Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

ONLINE QUANTIFICATION OF INPUT UNCERTAINTY FOR PARAMETRIC MODELS

Enlu Zhou
Tianyi Liu

School of Industrial & Systems Engineering
Georgia Institute of Technology
755 Ferst Drive NW
Atlanta, GA 30332, USA

ABSTRACT

It has become increasingly important to assimilate "online data" that arrive sequentially in time for real-time decision. Input uncertainty quantification in stochastic simulation has been developed extensively for batch data that are available all at once, but little has been studied for online data. In this paper, we propose a computationally efficient method to incorporate online data in real time for input uncertainty quantification of parametric models. We show finite-sample bounds and asymptotic convergence for the proposed method, and demonstrate its performance on a simple numerical example.

1 INTRODUCTION

Stochastic simulation is driven by the input model, which is a collection of distributions that model the randomness in the system. The input model is often estimated from data, and therefore the estimation error in the input model introduces the so-called input model uncertainty (or simply as input uncertainty) to the simulation output. It is important to quantify the impact of input uncertainty on the simulation output, since it tells the simulation user how to interpret the output by separating the input model uncertainty from the intrinsic uncertainty of the system itself. To quantify input uncertainty, one needs to start with the input data. Regarding the arriving pattern of input data, there are often two types: 1) "batch data" that are available all at once; 2) "online data" that arrive sequentially in time. Online data occur frequently in many applications, such as customer arrivals in time observed in a service system, and customer demands as the selling season progresses. Business applications nowadays often require to incorporate online data in real time for fast and up-to-date business decisions.

An extensive array of methods have been developed for input uncertainty quantification based on batch data. The methods developed so far in the literature can be roughly grouped into three major categories. First is the frequentist methods that allow nonparametric input distributions and use direct or bootstrap resampling techniques to assess input uncertainty (e.g., Barton and Schruben 1993; Barton and Schruben 2001; Cheng and Holloand 1997). Second is the Bayesian methods (e.g., Chick 2001; Zouaoui and Wilson 2003; Zouaoui and Wilson 2004; Xie et al. 2014) that assume a parametric model and use the posterior distribution of unknown parameters as the sampling distribution during simulation process. Third is the delta method: Cheng and Holloand (1997) uses the delta method to decompose the variance of simulation output into two parts respectively corresponding to stochastic uncertainty and input uncertainty; a robust sensitivity analysis approach developed by Lam (2016) can be viewed as the delta method with respect to a distributional perturbation. Recent advances in stochastic kriging (Ankenman et al. 2010) also give rise to the application of meta-model assisted methods to quantify input uncertainty (e.g., Barton et al. 2013; Xie et al. 2014). Other related work includes Song and Nelson (2015) that proposes a method for quickly assessing the relative contribution of each input distribution to the overall effect of input uncertainty These aforementioned methods take input data in a batch. Although they can be extended to work with online

data by naively repeating the method each time when a new data point comes in, it is obviously not an efficient way because it requires to run new simulation experiments every time when the method is used and these simulations experiments are often computationally expensive. It still remains an open challenge to efficiently work with online data in the domain of input uncertainty quantification.

In this paper we propose a computationally efficient method to assimilate online data for input uncertainty quantification in real time. We assume that the input model takes a parametric form, and take a Bayesian approach to estimate the unknown input parameters from data. When a new data point arrives in the current time stage, the Bayesian posterior distribution on the input parameter is updated. Our main idea is to use importance sampling to transform a set of samples under the previous posterior distribution to a new set of samples under the current posterior, such that the simulation output estimates from previous time stage can be reused and thus new simulation experiments can be avoided. The challenge here is the long time horizon, which can cause sample degeneracy (i.e., most samples have trivial weights close to zero, and only very few samples are non-trivially weighted) if we keep doing importance sampling to track the sequence of posterior distributions over the time. To alleviate sample degeneracy, following the importance sampling step we resample with replacement to generate a new set of samples with equal weights. To avoid the estimation error accumulating over time when using a fixed sample size, we further introduce a restarting mechanism to enhance the performance of the method. Our theoretical and empirical results both show that our proposed method can asymptotically track the true posterior distribution on the output performance measure over long time horizon.

Our proposed method is closely related to particle filtering, also known as sequential Monte Carlo methods (Doucet et al. 2000). Different from the goal of particle filtering to track the unobserved state and/or unknown parameters in a state-space model, our main objective here is to reuse simulation output estimates from previous time stages while estimating the unknown input parameters at the same time. This limits us to the same sample values and consequently causes the estimation error to accumulate over time as shown by our error-bound analysis, which in turn motivates us to introduce a restarting mechanism when necessary. Our method is also related to "green simulation" proposed by Feng and Staum (2015) and Feng and Staum (2017) in the sense of reusing simulation outputs from previous experiments, but they do not consider the accumulated error in the long time horizon, which is our focus. Similar to Xie et al. (2014), we also take a Bayesian perspective to quantify input uncertainty, but our estimation method is different, and more importantly, tailors to the online setting.

2 ONLINE QUANTIFICATION FOR PARAMETRIC INPUT MODELS

Assume the input distribution takes a parametric form $F(\cdot;\theta)$, $\theta \in \Theta \subseteq \mathbb{R}^n$. The true value of the parameter is denoted by θ^c, but it is unknown to us. The system performance is measured by $H = E[h(\xi_{\theta^c})]$, where ξ_{θ^c} is a random variable that represents the stochastic uncertainty in the system and follows the distribution $F(\cdot;\theta^c)$, and the expectation is taken with respect to $F(\cdot;\theta^c)$. For complex systems we usually do not have the analytical form of h, so we estimate the performance measure H by simulation outputs $h(\xi^i_{\hat{\theta}})$'s, where $\hat{\theta}$ is the parameter of the input model that drives the simulation.

The input parameter needs to be estimated from input data. Here we consider the online scenario where the input data arrives one at a time, i.e., at time stage t we observe one data point ξ_t, $t = 1,2,\ldots$, where $\{\xi_t, t = 1,2,\ldots\}$ are independent and identically distributed (i.i.d.) from $F(\cdot;\theta^c)$. We take a Bayesian approach to incorporate data sequentially in time. Specifically, we treat the unknown parameters as a random vector (r.v.) θ and assume a prior distribution π_0. The prior distribution π_0 can be constructed from historical data that were collected prior to time 0. At time stage t, the posterior distribution on θ is $\pi_t \triangleq p(\theta|\xi_1,\ldots,\xi_t)$. Then at the next time stage $t+1$, a new data point ξ_{t+1} comes in, and the posterior distribution is updated according to

$$\pi_{t+1}(\theta) \triangleq p(\theta|\xi_1,\ldots,\xi_{t+1}) = \frac{\pi_t(\theta)p(\xi_{t+1}|\theta)}{\int \pi_t(\theta)p(\xi_{t+1}|\theta)d\theta}.$$

where $p(\xi_{t+1}|\theta)$ is the probability density function (p.d.f.) of $F(\cdot;\theta)$ evaluated at ξ_{t+1}. We also introduce the notation $H(\theta) \triangleq E[h(\xi_\theta)]$, which is a random variable induced by the r.v. θ. The cumulative distribution function (c.d.f.) of the induced posterior distribution on $H(\theta)$ at time t is then defined as

$$G_t(y) \triangleq Pr(H(\theta) \leq y|\xi_1,\ldots,\xi_t).$$

Let's first fix the time t. The posterior distribution G_t represents our knowledge about the performance measure based on all the input data up to time t, and hence G_t provides an uncertainty quantification solely due to the input data without any additional error. The input uncertainty can be quantified by the credible interval (sometimes also called the Bayesian confidence interval) of G_t. Specifically, the $(1-\alpha)100\%$ credible interval

$$CrI = [q_{t,\alpha/2}, q_{t,1-\alpha/2}]$$

is defined such that $G_t(q_{t,1-\alpha/2}) - G_t(q_{t,\alpha/2}) = 1-\alpha$, which contains $(1-\alpha)$ probability mass of G_t. To avoid technicalities, we assume $q_{t,\alpha/2}$ and $q_{t,1-\alpha/2}$ are respectively the $\alpha/2$ and $1-\alpha/2$ quantiles of G_t. However, G_t is not directly computable due to the complex input-to-output mapping and needs to be estimated by simulation. A direct method is to generate M i.i.d. samples $\{\theta_t^1,\ldots,\theta_t^M\}$ from π_t, and run simulations to obtain the corresponding performance estimates $\{\hat{H}_t^1,\ldots,\hat{H}_t^M\}$, which forms an empirical distribution for G_t. From this empirical distribution, the $(1-\alpha)100\%$ credible interval CrI defined above can be estimated by $[\hat{H}_t^{(\lceil M\frac{\alpha}{2}\rceil)}, \hat{H}_t^{(\lceil M(1-\frac{\alpha}{2})\rceil)}]$, where $\hat{H}_t^{(1)} \leq \ldots \leq \hat{H}_t^{(M)}$ are the ordered values of $\{\hat{H}_t^1,\ldots,\hat{H}_t^M\}$.

Before proposing our method, we first discuss a naive extension of the above direct Bayesian method to online quantification: we take in the new input data to update the posterior distribution on the input parameter, and then run simulation experiments under sampled input parameters drawn from the updated posterior. This naive approach is computationally inefficient, since it requires running new (expensive) simulation experiments whenever a new data point is assimilated. Moreover, a sufficiently large number of simulation replications are needed for a reasonably accurate estimation. Therefore, as easy as the naive online approach sounds, it is not applicable in practical problems. Our proposed method below will work around the issue of the naive extension by reusing simulation outputs from previous time stages.

2.1 Main Idea and Algorithm

The general idea of our method consists of a sequence of three steps whenever a new data point is assimilated. These three steps include importance sampling, resampling, and restarting when some criterion is satisfied. The purpose of importance sampling step is to reuse performance outputs generated by previous simulation experiments. The importance sampling step transforms the set of i.i.d. samples following the previous posterior distribution to a new weighted set of samples following the current updated posterior distribution, and hence the performance estimates of the previous sample set will be inherited by the new weighted sample set. Following this importance sampling step, a resampling step is taken to sample (with replacements) from the weighted set of samples to generate a new set of i.i.d. samples. The purpose of the resampling step is to help preventing the issue of sample degeneracy: without resampling, after a few iterations only very few samples will have significant weights while the majority of samples all have very small weights close to zero. Even with resampling, as shown by our error-bound analysis later, the estimation error will accumulate over time when using a constant number of samples per time stage, so we further introduce a restarting step to prevent the accumulated error from blowing up. When the accumulated error becomes too large, the restarting step discards previous samples, and generates a new set of i.i.d. samples from the current posterior distribution and carries out simulation experiments to estimate the performance measure at these new samples. Following the idea outlined above, below we describe our algorithm in details.

Online Quantification of Input Uncertainty: **Importance Sampling Resampling with Restarting (ISRR)**

- At time stage $t = 0$, we start with a prior distribution π_0. Draw i.i.d. samples $\{\theta_0^1, \ldots, \theta_0^M\}$ from π_0. For each $i = 1, \ldots, M$, run N simulation experiments at θ_0^i to obtain a performance estimate $\hat{H}_0^i$, i.e., $\hat{H}_0^i = \frac{1}{N}\sum_{j=1}^N h(\xi^{i,j})$ with $\xi^{i,j} \overset{\text{iid}}{\sim} F(\cdot; \theta_0^i), j = 1, \ldots, N$.
- At time stage $t+1 (t \geq 0)$, a new data point ξ_{t+1} arrives. The following steps are carried out.
  1. **Importance sampling**: compute the importance sampling weights $\{w_t^1, \ldots, w_t^M\}$ according to:

$$w_t^i = \frac{\pi_{t+1}(\theta_t^i)}{\pi_t(\theta_t^i)} \propto p(\xi_{t+1}|\theta_t^i), i = 1, \ldots, M; \quad \sum_{i=1}^M w_t^i = 1.$$

  2. **Resampling**: sample (with replacements) from $\{\theta_t^i, i = 1, \ldots, M\}$ according to the weights $\{w_t^i, i = 1, \ldots, M\}$ to generate i.i.d. samples $\{\theta_{t+1}^1, \ldots, \theta_{t+1}^M\}$, and for each θ_{t+1}^i record its corresponding sample value $\theta_{t+1}^i = \theta_t^{i'}$ and performance estimate $\hat{H}_{t+1}^i = \hat{H}_t^{i'}$.
  3. **Restarting**: if $\widehat{\text{Var}}_{t+1} < \beta \text{Var}_{t+1}$, $\beta \in (0,1)$, then discard the current samples and generate a new set of i.i.d. samples $\{\theta_{t+1}^1, \ldots, \theta_{t+1}^M\}$ from π_{t+1}. For each $i = 1, \ldots, M$, run N simulation experiments at θ_{t+1}^i to obtain performance estimate $\hat{H}_{t+1}^i$, i.e., $\hat{H}_{t+1}^i = \frac{1}{N}\sum_{j=1}^N h(\xi^{i,j})$ with $\xi^{i,j} \overset{\text{iid}}{\sim} F(\cdot; \theta_{t+1}^i), j = 1, \ldots, M$.
  4. **Quantification**: sort performance estimates $\hat{H}_{t+1}^{(1)} \leq \ldots \leq \hat{H}_{t+1}^{(M)}$, and $[\hat{H}_{t+1}^{(\lceil M\frac{\alpha}{2} \rceil)}, \hat{H}_{t+1}^{(\lceil M(1-\frac{\alpha}{2}) \rceil)}]$ is an approximate $(1-\alpha)100\%$ credible interval (under the posterior G_{t+1}) for the true performance.

Step 1 (importance sampling) computes the weights $\{w_t^1, \ldots, w_t^M\}$ such that they sum up to 1, which is often referred to as self-normalized importance sampling (Cappé et al. 2005). This gives a discrete distribution $\{Prob(\theta_t^i) = w_t^i, i = 1, \ldots, M\}$; so in Step 2 (resampling) a new set of samples are generated by drawing each θ_t^i with probability w_t^i. In Step 3 (restarting), the restarting criterion is set to be $\widehat{\text{Var}}_{t+1} < \beta \text{Var}_{t+1}$, $\beta \in (0,1)$, where $\widehat{\text{Var}}_{t+1}$ denotes the sample variance of the samples $\{\theta_{t+1}^1, \ldots, \theta_{t+1}^M\}$ after Step 2 (resampling), and Var_{t+1} denotes the variance of π_{t+1}. This is motivated by the fact that the difference between the sample variance and true variance is an indicator of the estimation error. Intuitively, the estimation error increases if we use the same set of sample values (even with different sets of weights), because the posterior distribution π_t becomes more and more different from the prior distribution π_0 as time progresses, and the samples drawn initially from π_0 may become a poor representation of π_t (for example, most of the initial samples lie in the tails of π_t). Note here the constant β is often chosen to be close to 1. A larger β leads to more frequent restarts, and consequently smaller estimation error over time for a fixed sample size M. On the other hand, for a fixed β, a larger sample size M leads to less frequent restarts. We also note that there might be other better restarting criteria, which we leave to our future work.

2.2 Error Bounds and Asymptotic Convergence

First, we introduce some notations. Let's denote the Dirac delta function by $\delta(\cdot)$. The empirical distributions at the beginning of the time $t+1$, after Step 1 (importance sampling) at time $t+1$, and after Step 2 (resampling) at time $t+1$ are respectively denoted as: $\hat{\pi}_t(\theta) \triangleq \frac{1}{M}\sum_{i=1}^M \delta(\theta = \theta_t^i)$, $\hat{\pi}_{t+1|t}(\theta) \triangleq \sum_{i=1}^M w_t^i \delta(\theta = \theta_t^i)$, $\hat{\pi}_{t+1}(\theta) \triangleq \frac{1}{M}\sum_{i=1}^M \delta(\theta = \theta_{t+1}^i)$. Recall that $G_t(y) \triangleq Pr(H(\theta) \leq y|\xi_1, \ldots, \xi_t)$. Let g_t be the p.d.f. of G_t. The empirical posterior distributions on the performance measure at the beginning of time $t+1$, after Step 1 (importance sampling) at time $t+1$, and after Step 2 (resampling) at time $t+1$ are respectively denoted as: $\tilde{g}_t(y) \triangleq \frac{1}{M}\sum_{i=1}^M \delta(y = \hat{H}_t^i)$, $\tilde{g}_{t+1|t}(y) \triangleq \sum_{i=1}^M w_t^i \delta(y = \hat{H}_t^i)$, $\tilde{g}_{t+1} \triangleq \frac{1}{M}\sum_{i=1}^M \delta(y = \hat{H}_{t+1}^i)$, where $\hat{H}_t^i = \frac{1}{N}\sum_{j=1}^N h(\xi^{i,j})$ with $\xi^{i,j} \overset{\text{iid}}{\sim} F(\cdot; \theta_t^i), j = 1, \ldots, M$. The likelihood function at time $t+1$ is denoted as $l_t \triangleq p(\xi_{t+1}|\theta)$.

Let $C_b(\mathbb{R}^n)$ be the set of all bounded and continuous functions $\phi : \mathbb{R}^n \to \mathbb{R}$. Let $\|\cdot\|$ denote the supremum norm on $C_b(\mathbb{R}^n)$, i.e., $\|\phi\| \triangleq \sup_{x \in \mathbb{R}^n} |\phi(x)|$, $\phi \in C_b(\mathbb{R}^n)$. Given that $\pi(x)$ is a p.d.f. and $\phi : \mathbb{R}^n \to \mathbb{R}$ is an

integrable function with respect to π, we define

$$\langle \pi, \phi \rangle \triangleq \int \phi(x)\pi(x)dx,$$

and we use the shorthand notation $\langle \pi - \pi', \phi \rangle \triangleq \langle \pi, \phi \rangle - \langle \pi', \phi \rangle$. Hence, the posterior updating can be written as

$$\pi_{t+1} = \frac{\pi_t l_t}{\langle \pi_t, l_t \rangle}.$$

The empirical posterior updating can be written as

$$\hat{\pi}_{t+1|t}(\theta) = \frac{\sum_{i=1}^{M} \delta(\theta_t^i = \theta) p(\xi_{t+1}|\theta)}{\sum_{i=1}^{M} p(\xi_{t+1}|\theta_t^i)} = \frac{\hat{\pi}_t l_t}{\langle \hat{\pi}_t, l_t \rangle}.$$

We need the following mild assumptions to ensure posterior distributions are well defined and some quantities of simulation outputs are well regularized.

Assumption 1
(i) The likelihood function l_t is continuous, bounded, and strictly positive.
(ii) The expectation $E[h(\xi_\theta)]$ (taken with respect to $\xi_\theta \sim F(\cdot;\theta)$) is continuous and bounded on Θ, and the standard deviation $std(h(\xi_\theta))$ (taken with respect to $\xi_\theta \sim F(\cdot;\theta)$) is bounded on Θ.

The following lemma bounds the error between the true posterior distribution and the empirical distribution after the importance sampling step at time $t+1$. Its proof technique is similar to that for Lemma 4 in Crisan and Doucet (2002).

Lemma 1 Suppose Assumption 1(i) holds. For any $\phi \in C_b(\mathbb{R}^n)$, if $E[|\langle \hat{\pi}_t - \pi_t, \phi \rangle|] \le e_t \|\phi\|$, then

$$E\left[\left|\left\langle \hat{\pi}_{t+1|t}, \phi \right\rangle - \langle \pi_{t+1}, \phi \rangle\right|\right] \le \gamma_t e_t \|\phi\|,$$

where $\gamma_t \triangleq \frac{2\|l_t\|}{\langle \pi_t, l_t \rangle}$.

Proof. From Assumption 1(i), $l_t\phi \in C_b(\mathbb{R}^n)$, $\langle \pi_t, l_t \rangle > 0$, and $\langle \hat{\pi}_t, l_t \rangle > 0$.

$$\begin{aligned}
& \left|\left\langle \hat{\pi}_{t+1|t}, \phi \right\rangle - \langle \pi_{t+1}, \phi \rangle\right| \\
= & \left| \frac{\langle \hat{\pi}_t, l_t\phi \rangle}{\langle \hat{\pi}_t, l_t \rangle} - \frac{\langle \pi_t, l_t\phi \rangle}{\langle \pi_t, l_t \rangle} \right| \\
\le & \left| \frac{\langle \hat{\pi}_t, l_t\phi \rangle}{\langle \hat{\pi}_t, l_t \rangle} - \frac{\langle \hat{\pi}_t, l_t\phi \rangle}{\langle \pi_t, l_t \rangle} \right| + \left| \frac{\langle \hat{\pi}_t, l_t\phi \rangle}{\langle \pi_t, l_t \rangle} - \frac{\langle \pi_t, l_t\phi \rangle}{\langle \pi_t, l_t \rangle} \right|.
\end{aligned}$$

The first term above can be bounded as

$$\left| \frac{\langle \hat{\pi}_t, l_t\phi \rangle}{\langle \hat{\pi}_t, l_t \rangle} - \frac{\langle \hat{\pi}_t, l_t\phi \rangle}{\langle \pi_t, l_t \rangle} \right| = \frac{|\langle \hat{\pi}_t, l_t\phi \rangle (\langle \pi_t, l_t \rangle - \langle \hat{\pi}_t, l_t \rangle)|}{\langle \hat{\pi}_t, l_t \rangle \langle \pi_t, l_t \rangle} \le \frac{\|\phi\|}{\langle \pi_t, l_t \rangle} |\langle \pi_t, l_t \rangle - \langle \hat{\pi}_t, l_t \rangle|.$$

Hence, we get

$$\begin{aligned}
& E\left[\left|\left\langle \hat{\pi}_{t+1|t}, \phi \right\rangle - \langle \pi_{t+1}, \phi \rangle\right|\right] \\
\le & \frac{\|\phi\|}{\langle \pi_t, l_t \rangle} E\left[|\langle \pi_t, l_t \rangle - \langle \hat{\pi}_t, l_t \rangle|\right] + \frac{E[|\langle \hat{\pi}_t, l_t\phi \rangle - \langle \pi_t, l_t\phi \rangle|]}{\langle \pi_t, l_t \rangle} \\
\le & \frac{2e_t\|l_t\|}{\langle \pi_t, l_t \rangle}\|\phi\| = \gamma_t e_t \|\phi\|.
\end{aligned}$$

□

The following lemma bounds the sampling error due to the resampling step at time $t+1$. Its proof technique is similar to that for Lemma 5 in Crisan and Doucet (2002).

Lemma 2 For any $\phi \in C_b(\mathbb{R}^n)$,

$$E\left[\left|\langle \hat{\pi}_{t+1}, \phi\rangle - \left\langle \hat{\pi}_{t+1|t}, \phi\right\rangle\right|\right] \leq \frac{\|\phi\|}{\sqrt{M}}.$$

Proof. Let $\mathscr{F}_t$ be the σ-field generated by $\{\theta_t^i, i=1,\ldots,M\}$. Due to i.i.d. sampling we have unbiased estimates, i.e., $E[\phi(\theta_{t+1}^i)|\mathscr{F}_t] = \left\langle \hat{\pi}_{t+1|t}, \phi\right\rangle, i=1,\ldots,M$, where the expectation is taken with respect to the randomness in sampling. Hence,

$$\begin{aligned}
& E\left[\left|\left\langle \hat{\pi}_{t+1} - \hat{\pi}_{t+1|t}, \phi\right\rangle\right|\right]^2 \\
\leq\ & E\left[\left|\left\langle \hat{\pi}_{t+1} - \hat{\pi}_{t+1|t}, \phi\right\rangle\right|^2\right] \\
=\ & \frac{1}{M^2} E\left[E\left[\sum_{i=1}^{M}\left(\phi(\theta_{t+1}^i) - \left\langle \hat{\pi}_{t+1|t}, \phi\right\rangle\right)^2 \middle| \mathscr{F}_t\right]\right] \\
=\ & \frac{1}{M} E\left[\left\langle \hat{\pi}_{t+1|t}, \phi^2\right\rangle - \left\langle \hat{\pi}_{t+1|t}, \phi\right\rangle^2\right] \frac{1}{M}\|\phi\|^2.
\end{aligned}$$

By taking square root on both sides of the inequality above, we prove the lemma. □

Based on the lemmas above, we prove the following theorem, which bounds the accumulated error between the true posterior distribution and the empirical distribution at time $t+1$.

Theorem 1 Suppose Assumption 1(i) holds. For any $\phi \in C_b(\mathbb{R}^n)$,

$$\begin{aligned}
E\left[\left|\left\langle \hat{\pi}_{t+1|t} - \pi_{t+1}, \phi\right\rangle\right|\right] &\leq c_{t+1|t}\frac{\|\phi\|}{\sqrt{M}} \\
E\left[\left|\langle \hat{\pi}_{t+1} - \pi_{t+1}, \phi\rangle\right|\right] &\leq c_{t+1}\frac{\|\phi\|}{\sqrt{M}},
\end{aligned}$$

where the constants $c_{t+1|t}$ and c_{t+1} are respectively

$$c_{t+1|t} = \sum_{j=0}^{t}\left(\prod_{i=j}^{t}\gamma_i\right), \quad c_{t+1} = \sum_{j=0}^{t}\left(\prod_{i=j}^{t}\gamma_i\right) + 1.$$

Proof.

$$\begin{aligned}
& E\left[\left|\langle \hat{\pi}_{t+1} - \pi_{t+1}, \phi\rangle\right|\right] \\
\leq\ & E\left[\left|\left\langle \hat{\pi}_{t+1} - \hat{\pi}_{t+1|t}, \phi\right\rangle\right|\right] + E\left[\left|\left\langle \hat{\pi}_{t+1|t} - \pi_{t+1}, \phi\right\rangle\right|\right] \\
\leq\ & \frac{1}{\sqrt{M}}\|\phi\| + \gamma_t e_t\|\phi\| = e_{t+1}\|\phi\|,
\end{aligned}$$

where $e_{t+1} = \gamma_t e_t + \frac{1}{\sqrt{M}}$. Note that $E[|\langle \hat{\pi}_0 - \pi_0, \phi\rangle|] \leq \frac{\|\phi\|}{\sqrt{M}}$, so $e_0 = \frac{1}{\sqrt{M}}$. By induction, we get

$$e_{t+1} = \left\{\sum_{j=0}^{t}\left(\prod_{i=j}^{t}\gamma_i\right) + 1\right\}\frac{1}{\sqrt{M}} = c_{t+1}\frac{1}{\sqrt{M}}.$$

Similarly, we get

$$E\left[\left|\left\langle \hat{\pi}_{t+1|t} - \pi_{t+1}, \phi\right\rangle\right|\right] \leq \gamma_t e_t\|\phi\| = \left\{\sum_{j=0}^{t}\left(\prod_{i=j}^{t}\gamma_i\right)\right\}\frac{\|\phi\|}{\sqrt{M}} = c_{t+1|t}\frac{1}{\sqrt{M}}.$$

□

The following corollary shows the weak convergence of empirical posterior distributions to the true posterior distributions as the sample number M goes to infinity.

Corollary 2 Suppose Assumption 1(i) holds. For any fixed time t, $\hat{\pi}_{t+1|t}$ and $\hat{\pi}_{t+1}$ converge weakly to π_{t+1} in mean as $M \to \infty$.

Proof. By Portmanteau theorem, the corollary statement is equivalent to showing $\lim_{M\to\infty} \langle \hat{\pi}_{t+1|t}, \phi \rangle = \langle \pi_{t+1}, \phi \rangle$ ($\lim_{M\to\infty} \langle \hat{\pi}_{t+1}, \phi \rangle = \langle \pi_{t+1}, \phi \rangle$) in mean for any $\phi \in C_b(\mathbb{R}^n)$. Hence, it follows directly from Theorem 1. □

The following theorem shows that the induced empirical distribution converges weakly to the true posterior distribution on the performance measure as the sample number M and number of simulation replications N go to infinity.

Theorem 3 Suppose Assumption 1 holds. For any fixed time t, $\tilde{g}_t$ converges weakly to g_t in mean as $M,N \to \infty$.

Proof. For simplicity, we define $\hat{H}(\theta) \triangleq \frac{1}{N}\sum_{j=1}^{N} h(\xi_\theta^j)$, where $\xi_\theta^j \overset{\text{iid}}{\sim} F(\cdot;\theta)$. Assume $\varphi : \mathbb{R} \to \mathbb{R}$ is bounded and Lipschitz continuous with a Lipschitz constant L_φ. We use $\varphi \circ H$ to denote the composition of φ and H, i.e., $\varphi \circ H = \varphi(H(\cdot))$. Recall that $\tilde{g}_t(y) = \frac{1}{M}\sum_{i=1}^{M} \delta(y = \hat{H}_t^i)$.

$$|\langle g_t, \varphi \rangle - \langle \tilde{g}_t, \varphi \rangle| = |\langle \pi_t, \varphi \circ H \rangle - \langle \hat{\pi}_t, \varphi \circ \hat{H} \rangle| \le |\langle \pi_t, \varphi \circ H \rangle - \langle \hat{\pi}_t, \varphi \circ H \rangle| + \left|\langle \hat{\pi}_t, \varphi \circ H \rangle - \langle \hat{\pi}_t, \varphi \circ \hat{H} \rangle\right|$$

From Assumption 1(ii), $H(\theta) \in C_b(\mathbb{R}^n)$. Hence, $\varphi \circ H \in C_b(\mathbb{R}^n)$, and it then follows from Theorem 1 that the first term above is bounded by $c_t \frac{\|\varphi \circ H\|}{\sqrt{M}}$. The expectation of the second term above can be written as

$$E\left[\left\langle \hat{\pi}_t, \left|\varphi \circ H - \varphi \circ \hat{H}\right| \right\rangle\right] = \left\langle \hat{\pi}_t, E\left[\left|\varphi \circ H - \varphi \circ \hat{H}\right|\right] \right\rangle,$$

where the interchange of expectation and integration follows from the bounded convergence theorem. Since φ is Lipschitz continuous, we have

$$\begin{aligned}
& E\left[\left|\varphi \circ H - \varphi \circ \hat{H}\right|\right] \\
\le\ & L_\varphi E\left[\left|E[h(\xi_\theta)] - \frac{1}{N}\sum_{j=1}^{N} h(\xi_\theta^j)\right|\right] \\
\le\ & L_\varphi E\left[\left(E[h(\xi_\theta)] - \frac{1}{N}\sum_{j=1}^{N} h(\xi_\theta^j)\right)^2\right]^{1/2} \\
\le\ & L_\varphi \frac{std(h(\xi_\theta))}{\sqrt{N}},
\end{aligned}$$

where $std(h(\xi_\theta))$ is bounded on Θ by Assumption 1(ii). Therefore,

$$E\left[|\langle g_t, \varphi \rangle - \langle \tilde{g}_t, \varphi \rangle|\right] \le \frac{c_t \|\varphi \circ H\|}{\sqrt{M}} + \frac{L_\varphi \|std(h(\xi)\|}{\sqrt{N}},$$

where $\|std(h(\xi)\| = \sup_{\theta\in\Theta} std(h(\xi_\theta)$. Hence, as $M,N \to \infty$, $\langle \tilde{g}_t, \phi \rangle \to \langle g_t, \phi \rangle$ in mean. Since it holds for any φ that is bounded Lipschitz, by Portmanteau theorem $\tilde{g}_t$ converges weakly to g_{t+1} in mean as $M,N \to \infty$. □

Denote the true $(1-\alpha)$ credible interval (under g_t) as $[q_{t,\alpha/2}, q_{t,1-\alpha/2}]$, where $q_{t,\alpha/2}$ and $q_{t,1-\alpha/2}$ are respectively the $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$ quantiles of g_t. In other words, this credible interval covers $(1-\alpha)$ probability mass of the distribution g_t, which represents our belief about the true performance measure at time t. From Theorem 3, it is straightforward to see the following corollary.

Corollary 4 Suppose Assumption 1 holds. For any fixed time t, $[\hat{H}_t^{(\lceil M\frac{\alpha}{2}\rceil)}, \hat{H}_t^{(\lceil M(1-\frac{\alpha}{2})\rceil)}]$ converges to the true $(1-\alpha)$ credible interval $[q_{t,\alpha/2}, q_{t,1-\alpha/2}]$ in mean as $M,N \to \infty$.

Remark 1: Our analysis above treats Algorithm ISRR without the restarting step. It is straightforward to extend the error-bound results to Algorithm ISRR (including the restarting step) simply by resetting each restarting time to time zero in these results. The asymptotic convergence results also hold for Algorithm ISRR (including the restarting step) since they are proved for each fixed time t.

Remark 2: The error-bound results not only serve as a bridge to prove the asymptotic convergence, but also provide important insights into the algorithm finite-sample performance. As Theorem 1 indicates, the estimation error between empirical posterior distributions and the true posterior distributions increases over time if we use a constant sample number M. The error can be kept constant if we increase the sample number at an appropriate rate as time progresses. However, not only the exact increasing rate of the sample number is difficult to identify, increasing the sample number also means more computation for new simulation experiments at every time stage. This in turn has motivated us to introduce the restarting mechanism into the algorithm, so that the error can be reset to the initial value when the accumulated error after a certain time period becomes too large. We also note that uniform-in-time error bound is not likely (if not at all impossible) to be obtained for online parameter estimation such as our scenario here. Uniform-in-time error bound for sequential Monte Carlo methods can be obtained if the state-space model possesses the so-called exponential forgetting property (ref. Moral 2004, Chapter 4), which means that the true posterior distribution forgets exponentially fast its initial condition. However, this exponentially forgetting property obviously does not hold here when the state is a static parameter, and the most recent advances in sequential Monte Carlo methods for parameter estimation still suffer from the increasing-in-time error bounds (Kantas et al. 2015).

3 NUMERICAL EXAMPLES

We use a simple M/M/1 queue example to verify our theoretical results and demonstrate the performance of our algorithm. In particular, we are interested in estimating the average queue length in the system. The true service rate μ is known to both the judges and the experimenters. However, the true arrival rate λ^c of the customers is only known to judges but unknown to the experimenters. Here we set $\mu = 10$ and $\lambda^c = 5$, so the true average queue length is $\lambda^c/(\mu-\lambda^c) = 1$.

We take a Bayesian approach to estimate the unknown input parameter — the Poisson arrival rate λ. We assume there is a historical observation of $n = 50$ data points $\{x_1,...,x_n\}$ that were collected before time $t = 0$. Starting from time $t = 1$, there is one new data point ξ_t arriving at each time stage t. All these data points are i.i.d. from the true input distribution, i.e., the exponential distribution with rate parameter λ^c. Assuming a non-informative prior for λ, i.e., $\pi_{-1}(\lambda) \propto 1/\lambda$, the posterior distribution π_0 based on historic data at time $t = 0$ is a Gamma distribution with shape parameter n and scale parameter $1/(\sum_{i=1}^n x_i)$. Similarly, at each following time t, the posterior distribution π_t is a Gamma distribution with shape parameter $n+t$ and scale parameter $1/(\sum_{i=1}^n x_i + \sum_{s=1}^t \xi_s)$.

We apply the proposed algorithm ISRR and the naive Bayesian method (mentioned right before Section 2.1) to construct the $100(1-\alpha)\%$ credible interval (CrI) $[\hat{q}_{t,\alpha/2}, \hat{q}_{t,1-\alpha/2}]$ of the average queue length and compare their performance. Here, $\hat{q}_{t,\alpha/2}$, $\hat{q}_{t,1-\alpha/2}$ are the estimated $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$ quantiles at time t respectively. The true credible interval $[q_{t,\alpha/2}, q_{t,1-\alpha/2}]$ obtained from the true posterior distribution G_t serves as a benchmark in this comparison. We also use $\hat{p}_{t,\alpha}$ to denote the coverage probability of the empirical credible interval, which is equal to the probability mass of $[\hat{q}_{t,\alpha/2}, \hat{q}_{t,1-\alpha/2}]$ under G_t. In the experiments, we set $\alpha = 0.1$, i.e., our target coverage probability is 90%.

In ISRR, we set $\beta = 0.95$, i.e. ISRR restarts when $\widehat{\text{Var}}_{t+1} < 0.95\text{Var}_{t+1}$. In both ISRR and the naive method, we first use linear interpolation (between samples) to smooth the empirical distribution of the samples, and then use the quantiles of the smoothed empirical distribution as estimates of the true quantiles. The linear interpolation is used to improve the accuracy of quantile estimates, especially when there is

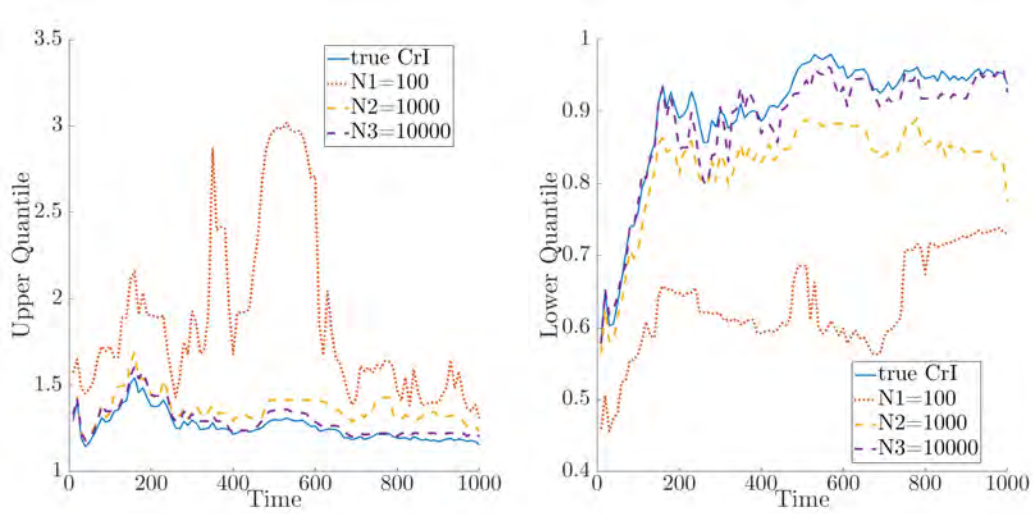


Figure 1: CrI quantiles estimates by ISRR when $N = 100, 1000, 10000$ and $M = 10000$.

only a small number of samples in the empirical distribution. We run both algorithms over a long time horizon $T = 1000$.

Experiment 1 In this experiment, we verify our asymptotic convergence results, specifically Theorem 3 and Corollary 4. We show that ISRR's empirical credible interval converges to the true credible interval as M (sample number) and N (number of simulation replications) increase, and moreover, ISRR's empirical credible interval can track the true one well over long time horizon.

a. We first study the influence of N. We fix $M = 10000$ and run the algorithm with $N = 100, 1000, 10000$. As expected, it can be seen in Figure 1 that when we increase N, both upper and lower quantiles estimates become closer to the true ones. Moreover, the error does not blow up as times goes by, and our algorithm can track the true credible interval properly.

b. We then study the influence of M. Here, M is chosen to be $100, 1000, 100000$, and we evaluate the exact performance measure $\{\lambda_t^i/(\mu - \lambda_t^i)\}_{i=1}^M$ instead of using simulation estimation for each λ_t^i. In other words, N here is set as infinity. In Figure 2, it is clear that when we increase M, the empirical credible intervals become closer to the true ones for all time stages; when $M = 100000$, the empirical credible intervals almost coincide with the true ones. Table 1 lists the number of restarts and corresponding number of evaluations. As we expect, a larger M often leads to less restarts. However, what really matters to the computational cost is the total number of evaluations (which is equal to $(\text{number of restarts} + 1) \times M$). In this case a smaller M incurs less total evaluations.

Table 1: Experiment 1.b results (fixed N).

M	# Restarts	Total # Evaluations
100	93	9400
1000	48	49000
100000	3	400000

Experiment 2 At last, we compare ISRR with the naive Bayesian method. Since running simulation replications dominates the computational time of other steps in those methods, a fair comparison is to use the same total number of simulation replications. With a total number of evaluations of $(\text{number of restarts} + 1) \times M$, ISRR generates M samples of θ at each restart, and the naive method generates $(\text{number of restarts} +$

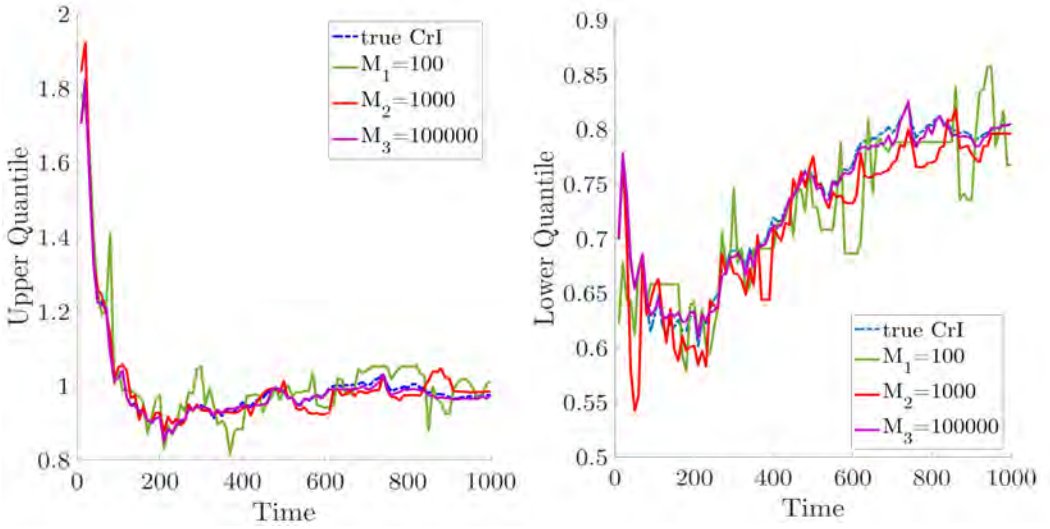
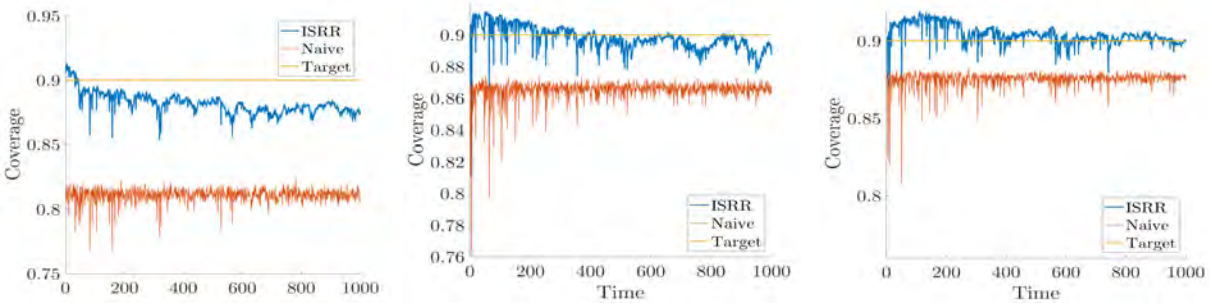


Figure 2: CrI quantiles estimates when $M = 100, 1000, 100000$ (No stochastic uncertainty, i.e. $N = \infty$).

$1) \times M/T$ samples in every iteration. Similar to Experiment 1, we choose $N = \infty$ and consider three cases $M = 100, 500, 1000$. For each M, both of methods are run for 1000 independent macro-replications to see their mean behaviors. Table 2 shows the errors of CrI quantiles estimates and coverage probabilities at time $t = 200, 600, 1000$. Here, $e_{\hat{q}_\gamma} = \hat{q}_\gamma - q_\gamma$ ($\gamma = \alpha/2, 1-\alpha/2$), which is the error of the quantile estimate, and $e_{\hat{p}_\alpha} = \hat{p}_\alpha - (1-\alpha)$, which is the difference between the empirical coverage probability and the target coverage $1-\alpha$. We have the following observations:



(a) $M = 100$ (b) $M = 500$ (c) $M = 1000$

Figure 3: Coverage probabilities when $M = 100, 500, 1000$ and $\alpha = 0.1$.

a. When M is small, both methods have "under-coverage" due to the lack of θ **samples**. However, ISRR performs much better than the naive approach. Specifically, ISRR's quantile estimates is very close to the true ones, and its credible interval achieves nearly 90% coverage ($e_{\hat{p}_\alpha} \approx -0.02$). However, under the same computational budget, the coverage probability of the naive approach is around 80% ($e_{\hat{p}_\alpha} \approx -0.1$), which is much smaller than the target 90%. Thus, when the computational budget is limited, ISRR shows its clear advantage over the naive approach.

b. When M gets bigger, the difference between two approaches decreases. This is because the naive approach has a large enough number of evaluations to get a good estimate for this simple example. However, when $M = 1000$, the naive approach remains "under-coverage" ($e_{\hat{p}_\alpha} \approx -0.02$) compared to ISRR. Therefore, our method is still better than the naive approach.

Table 2: Errors of CrI quantiles estimates and coverage probabilities when $M = 100, 500, 1000$, where $e_{\hat{q}_{\alpha/2}} = \hat{q}_\gamma - q_\gamma$, $\gamma = \alpha/2, 1-\alpha/2$, $e_{\hat{p}_{1-\alpha}} = \hat{p}_\alpha - (1-\alpha)$, and SD represents standard deviation.

		$e_{\hat{q}_{\alpha/2}}$ mean	$e_{\hat{q}_{\alpha/2}}$ SD	$e_{\hat{q}_{1-\alpha/2}}$ mean	$e_{\hat{q}_{1-\alpha/2}}$ SD	$e_{\hat{p}_\alpha}$ mean	$e_{\hat{p}_\alpha}$ SD
$M = 100, t = 200$	ISRR	–0.0192	0.0426	0.018	0.0508	–0.01507	0.0891
	Naive	0.00276	0.0410	–0.0346	0.0510	–0.0918	0.106
$M = 100, t = 600$	ISRR	–0.0128	0.0327	0.00518	0.0419	–0.0185	0.0953
	Naive	0.00145	0.0323	–0.0222	0.0330	–0.0856	0.1032
$M = 100, t = 1000$	ISRR	–0.00828	0.0305	0.00503	0.0330	–0.0240	0.0950
	Naive	0.000923	0.0276	–0.0196	0.0282	–0.0911	0.1075
$M = 500, t = 200$	ISRR	–0.0401	0.0598	0.0171	0.101	0.00332	0.0575
	Naive	–0.0167	0.0558	–0.0418	0.0831	–0.0381	0.0685
$M = 500, t = 600$	ISRR	–0.0219	0.0367	0.007	0.0501	–0.00659	0.0784
	Naive	–0.0105	0.0317	–0.0178	0.0362	–0.0397	0.0716
$M = 500, t = 1000$	ISRR	–0.0121	0.0287	0.00505	0.0350	–0.0129	0.0910
	Naive	–0.00738	0.0235	–0.0115	0.0247	–0.0366	0.0704
$M = 1000, t = 200$	ISRR	–0.0359	0.0520	0.0266	0.0879	0.0125	0.0398
	Naive	–0.0118	0.0475	–0.0282	0.0758	–0.0268	0.0619
$M = 1000, t = 600$	ISRR	–0.0177	0.0295	0.00510	0.0378	0.00449	0.0505
	Naive	–0.00716	0.0245	–0.0146	0.0293	–0.0252	0.0585
$M = 1000, t = 1000$	ISRR	–0.0172	0.0274	0.00575	0.0343	-6.86×10^{-5}	0.0726
	Naive	–0.00662	0.0207	–0.00955	0.0224	–0.025	0.0614

c. Similar observations can be made from Figure 3. Both methods achieve better coverage probabilities when M is larger. Despite that ISRR's advantage becomes less obvious as M increases, it performs better than the naive approach in all these three scenarios.

4 CONCLUSIONS

Assuming the input model of a stochastic simulation model takes a parametric form, we propose a method for online quantification of input uncertainty when input data arrive sequentially in time. The method has two features. First is its computational efficiency for real-time updating: at each time stage, by utilizing the simulation outputs from previous time stages, the method updates the quantification estimates with very little computational cost. Second is its good performance over a long time horizon: by introducing a resampling and restarting mechanism, the method effectively addresses the challenge of accumulated estimation error over a long time horizon. Our theoretical analysis proves the error bound and asymptotic convergence of the method. We compare the proposed method with a naive Bayesian method on an M/M/1 queue example, which shows the proposed method performs much better than the naive method especially under a limited computing budget. A future direction is to extend this online method for nonparametric input models.

ACKNOWLEDGMENTS

This work was supported by National Science Foundation under Grant CAREER CMMI-1453934.

REFERENCES

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

Barton, R. R., and L. W. Schruben. 1993. "Uniform and Bootstrap Resampling of Empirical Distributions". In *Proceedings of the 1993 Winter Simulation Conference*, edited by G. W. Evans et al., 503–508. Piscataway, New Jersey: IEEE.

Barton, R. R., and L. W. Schruben. 2001. "Resampling Methods for Input Modeling". In *Proceedings of the 2001 Winter Simulation Conference*, edited by B. A. Peters et al., 372–378. Piscataway, New Jersey: IEEE.

Barton, R. R., B. L. Nelson, and W. Xie. 2013. "Quantifying Input Uncertainty via Simulation Confidence Intervals". *INFORMS Journal on Computing* 26(1):74–87.

Cappé, O., E. Moulines, and T. Rydén. 2005. *Inference in Hidden Markov Models*. Springer Series in Statistics. New York: Springer.

Cheng, R. C., and W. Holloand. 1997. "Sensitivity of Computer Simulation Experiments to Errors in Input Data". *Journal of Statistical Computation and Simulation* 57(1-4):219–241.

Chick, S. E. 2001. "Input Distribution selection for Simulation Experiments: accounting for Input Uncertainty". *Operations Research* 49(5):744–758.

Crisan, D., and A. Doucet. 2002. "A survey of convergence results on Particle Filtering methods for practitioners". *IEEE Transaction on Signal Processing* 50(3):736–746.

Doucet, A., S. Godsill, and C. Andrieu. 2000. "On Sequential Monte Carlo Sampling Methods for Bayesian Filtering". *Statistics and Computing* 10(3):197–208.

Feng, M., and J. Staum. 2015. "Green Simulation Designs for Repeated Experiments". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz et al., 403–413. Piscataway, New Jersey: IEEE.

Feng, M., and J. Staum. 2017. "Green Simulation: Reusing the Output of Repeated Experiments". *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 27(23):1–28.

Kantas, N., A. Doucet, S. S. Singh, J. Maciejowski, and N. Chopin. 2015. "On Particle methods for Parameter Estimation in State-Space models". *Statistical Science* 30(3):328–351.

Lam, H. 2016. "Robust Sensitivity Analysis for Stochastic Systems". *Mathematics of Operations Research* 41(4):1248–1275.

Moral, P. D. 2004. *Feynman-Kac Formulae: genealogical and interacting Particle Systems with applications*. New York: Springer.

Song, E., and B. L. Nelson. 2015. "Quickly assessing contributions to Input Uncertainty". *IIE Transactions* 47(9):893–909.

Xie, W., B. L. Nelson, and R. R. Barton. 2014. "A Bayesian Framework for Quantifying Uncertainty in Stochastic Simulation". *Operations Research* 62(6):1439–1452.

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.

AUTHOR BIOGRAPHIES

ENLU ZHOU is an Associate Professor in the H. Milton Stewart School of Industrial and Systems Engineering at Georgia Institute of Technology. She received the B.S. degree with highest honors in electrical engineering from Zhejiang University, China, in 2004, and received the Ph.D. degree in electrical engineering from the University of Maryland, College Park, in 2009. Her research interests include stochastic control and simulation optimization. Her email address is enlu.zhou@isye.gatech.edu.

TIANYI LIU is a Ph.D. student in the H. Milton Stewart School of Industrial and Systems Engineering at Georgia Institute of Technology. He received his B.S. in Mathematics from Nanjing University, China, in 2016. His research interests include machine learning, stochastic optimization, and simulation optimization. His e-mail address is tliu341@gatech.edu.