ell(n_{0\ell} + n_\ell)$. Substitute Equation (21) into this expression to obtain

$$(n_{0i} + n_i + m_i) \sum_{j=1}^{k} \left(\frac{\xi_j c_i c_j}{\xi_i}\right)^{1/2} = B + \sum_{\ell=1}^{k} c_\ell(n_{0\ell} + n_\ell). \tag{22}$$

Algebraic rearranging gives the claimed result for the optimal m_i^*,

$$m_i^* = \frac{B + \sum_{\ell=1}^{k} c_\ell(n_{0\ell} + n_\ell)}{\sum_{j=1}^{k} \left(\frac{\xi_j c_i c_j}{\xi_i}\right)^{1/2}} - (n_{0i} + n_i).$$

When B is sufficiently large, the constraint $m_i^* \geq 0$ is satisfied as desired.

Acknowledgments: We thank Jim Wilson for thorough and helpful feedback.

Bibliography

Barton, R. R. and L. W. Schruben (2001). Simulating real systems. in submission.

Bernardo, J. M. and A. F. M. Smith (1994). *Bayesian Theory*. Chichester, UK: Wiley.

Box, G. and N. Draper (1987). *Empirical Model-Building and Response Surfaces*. New York: Wiley.

Box, G., W. Hunter, and J. Hunter (1978). *Statistics for Experimenters. An Introduction to Design, Data Analysis and Model Building*. New York: Wiley.

Buzacott, J. A. and J. G. Shanthikumar (1993). *Stochastic Models of Manufacturing Systems*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Cheng, R. C. H. and W. Holland (1997). Sensitivity of computer simulation experiments to errors in input data. *J. Statistical Computing and Simulation 57*, 219–241.

Cheng, R. C. H. and W. Holland (2004). Calculation of confidence intervals for simulation output. *ACM Transactions on Modeling and Computer Simulation 14*(4), 344–362.

Chick, S. E. (1997). Bayesian analysis for simulation input and output. In S. Andradóttir, K. Healy, D. Withers, and B. Nelson (Eds.), *Proc. 1997 Winter Simulation Conference*, Piscataway, NJ, pp. 253–260. IEEE, Inc.

Chick, S. E. (2001). Input distribution selection for simulation experiments: Accounting for input uncertainty. *Operations Research 49*(5), 744–758.

Draper, D. (1995). Assessment and propogation of model uncertainty (with discussion). *J. Royal Statistical Society, Series B 57*(1), 45–97.

Edwards, A. W. F. (1984). *Likelihood*. Cambridge University Press.

Gelman, A., J. B. Carlin, H. S. Stern, and D. R. Rubin (1995). *Bayesian Data Analysis*. London: Chapman & Hall.

Kleijnen, J. P. C. (2001). Comments on M.C. Kennedy & A. O'Hagan's 'Bayesian calibration of computer models'. *Journal Royal Statistical Society, Series B 63*(3), in press.

Law, A. M. and W. D. Kelton (2000). *Simulation Modeling & Analysis* (3rd ed.). New York: McGraw-Hill, Inc.

Mendoza, M. (1994). Asymptotic normality under transformations. A result with Bayesian applications. *TEST 3*(2), 173–180.

Myers, R., A. Khuri, and W. Carter (1989). Response surface methodology: 1966-1988. *Technometrics 31*(2), 137–157.

Ng, S.-H. and S. E. Chick (2001). Reducing input distribution uncertainty for simulations. In B. Peters, J. Smith, M. Rohrer, and D. Madeiros (Eds.), *Proc. 2001 Winter Simulation Conference*, Piscataway, NJ, pp. 364–371. IEEE, Inc.

Raftery, A. E., D. Madigan, and J. A. Hoeting (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association 92*(437), 179–191.

Scarf, H. (1959). Bayes solutions of the statistical inventory problem. *Annals of Mathematical Statistics 30*, 490–508.

Schervish, M. J. (1995). *Theory of Statistics*. New York: Springer.

Schruben, L. W. and B. H. Margolin (1978). Pseudorandom number assignment in statistically designed simulation and distribution sampling experiments. *Journal of the American Statistical Association 73*(363), 504–525.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Zouaoui, F. and J. R. Wilson (2001). Accounting for input model and parameter uncertainty in simulation input modeling. In B. Peters, J. Smith, M. Rohrer, and D. Madeiros (Eds.), *Proc. 2001 Winter Simulation Conference*, Piscataway, NJ, pp. 290–299. IEEE, Inc.

Zouaoui, F. and J. R. Wilson (2003). Accounting for parameter uncertainty in simulation input modeling. *IIE Transactions 35*, 781–792.

Zouaoui, F. and J. R. Wilson (2004). Accounting for input-model and input-parameter uncertainties in simulation. *IIE Transactions 36*(11), 1135–1151.