Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

CALIBRATING INPUT PARAMETERS VIA ELIGIBILITY SETS

Yuanlu Bai
Henry Lam

Industrial Engineering and Operations Research
Columbia University
500 West 120th Street
New York, NY 10027, USA

ABSTRACT

Reliable simulation analysis requires accurately calibrating input model parameters. While there has been a sizable literature on parameter calibration that utilizes directly observed data, much less attention has been paid to the situation where only output-level data are available to justify input parameter choices. This latter problem, which is known as the inverse problem and relates to the model validation literature, involves several new challenges, one of which is the non-identifiability issue. In this paper we introduce the concept of *eligibility set* to bypass non-identifiability, by relaxing the need of consistent estimation to obtaining bounds on the input parameter values. We reason this concept from the worst-case notion in robust optimization, and demonstrate how to compute eligibility set via empirical matching between the simulated and the real outputs. We substantiate our procedure with theoretical error analysis and validate its effectiveness via numerical experiments.

1 INTRODUCTION

Stochastic simulation is widely-used in system performance evaluation and decision-making. It operates by generating random variates, fed through the system logic, to obtain random outputs that are utilized for subsequent analysis. To ensure the reliability of the analysis, the input parameters in the simulation need to be accurately calibrated. These parameters can arise in the probability distributions of the input random variates and the system logic, and are typically obtained via data or domain knowledge. The literature on input modeling and input uncertainty, which has grown substantially in recent years, precisely handles these issues (see, for example, the surveys Henderson (2003), Chick (2006), Barton (2012), Song et al. (2014), Lam (2016)).

While calibrating input parameters from direct data has been actively studied, much less attention has been paid to situations where only output-level data, instead of direct input-level data, are available to assist input parameter selection. This nonetheless happens ubiquitously in practice. As a simple example, models of inter-arrival times and service times are needed to simulate generic service queues. Some service centers, however, may not have advanced electronic systems to track each customer at all times, and may only record the times a customer arrives or leaves the building. It is thus unclear how the sojourn time is contributed between the waiting time and the service time, yet the latter is needed to calibrate the model for use. One could imagine analogous issues can get substantially more complicated as the system size grows.

In this paper, we study input parameter calibration using only output-level data. In scientific modeling, the task of inferring input from output is often referred to as an "inverse problem" (Tarantola 2005). Related investigations include model error inference or uncertainty quantification, typically addressed via Bayesian approaches (Kennedy and O'Hagan 2001; Currin et al. 1991). In stochastic simulation, these issues are

closely related to so-called model calibration and validation (Sargent 2010; Kleijnen 1995). Upon building a simulation model, the modeler checks the differences of the simulated outputs with real-world outputs ("validation"), and makes refinements to the simulation model to close the gap ("calibration"). This process is conducted iteratively, each time possibly with expansive effort to collect more real data, and using expert (or otherwise ad hoc) judgements to decide how to validate and what data to collect. Some statistical tests have been suggested for validation, such as the two-sample mean-difference tests (Balci and Sargent 1982) and the Schruben-Turing test (Schruben 1980). However, despite these works and the importance of model calibration and validation, there appears relatively few studies on concrete general-purpose methodologies for these tasks (Nelson 2016).

A challenge in inverse problem or model calibration is the so-called non-identifiability (Tarantola 2005). That is, there exist more than one input parameter value that exactly matches the output behavior. This is an intrinsic barrier in inverse model calibration, in the sense that this cannot be resolved even if one has an infinite amount of data. It stems, instead, from the possibility that the input-output map is non-injective. Compared to standard statistical inference, one could also view non-identifiability as arising because the data is only partially observed. Non-identifiability appears generally in inverse problems and is not only restricted to the setting of stochastic simulation. In the latter, however, one additional challenge is that the output is represented by a probability distribution. Thus, when speaking of a "match" between the model and real system behaviors, one needs to think of a close distance between probability distributions.

Our main goal of this paper is to introduce a notion of *eligibility set* to fundamentally bypass the non-identifiabliity issue. The idea is that, instead of insisting on consistent parameter estimation which could be intrinsically impossible as discussed, we relax our "target" to obtaining tight bounds or regions on all the possible parameter values given the data. Intuitively, when this "eligibility set" is big, this means the problem is very "ill-posed" and it is difficult to pinpoint exactly the parameter location. On the other hand, if the set is small, this means we are capable of accurately inferring the true location. Regardless of the set size, however, the key of this concept is that it is statistically valid in all circumstances. In other words, the eligibility set contains the true parameter with certain given confidence level, and thus is correct in a rigorous statistical sense. The set size dictates only the conservativeness level rather than correctness. Constructing a valid eligibility set needs to handle both the input-output map and the statistical noise from the simulation and real data. We will demonstrate how to do so, and also derive some theoretical results regarding our calibration errors. Finally, we validate the performances of our method via numerical experiments.

We discuss some related works. First, our construction of eligibility set resembles concepts in robust optimization (RO) (Bertsimas et al. 2011; Ben-Tal et al. 2009), and relatedly distributionally robust optimization (DRO) (Delage and Ye 2010; Wiesemann et al. 2014). This literature advocates decision-making under the worst-case scenario, in situations where the optimization problem contains unknown or uncertain parameters. The worst-case scenario is typically obtained among a set that likely contains the true parameter value, which is often known as the uncertainty set or ambiguity set. In DRO in particular, the unknown parameter is the underlying probability distribution in a stochastic problem. Our eligibility set utilizes, in a sense, the concept of uncertainty set that attempts to capture the truth with high likelihood. Second, our work is related to a couple of recent works on calibration in the simulation literature. Like us, Goeva et al. (2019) investigate the calibration of input models from output-level data. They focus on a nonparametric setting in which the input model is represented as a general probability distribution, and employ DRO to calibrate the input model by providing bounds to certain target input-dependent quantities. However, without an explicit input parameter in their framework, the concept of eligibility set is not clearly defined there. As a result, their construction of bounds is different from our presented methodology. Moreover, despite the study of asymptotic effectiveness, they do not provide instructions on how to choose the simulation size corresponding to the data size as we do here. Other related works include Lam et al. (2017), Zhang and Zou (2016), Plumlee and Lam (2017), which consider the inference on the model error between the simulation model and the real world using Bayesian approaches or machine

learning methods. Lastly, there is a line of work using maximum entropy to calibrate distributions, in pricing financial derivatives (Avellaneda et al. 2001; Glasserman and Yu 2005), and also in simulation and probabilistic analysis (Kraan and Bedford 2005; Goeva et al. 2014).

The remainder of this paper is as follows. Section 2 formulates the problem setting and introduces notations. Sections 3 and 4 describe in detail our concepts and procedure respectively. Section 5 presents theoretical statistical guarantees. Section 6 demonstrates the numerical results. Section 7 concludes this paper.

2 PROBLEM SETTING AND NOTATIONS

We consider a family of output probability distributions P^θ where θ denotes the finite-dimensional parameter that we want to calibrate. We suppose that the true value of θ, denoted by θ_0, is known to lie in the parameter space Θ. We assume that real-world output data, denoted $X_1, X_2, \cdots, X_N \sim P^{\theta_0}$, are available to us. On the other hand, we also have the capability to simulate a random sample $Y_1^\theta, Y_2^\theta, \cdots, Y_n^\theta \sim P^\theta$, given any input value $\theta \in \Theta$. The latter condition is a basic paradigm in stochastic simulation that is generally satisfied and essential to our developments. In practice, besides parameter uncertainty, the risk of model misspecification is also a common issue, but in this paper we focus on solving the former problem assuming the correctness of the parametric model.

It is noteworthy that we make no assumptions on the relationship between θ and the output distribution P^θ. All we need is the simulability of samples from P^θ given θ, and the underlying map can be extremely complicated. A generic example would be a general queuing system or network where the inter-arrival and service time distributions are known to be in certain parametric families. Suppose we have data from some output quantities such as the average sojourn times, in which case P^{θ_0} can denote their distribution. Given the parameter value, the system dynamics can be readily simulated to generate random samples, yet it is too complicated to write down analytical expression. In other cases, we could encounter a "black box" as the simulation model developed by a third party, in which case our framework would still apply.

To fix some notations, the real output data set $X_1, \cdots, X_N$ and the simulated output data set $Y_1^\theta, \cdots, Y_n^\theta$ each determines an empirical probability distribution, which we respectively write as $P_N^{\theta_0}$ and P_n^θ. More concretely, they are given by

$$P_N^{\theta_0}(\cdot) = \frac{1}{N}\sum_{k=1}^{N} \delta_{X_k}(\cdot)$$

and

$$P_n^\theta(\cdot) = \frac{1}{n}\sum_{j=1}^{n} \delta_{Y_j^\theta}(\cdot)$$

where $\delta_{X_k}(\cdot)$ and $\delta_{Y_j^\theta}(\cdot)$ denote the Dirac measures at X_k and Y_j^θ.

We also use F^θ to denote the cumulative distribution function of P^θ, so that

$$F^{\theta_0}(x) = P(X_1 \le x), x \in \mathbb{R}$$

and

$$F^\theta(x) = P(Y_1^\theta \le x), x \in \mathbb{R}.$$

Correspondingly, we denote the empirical distribution functions of $X_1, \cdots, X_N$ and $Y_1^\theta, \cdots, Y_n^\theta$ respectively as $F_N^{\theta_0}$ and F_n^θ. That is,

$$F_N^{\theta_0}(x) = \frac{1}{N}\sum_{k=1}^{N} I(X_k \le x), x \in \mathbb{R}$$

and

$$F_n^\theta(x) = \frac{1}{n}\sum_{j=1}^{n} I(Y_j^\theta \le x), x \in \mathbb{R}$$

where $I(\cdot)$ is the indicator function.

3 ELIGIBILITY SET: BASIC CONCEPTS

Our goal is to estimate θ_0 from the real output data $X_1, \cdots, X_N$ and using the capability to simulate samples given θ, namely $Y_1^\theta, \cdots, Y_n^\theta$ for any given θ of our choice, potentially repeatedly. The standard statistical approach for this task would be to obtain an estimate as close to θ_0 as possible, i.e., a consistent estimator with error shrinking to zero as the data size increases.

To understand some challenge of the above and how we can handle it, note that our real observations only inform us on P^{θ_0}, and our model allows us to observe P^θ. In other words, any estimation needs to be done at the output level, by matching these two distributions in some sense. To make it more precise, consider a statistical distance $d(\cdot,\cdot)$ between two probability distributions. The best scenario one could hope for is to obtain θ such that $d(P^\theta, P^{\theta_0}) = 0$. However, because of the intricacy of the input-output map, there could be other $\theta \neq \theta_0$ such that the distance is 0. This issue is known as non-identifiability. Moreover, even in the case that there is a unique θ_0 giving rise to the zero distance, in a finite sample situation there could still be many θ that gives small $d(P^\theta, P^{\theta_0})$, making the problem ill-posed. In scientific computing, these are usually handled via Bayesian approaches or some regularization to identify good candidates of θ, but since we are looking at the match of distributions in the stochastic simulation setting, it is not clear these approaches can apply easily.

Our idea is to relax the notion of consistent estimation, or to look for some "best" parameter, into a region that likely contains the true parameter value. We call this region the *eligibility set* of θ. To explain, suppose we have an infinite amount of real data, so that P^{θ_0} is fully known. The eligibility set in this case would be

$$\{\theta \in \Theta : d(P^\theta, P^{\theta_0}) = 0\} \tag{1}$$

It is trivial to see that (1) contains θ_0. Note that the choice of statistical distance $d(\cdot,\cdot)$ does not affect the correctness of this claim, but it can affect the *conservativeness*, i.e., some choices of $d(\cdot,\cdot)$ can give rise to a larger-volume set than others. Typically, the smaller the set size, the more successful it is considered in our calibration task.

To proceed, in practice we only have a finite real data size N, which incurs statistical error in informing P^{θ_0}. In this situation, we consider

$$\tilde{\Theta} = \{\theta \in \Theta : d(P^\theta, P_N^{\theta_0}) \leq \eta\} \tag{2}$$

where $\eta \in \mathbb{R}^+$ is a suitable constant. Suppose that $\{Q : d(Q, P_N^{\theta_0}) \leq \eta\}$ is a $(1-\alpha)$-level confidence region for the true output distribution P^{θ_0}, that is,

$$\mathbb{P}(d(P^{\theta_0}, P_N^{\theta_0}) \leq \eta) \geq 1 - \alpha \tag{3}$$

where $\mathbb{P}$ refers to the probability with respect to the real data $X_1, \ldots, X_N$. Thus we have

$$\mathbb{P}(\theta_0 \in \tilde{\Theta}) = \mathbb{P}(d(P^{\theta_0}, P_N^{\theta_0}) \leq \eta) \geq 1 - \alpha.$$

In other words, $\tilde{\Theta}$ is a $1-\alpha$ confidence region for θ_0. We summarize this as:

Theorem 1 Suppose $\{Q : d(Q, P_N^{\theta_0}) \leq \eta\}$ is a $(1-\alpha)$-level confidence region for the true output distribution P^{θ_0}. Then the set $\tilde{\Theta}$ in (2) is a $(1-\alpha)$-level confidence region for θ_0. Similarly, if the confidence guarantee for $\{Q : d(Q, P_N^{\theta_0}) \leq \eta\}$ holds asymptotically as N increases, then a corresponding asymptotic guarantee holds for $\tilde{\Theta}$.

Note that in general, one can replace $\{Q : d(Q, P_N^{\theta_0}) \leq \eta\}$ by any other confidence region of P^{θ_0}. However, the form $d(Q, P_N^{\theta_0}) \leq \eta$ is especially handy since $d(Q, P_N^{\theta_0})$ can represent a goodness-of-fit

statistic for Q, and the resulting form of confidence region is obtained via the standard duality from a hypothesis test to a confidence region. Moreover, the form depicted by $\tilde{\Theta}$ in (2) has a simple interpretation: If for some $\theta \in \Theta$, the distribution P^θ differs significantly from the true distribution P^{θ_0}, as indicated by a large $d(P^\theta, P_N^{\theta_0})$, then we do not accept this choice of θ, and vice versa.

The way we construct $\tilde{\Theta}$ above bears resemblance to the RO and DRO literature. This literature considers decision-making under the worst-case scenario for problems with uncertain or unknown parameters, where the worst case is taken over a set that likely contains the true parameter value. This latter set is often known as the uncertainty set or ambiguity set. DRO, in particular, focuses on stochastic problems where the uncertainty is on some underlying probability distributions. Furthermore, data-driven RO or DRO refers to situations where the uncertainty set is constructed from the data, typically as a confidence region for the unknowns. Then, via an optimization problem, one translates the confidence guarantee from the uncertainty set to confidence bounds on the performance measure or objective of interest. In our discussion, the sets $\{Q : d(Q, P_N^{\theta_0}) \leq \eta\}$ and $\{\theta \in \Theta : d(P^\theta, P_N^{\theta_0}) \leq \eta\}$ can be viewed as uncertainty sets. Suppose we are interested in a quantity $\psi(P^{\theta_0})$ where ψ is a deterministic function, and then we may evaluate

$$\begin{aligned} &\text{maximize/minimize } \psi(P^\theta) \\ &\quad \text{subject to } d(P^\theta, P_N^{\theta_0}) \leq \eta \quad (\text{i.e., } \theta \in \tilde{\Theta}). \end{aligned} \tag{4}$$

Similar to the discussion above, if $\tilde{\Theta}$ is a $(1-\alpha)$-level confidence region, or equivalently that (3) holds, then the optimal values of the two optimization problems in (4) provide $(1-\alpha)$-level confidence upper and lower bounds to the true value $\psi(P^{\theta_0})$. Formulation (4), which aims to compute worst-case bounds under a confidence region, thus resembles closely the framework of data-driven RO (or DRO). Although we focus on the calibration of parameter θ in this paper, one may keep in mind that our framework can be generalized to tackle relevant problems in the form above.

4 ELIGIBILITY SET: CONSTRUCTION PROCEDURE

To construct an eligibility set with the guarantee presented in Theorem 1, we need to resolve two questions. First, how should we pick a proper discrepancy measure $d(\cdot,\cdot)$ and the associated η to achieve a given confidence level $1-\alpha$? Second, how can we determine whether $d(P^\theta, P_N^{\theta_0}) \leq \eta$ or not for any $\theta \in \Theta$, especially when the relationship between θ and P^θ is quite complicated?

For the first question, we acknowledge that there are many different ways to measure the discrepancy between two probability distributions, and also many of them could potentially be used here. Indeed, as long as the confidence property (3) is satisfied for some $\eta > 0$, then the pair (d, η) should satisfy our requirements. However, a good choice d should not only well capture the discrepancy between two distributions, but also efficient to compute (This is even more so if we were to tackle the optimization problem (4)). Similarly, η should be easy to calibrate to attain the confidence level $1-\alpha$. Based on these two considerations, a natural choice is

$$d(P_1, P_2) = \sup_{x \in \mathbb{R}} |F_1(x) - F_2(x)|$$

where P_1, P_2 are probability distributions and F_1, F_2 are respectively their cumulative distribution functions. Then we immediately get that

$$d(P^{\theta_0}, P_N^{\theta_0}) = \sup_{x \in \mathbb{R}} |F^{\theta_0}(x) - F_N^{\theta_0}(x)|,$$

which is exactly the the Kolmogorov-Smirnov (KS) statistic for the goodness-of-fit for F^{θ_0}. Correspondingly, we may choose $\eta = q_{1-\alpha}/\sqrt{N}$, where, $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of $\sup_{t \in [0,1]} |BB(t)|$ and $BB(\cdot)$ denotes a standard Brownian bridge. This choice satisfies the asymptotic version of the confidence property (3).

Now we address the second question on how to determine whether $d(P^\theta, P_N^{\theta_0}) \le \eta$ or not. We have assumed that P^θ comes potentially from a black-box, so that we should resort to sampling $Y_1^\theta, \cdots, Y_n^\theta \sim P^\theta$ to get P_n^θ to approximate P^θ. Correspondingly, we use F_n^θ to approximate F^θ, and in turn $d(P_n^\theta, P_N^{\theta_0})$ to approximate $d(P^\theta, P_N^{\theta_0})$. This approximation is utilizable if the simulation size n is much larger than the data size N, which we will justify in detail in Section 5. In other words, for this approach to work, we need enough computational capacity measured by the amount of simulation size we can obtain.

To summarize, we now choose $\tilde{\Theta}$ defined in (2) more explicitly as

$$\hat{\Theta} = \left\{ \theta \in \Theta : \sup_{x \in \mathbb{R}} |F_n^\theta(x) - F_N^{\theta_0}(x)| \le q_{1-\alpha}/\sqrt{N} \right\} \tag{5}$$

where $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of $\sup_{t\in[0,1]} |BB(t)|$.

We provide a more computationally friendly equivalent to (5), which is quite standard but we show here for user's convenience. The empirical distribution functions F_n^θ and $F_N^{\theta_0}$ are both non-decreasing staircase functions with $\lim_{x\to-\infty} F_n^\theta(x) = \lim_{x\to-\infty} F_N^{\theta_0}(x) = 0$ and $\lim_{x\to\infty} F_n^\theta(x) = \lim_{x\to\infty} F_N^{\theta_0}(x) = 1$. So it is easy to verify that

$$\begin{aligned} &\sup_{x\in\mathbb{R}} |F_n^\theta(x) - F_N^{\theta_0}(x)| \le q_{1-\alpha}/\sqrt{N} \\ \Leftrightarrow &|F_n^\theta(X_k) - F_N^{\theta_0}(X_k)| \le q_{1-\alpha}/\sqrt{N}, |F_n^\theta(X_k-) - F_N^{\theta_0}(X_k-)| \le q_{1-\alpha}/\sqrt{N}, \forall k = 1, \cdots, N \end{aligned}$$

where

$$F_n^\theta(u-) = \lim_{x\to u-} F_n^\theta(x) = \frac{1}{n}\sum_{j=1}^n I(Y_j^\theta < u), u \in \mathbb{R}$$

and similarly

$$F_N^{\theta_0}(u-) = \lim_{x\to u-} F_N^{\theta_0}(x) = \frac{1}{N}\sum_{k=1}^N I(X_k < u), u \in \mathbb{R}.$$

Therefore, (5) can be rewritten as

$$\hat{\Theta} = \left\{ \theta \in \Theta : |F_n^\theta(X_k) - F_N^{\theta_0}(X_k)| \le q_{1-\alpha}/\sqrt{N}, |F_n^\theta(X_k-) - F_N^{\theta_0}(X_k-)| \le q_{1-\alpha}/\sqrt{N}, \forall k = 1, \cdots, N \right\}. \tag{6}$$

Consequently, given any $\theta \in \Theta$, we may simulate $Y_1^\theta, \cdots, Y_n^\theta \sim P^\theta$ and readily check if it is inside $\hat{\Theta}$. If it is, then we say this θ is eligible. However, there are potentially infinitely many θ in Θ to test. To handle this, we may choose a finite number of values, say $\theta_1, \cdots, \theta_m \in \Theta$, as "representatives" and determine their eligibility. The resulting eligible points, collectively denoted

$$\hat{\Theta}_m = \left\{ \theta_1, \ldots, \theta_m \in \Theta : |F_n^{\theta_i}(X_k) - F_N^{\theta_0}(X_k)| \le q_{1-\alpha}/\sqrt{N}, |F_n^{\theta_i}(X_k-) - F_N^{\theta_0}(X_k-)| \le q_{1-\alpha}/\sqrt{N}, \forall k = 1, \cdots, N \right\},$$

provide a discretized approximation to $\hat{\Theta}$. Here, $\theta_1, \cdots, \theta_m$ can either be deterministic grid points in Θ or randomly sampled. For instance, when Θ is bounded, we may sample $\theta_1, \cdots, \theta_m \sim Unif(\Theta)$. Otherwise, we may sample θ_i's from some proper (heavy-tailed) distribution on Θ. In the high-dimensional case, $\theta_1, \ldots, \theta_m$ can be strategically chosen using stochastic gradient-based search with randomly sampled initial points.

Putting everything together, we propose Algorithm 1 to compute the eligibility set which can be utilized to calibrate the unknown parameter θ.

Algorithm 1 Constructing eligibility set of θ.

Require: The output data $X_1, \cdots, X_N$. The number of candidate θ's m. The simulation replication size n. The confidence level $1-\alpha$.

Ensure: An eligibility set $\hat{\Theta}_m$.

1: Generate $\theta_1, \cdots, \theta_m \in \Theta$;
2: **for** $i = 1, \cdots, m$ **do**
3: Generate a random sample $Y_1^{\theta_i}, \cdots, Y_n^{\theta_i} \sim P^{\theta_i}$;
4: **for** $k = 1, \cdots, N$ **do**
5: Compute $a_{i,k} = \left| \frac{1}{n}\sum_{j=1}^{n} I\left(Y_j^{\theta_i} \le X_k\right) - \frac{1}{N}\sum_{l=1}^{N} I(X_l \le X_k) \right|$;
6: Compute $b_{i,k} = \left| \frac{1}{n}\sum_{j=1}^{n} I\left(Y_j^{\theta_i} < X_k\right) - \frac{1}{N}\sum_{l=1}^{N} I(X_l < X_k) \right|$;
7: **end for**
8: **end for**
9: Construct $\hat{\Theta}_m = \{\theta_i : a_{i,k} \le q_{1-\alpha}/\sqrt{N}$ and $b_{i,k} \le q_{1-\alpha}/\sqrt{N}, \forall k = 1, \cdots, N\}$;
10: **return** The eligibility set $\hat{\Theta}_m$.

5 THEORETICAL STATISTICAL GUARANTEES

As discussed in Section 4, our choice of (d, η) satisfies the asymptotic version of (3), that is,

$$\underline{\lim}_{N\to\infty} \mathbb{P}\left(\sup_{x\in\mathbb{R}} |F^{\theta_0}(x) - F_N^{\theta_0}(x)| \le q_{1-\alpha}/\sqrt{N} \right) \ge 1-\alpha.$$

However, in Algorithm 1, we substitute F^{θ_0} with $F_n^{\theta_0}$, and thus additional randomness is introduced. To justify our approach, we need to show that a similar asymptotic confidence property holds with this substitution. Theorem 2 demonstrates a sufficient condition for this.

Theorem 2 Suppose that $X_1, \cdots, X_N$ is an i.i.d. random sample from P^{θ_0} and $Y_1^{\theta_0}, \cdots, Y_n^{\theta_0}$ is another i.i.d. random sample from P^{θ_0}. $F_N^{\theta_0}$ and $F_n^{\theta_0}$ are respectively the empirical distribution functions of the two random samples. F^{θ_0} denotes the cumulative distribution function of P^{θ_0}. If $n = \omega(N)$ as $N \to \infty$, then

$$\underline{\lim}_{n,N\to\infty} \mathbb{P}\left(\sup_{x\in\mathbb{R}} |F_n^{\theta_0}(x) - F_N^{\theta_0}(x)| \le q_{1-\alpha}/\sqrt{N} \right) \ge 1-\alpha.$$

Proof. Since

$$\begin{aligned}
\sup_{x\in\mathbb{R}} |F_n^{\theta_0}(x) - F_N^{\theta_0}(x)| &\le \sup_{x\in\mathbb{R}} \left(|F_n^{\theta_0}(x) - F^{\theta_0}(x)| + |F^{\theta_0}(x) - F_N^{\theta_0}(x)| \right) \\
&\le \sup_{x\in\mathbb{R}} |F_n^{\theta_0}(x) - F^{\theta_0}(x)| + \sup_{x\in\mathbb{R}} |F_N^{\theta_0}(x) - F^{\theta_0}(x)|,
\end{aligned}$$

we get that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{x\in\mathbb{R}} |F_n^{\theta_0}(x) - F_N^{\theta_0}(x)| > q_{1-\alpha}/\sqrt{N} \right) \\
\le &\mathbb{P}\left(\sup_{x\in\mathbb{R}} |F_n^{\theta_0}(x) - F^{\theta_0}(x)| + \sup_{x\in\mathbb{R}} |F_N^{\theta_0}(x) - F^{\theta_0}(x)| > q_{1-\alpha}/\sqrt{N} \right) \\
\le &\mathbb{P}\left(\sup_{x\in\mathbb{R}} |F_n^{\theta_0}(x) - F^{\theta_0}(x)| > \lambda q_{1-\alpha}/\sqrt{N} \right) + \mathbb{P}\left(\sup_{x\in\mathbb{R}} |F_N^{\theta_0}(x) - F^{\theta_0}(x)| > (1-\lambda) q_{1-\alpha}/\sqrt{N} \right)
\end{aligned}$$

for any $\lambda \in (0,1)$. It is known that $\sqrt{n}\sup_{x\in\mathbb{R}}|F_n^{\theta_0}(x)-F^{\theta_0}(x)| \Rightarrow \sup_{t\in\mathbb{R}}|BB(F^{\theta_0}(t))|$ and similarly $\sqrt{N}\sup_{x\in\mathbb{R}}|F_N^{\theta_0}(x)-F^{\theta_0}(x)| \Rightarrow \sup_{t\in\mathbb{R}}|BB(F^{\theta_0}(t))|$ as $n,N\to\infty$ where $\Rightarrow$ stands for convergence in distribution. $n=\omega(N)$ as $N\to\infty$ implies that $n/N\to\infty$ as $N\to\infty$, and thus

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}}|F_n^{\theta_0}(x)-F^{\theta_0}(x)|>\lambda q_{1-\alpha}/\sqrt{N}\right)=\mathbb{P}\left(\sqrt{n}\sup_{x\in\mathbb{R}}|F_n^{\theta_0}(x)-F^{\theta_0}(x)|>\lambda q_{1-\alpha}\sqrt{n/N}\right)\to 0$$

as $N\to\infty$ for any $\lambda\in(0,1)$. Hence,

$$\begin{aligned}\overline{\lim_{N\to\infty}}\,\mathbb{P}\left(\sup_{x\in\mathbb{R}}|F_n^{\theta_0}(x)-F_N^{\theta_0}(x)|>q_{1-\alpha}/\sqrt{N}\right) &\le \overline{\lim_{N\to\infty}}\,\mathbb{P}\left(\sup_{x\in\mathbb{R}}|F_N^{\theta_0}(x)-F^{\theta_0}(x)|>(1-\lambda)q_{1-\alpha}/\sqrt{N}\right)\\ &=\mathbb{P}\left(\sup_{t\in\mathbb{R}}|BB(F^{\theta_0}(t))|>(1-\lambda)q_{1-\alpha}\right)\\ &\le\mathbb{P}\left(\sup_{t\in[0,1}|BB(t)|>(1-\lambda)q_{1-\alpha}\right).\end{aligned}$$

By the arbitrariness of λ and the definition of $q_{1-\alpha}$, we finally get that

$$\overline{\lim_{N\to\infty}}\,\mathbb{P}\left(\sup_{x\in\mathbb{R}}|F_n^{\theta_0}(x)-F_N^{\theta_0}(x)|>q_{1-\alpha}/\sqrt{N}\right)\le\mathbb{P}\left(\sup_{t\in[0,1]}|BB(t)|\ge q_{1-\alpha}\right)=\alpha.$$

□

Note that the probability of interest in Theorem 2 relates closely to the type I error in statistical inference. According to the theorem, conditional on being chosen as a "representative", θ_0 is correctly selected to put inside the eligibility set with probability at least $1-\alpha$ asymptotically. In this sense, we may regard the discretized eligibility set $\hat{\Theta}_m$ as (upon a suitable interpolation) an approximate confidence region for θ_0.

On the other hand, we are also interested in the type II error, that is, the probability that some representative does not belong to the true (discretized) eligibility set yet is accepted. The type II error here governs the conservativeness level of $\hat{\Theta}_m$ as an approximate confidence region. Note that, in the case $F^{\theta}\neq F^{\theta_0}$ and as n and N go to infinity, the difference between F_n^{θ} and $F_N^{\theta_0}$ gradually converges to the difference between F^{θ} and F^{θ_0}, which is a positive constant. Theorem 3 captures the rate under which this occurs and provides an upper bound for the probability that $F^{\theta}\neq F^{\theta_0}$ yet θ is eligible.

Theorem 3 Suppose that $X_1,\cdots,X_N$ is an i.i.d. random sample from P^{θ_0} and $Y_1^{\theta},\cdots,Y_n^{\theta}$ is an i.i.d. random sample from P^{θ}. $F_N^{\theta_0}$ and F_n^{θ} are respectively the empirical distribution functions of the two random samples. F^{θ_0} and F^{θ} denote the cumulative distribution functions of P^{θ_0} and P^{θ}. Suppose that $\sup_{x\in\mathbb{R}}|F^{\theta}(x)-F^{\theta_0}(x)|>0$. For any $\varepsilon_1,\varepsilon_2>0$ such that $\varepsilon_1+\varepsilon_2<\sup_{x\in\mathbb{R}}|F^{\theta}(x)-F^{\theta_0}(x)|$, if

$$N>\left(\frac{q_{1-\alpha}}{\sup_{x\in\mathbb{R}}|F^{\theta}(x)-F^{\theta_0}(x)|-\varepsilon_1-\varepsilon_2}\right)^2,$$

then

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}}|F_n^{\theta}(x)-F_N^{\theta_0}(x)|\le q_{1-\alpha}/\sqrt{N}\right)\le 2\left(e^{-2n\varepsilon_1^2}+e^{-2N\varepsilon_2^2}\right).$$

Proof. We know that

$$|F_n^{\theta}(x)-F_N^{\theta_0}(x)|\ge|F^{\theta}(x)-F^{\theta_0}(x)|-|F_n^{\theta}(x)-F^{\theta}(x)|-|F_N^{\theta_0}(x)-F^{\theta_0}(x)|,$$

and then we get that

$$\sup_{x\in\mathbb{R}}|F_n^{\theta}(x)-F_N^{\theta_0}(x)| \geq \sup_{x\in\mathbb{R}}|F^{\theta}(x)-F^{\theta_0}(x)| - \sup_{x\in\mathbb{R}}|F_n^{\theta}(x)-F^{\theta}(x)| - \sup_{x\in\mathbb{R}}|F_N^{\theta_0}(x)-F^{\theta_0}(x)|.$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{x\in\mathbb{R}}|F_n^{\theta}(x)-F_N^{\theta_0}(x)| \leq q_{1-\alpha}/\sqrt{N}\right)\\
\leq&\mathbb{P}\left(\sup_{x\in\mathbb{R}}|F^{\theta}(x)-F^{\theta_0}(x)| - \sup_{x\in\mathbb{R}}|F_n^{\theta}(x)-F^{\theta}(x)| - \sup_{x\in\mathbb{R}}|F_N^{\theta_0}(x)-F^{\theta_0}(x)| \leq q_{1-\alpha}/\sqrt{N}\right)\\
\leq&\mathbb{P}\left(\sup_{x\in\mathbb{R}}|F_n^{\theta}(x)-F^{\theta}(x)| > \varepsilon_1\right) + \mathbb{P}\left(\sup_{x\in\mathbb{R}}|F_N^{\theta_0}(x)-F^{\theta_0}(x)| > \varepsilon_2\right).
\end{aligned}$$

The last inequality is obtained using the fact that $q_{1-\alpha}/\sqrt{N} < \sup_{x\in\mathbb{R}}|F^{\theta}(x)-F^{\theta_0}(x)| - \varepsilon_1 - \varepsilon_2$ under the conditions in the theorem. By the refined Dvoretzky–Kiefer–Wolfowitz (DKW) inequality (Kosorok 2008), we get that

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}}|F_n^{\theta}(x)-F^{\theta}(x)| > \varepsilon_1\right) \leq 2e^{-2n\varepsilon_1^2}, \mathbb{P}\left(\sup_{x\in\mathbb{R}}|F_N^{\theta_0}(x)-F^{\theta_0}(x)| > \varepsilon_2\right) \leq 2e^{-2N\varepsilon_2^2},$$

which concludes the proof. □

With Theorem 3, we may easily develop a type II error guarantee for Algorithm 1. More specifically, we can bound the probability that $\hat{\Theta}_m$ contains some $\theta \in \Theta$ such that $\sup_{x\in\mathbb{R}}|F^{\theta}(x)-F^{\theta_0}(x)|$ is large. Theorem 4 shows the details.

Theorem 4 We follow Algorithm 1 to obtain $\hat{\Theta}_m$. For any $\varepsilon, \varepsilon_1, \varepsilon_2 > 0$ such that $\varepsilon_1 + \varepsilon_2 < \varepsilon$, if

$$N > \left(\frac{q_{1-\alpha}}{\varepsilon - \varepsilon_1 - \varepsilon_2}\right)^2,$$

then

$$\mathbb{P}\left(\exists i = 1,\cdots,m \text{ s.t. } \sup_{x\in\mathbb{R}}|F^{\theta_i}(x)-F^{\theta_0}(x)| > \varepsilon, \theta_i \in \hat{\Theta}_m\right) \leq 2m\left(e^{-2n\varepsilon_1^2} + e^{-2N\varepsilon_2^2}\right).$$

Proof. By applying Theorem 3, we have that

$$\begin{aligned}
&\mathbb{P}\left(\exists i = 1,\cdots,m \text{ s.t. } \sup_{x\in\mathbb{R}}|F^{\theta_i}(x)-F^{\theta_0}(x)| > \varepsilon, \theta_i \in \hat{\Theta}_m\right)\\
\leq&\sum_{i=1}^{m}\mathbb{P}\left(\sup_{x\in\mathbb{R}}|F^{\theta_i}(x)-F^{\theta_0}(x)| > \varepsilon, \theta_i \in \hat{\Theta}_m\right)\\
\leq&\sum_{i=1}^{m}\mathbb{P}\left(\sup_{x\in\mathbb{R}}|F_n^{\theta_i}(x)-F_N^{\theta_0}(x)| \leq q_{1-\alpha}/\sqrt{N}\middle|\sup_{x\in\mathbb{R}}|F^{\theta_i}(x)-F^{\theta_0}(x)| > \varepsilon\right)\\
\leq&2m\left(e^{-2n\varepsilon_1^2} + e^{-2N\varepsilon_2^2}\right).
\end{aligned}$$

□

Theorems 2 and 4 together justify that the eligibility set $\hat{\Theta}_m$ given by Algorithm 1 can help us calibrate the unknown parameter θ, especially when m, n, N are sufficiently large and chosen properly. More specifically, the following corollaries provide instructions on how to choose the sizes.

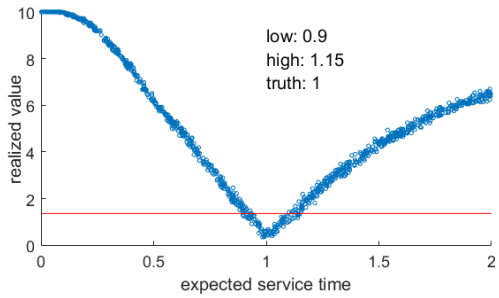


Figure 1: Apply Algorithm 1 to calibrate μ ($\sqrt{N}\sup_{x\in\mathbb{R}}|F_n^{\theta_i}(x)-F_N^{\theta_0}(x)|$ against θ_i).

Corollary 5 We follow Algorithm 1 to obtain $\hat{\Theta}_m$. If $\log m = o(N)$ and $n = \Omega(N)$ as $N\to\infty$, then for any $\varepsilon > 0$,

$$\lim_{m,n,N\to\infty}\mathbb{P}\left(\exists i = 1,\cdots,m \text{ s.t. } \sup_{x\in\mathbb{R}}|F^{\theta_i}(x)-F^{\theta_0}(x)| > \varepsilon, \theta_i\in\hat{\Theta}_m\right) = 0.$$

Corollary 6 We follow Algorithm 1 to obtain $\hat{\Theta}_m$. If $m = o(N)$ and $n = \Omega(N)$ as $N\to\infty$, then

$$\lim_{m,n,N\to\infty}\mathbb{P}\left(\exists i = 1,\cdots,m \text{ s.t. } \sup_{x\in\mathbb{R}}|F^{\theta_i}(x)-F^{\theta_0}(x)| > \sqrt{\log m/m}, \theta_i\in\hat{\Theta}_m\right) = 0.$$

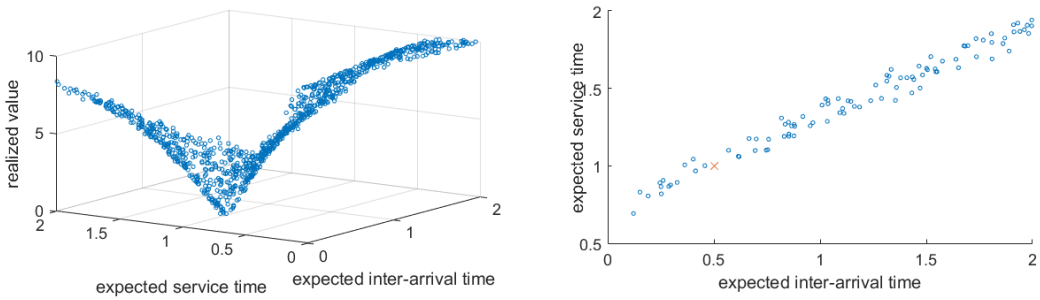
6 NUMERICAL RESULTS

We conduct some numerical experiments to validate our approach. We consider a $G/G/1$ queuing system where the inter-arrival time is known to follow a $Gamma(\lambda,1)$ distribution while the service time follows a $Weibull(\mu,1)$ distribution. Then the expected inter-arrival time and the expected service time are respectively λ and μ. We assume that we know $\lambda,\mu\in(0,2)$ and that the true values are $\lambda_0 = 1/2$ and $\mu_0 = 1$. Suppose that we are only given a random sample of the average sojourn time of the first $c = 10$ customers in the system. Clearly the distribution of this output quantity is uniquely determined by λ and μ, and conversely, given the values of λ,μ, the average sojourn time is easy to simulate. Thus this parameter calibration problem falls into our settings.

We first consider the case where λ is known and we want to calibrate μ (i.e., $\theta := \mu$). Suppose we have $N = 100$ output observations generated from the unknown $\mu_0 = 1$, which we mimic by running simulation. We follow Algorithm 1 by setting $m = n = 1000$ and sampling $\theta_1,\cdots,\theta_m \sim Unif(0,2)$. For each θ_i, we simulate $n = 1000$ versions of average sojourn time with $\lambda = 1/2$ and $\mu = \theta_i$, and then compute $\sup_{x\in\mathbb{R}}|F_n^{\theta_i}-F_N^{\theta_0}|$. We choose $1-\alpha = 0.95$. By comparing the realized value, $\sqrt{N}\sup_{x\in\mathbb{R}}|F_n^{\theta_i}-F_N^{\theta_0}|$, with $q_{1-\alpha}$, we determine whether θ_i is eligible. Figure 1 plots $\sqrt{N}\sup_{x\in\mathbb{R}}|F_n^{\theta_i}(x)-F_N^{\theta_0}(x)|$ against θ_i, and the red horizontal line corresponds to $q_{1-\alpha}\approx 1.36$. Thus the dots below the red line constitute the eligibility set $\hat{\Theta}_m$. From the graph, we find that all the eligible θ_i's are relatively close to the true value $\theta_0 = 1$. More specifically, the smallest and the largest values in the eligibility set are respectively 0.90 and 1.15, which has greatly shrunk the original a priori known range $(0,2)$.

Additionally, in the above setting, we may construct an approximate confidence interval with the smallest and the largest values in $\hat{\Theta}_m$. After repeating the above process for $L = 100$ times, we find that among the confidence intervals that we obtain, 99 cover the true value 1. Given that the target confidence level is 95%, the coverage probability is satisfactory, though a little conservative.

Now we consider the case that both λ and μ are unknown and we calibrate them simultaneously (i.e., $\theta := (\lambda,\mu)$). This problem is non-identifiable. It is easy to verify that, as λ and μ grow simultaneously in a certain way, the distribution of the average sojourn time of the first $c = 10$ customers does not change. Similar to the previous setup, given an output data set of size $N = 100$, we run Algorithm 1 with $m = n = 1000$, $1-\alpha = 0.95$, but now we generate $\theta_1,\cdots,\theta_m \sim Unif(0,2)^2$. In Figure 2, we show the



(a) $\sqrt{N}\sup_{x\in\mathbb{R}}|F_n^{\theta_i}(x)-F_N^{\theta_0}(x)|$ against θ_i.

(b) The eligibility set $\hat{\Theta}_m$.

Figure 2: Apply Algorithm 1 to calibrate (λ,μ).

three-dimensional scatter plot of $\sqrt{N}|F_n^{\theta_i}(x)-F_N^{\theta_0}(x)|$ against $\theta_i=(\lambda_i,\mu_i)$ as well as the two-dimensional visualization of the eligibility set $\hat{\Theta}_m$. Particularly, in Figure 2(b), the red cross mark denotes the true value $(1/2,1)$, which lies approximately in the band. We therefore see that, when faced with non-identifiability, our approach can still result in a region that contains the truth, where the size of the region is related to the degree of non-identifiability (i.e., the range of parameters that exactly match in terms of the output distribution). This demonstrates how our approach can bypasss non-identifiability and provides empirical support to our proposed framework.

7 CONCLUSION

In this paper, we have studied a framework to calibrate unknown parameter only using the output-level data, which only makes very mild assumptions on the model. Indeed, theoretically our framework may deal with the situation where the relationship between the input parameter and the output distribution is complicated or even a black box.

In particular, we consider a finite number of values of the parameter as "representatives" and then for each of them, we generate a simulated sample, which may be compared to the real data using the KS statistic. The eligibility set consists of the "representatives" whose simulated sample is "close enough" to the true data sample, and then we employ this set to estimate the parameter of interest.

We have also developed statistical guarantees for this procedure. To be more specific, we analyze the type I error as well as the type II error of the eligibility set in terms of the data size and the simulation size. These theoretical results may provide some guidance on how to choose the simulation size according to the data size. Moreover, the numerical results support that our framework works well even under complicated settings.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the National Science Foundation under grants CAREER CMMI-1834710 and IIS-1849280.

REFERENCES

Avellaneda, M., R. Buff, C. Friedman, N. Grandechamp, L. Kruk, and J. Newman. 2001. "Weighted Monte Carlo: a New Technique for Calibrating Asset-Pricing Models". *International Journal of Theoretical and Applied Finance* 4(01):91–119.

Balci, O., and R. G. Sargent. 1982. "Some Examples of Simulation Model Validation Using Hypothesis Testing". In *Proceedings of the 14th Conference on Winter Simulation - Volume 2*, WSC '82, 621–629: Winter Simulation Conference.

Barton, R. R. 2012. "Tutorial: Input Uncertainty in Output Analysis". In *Proceedings of the 2012 Winter Simulation Conference (WSC)*, 1–12. IEEE.

Ben-Tal, A., L. El Ghaoui, and A. Nemirovski. 2009. *Robust Optimization*, Volume 28. Princeton University Press.

Bertsimas, D., D. B. Brown, and C. Caramanis. 2011. "Theory and Applications of Robust Optimization". *SIAM review* 53(3):464–501.

Chick, S. E. 2006. "Bayesian Ideas and Discrete Event Simulation: Why, What and How". In *Proceedings of the 2006 winter simulation conference*, 96–106. IEEE.

Currin, C., T. Mitchell, M. Morris, and D. Ylvisaker. 1991. "Bayesian Prediction of Deterministic Functions, with Applications to the Design and Analysis of Computer Experiments". *Journal of the American Statistical Association* 86(416):953–963.

Delage, E., and Y. Ye. 2010. "Distributionally Robust Optimization under Moment Uncertainty with Application to Data-Driven Problems". *Operations research* 58(3):595–612.

Glasserman, P., and B. Yu. 2005. "Large Sample Properties of Weighted Monte Carlo Estimators". *Operations Research* 53(2):298–312.

Goeva, A., H. Lam, H. Qian, and B. Zhang. 2019, aug. "Optimization-Based Calibration of Simulation Input Models". *Operations Research* 67(5):1362–1382.

Goeva, A., H. Lam, and B. Zhang. 2014. "Reconstructing Input Models via Simulation Optimization". In *Proceedings of the Winter Simulation Conference 2014*, 698–709. IEEE.

Henderson 2003. "Input Model Uncertainty: Why Do We Care and What Should We Do about It?". In *Proceedings of the 2003 Winter Simulation Conference, 2003.*, Volume 1, 90–100 Vol.1.

Kennedy, M. C., and A. O'Hagan. 2001. "Bayesian Calibration of Computer Models". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63(3):425–464.

Kleijnen, J. P. 1995. "Verification and Validation of Simulation Models". *European journal of operational research* 82(1):145–162.

Kosorok, M. R. 2008. *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer.

Kraan, B., and T. Bedford. 2005. "Probabilistic Inversion of Expert Judgments in the Quantification of Model Uncertainty". *Management science* 51(6):995–1006.

Lam, H. 2016. "Advanced Tutorial: Input Uncertainty and Robust Analysis in Stochastic Simulation". In *2016 Winter Simulation Conference (WSC)*, 178–192. IEEE.

Lam, H., X. Zhang, and M. Plumlee. 2017. "Improving Prediction from Stochastic Simulation via Model Discrepancy Learning". In *2017 Winter Simulation Conference (WSC)*, 1808–1819. IEEE.

Nelson, B. L. 2016, feb. "'Some Tactical Problems in Digital Simulation' for the Next 10 Years". *Journal of Simulation* 10(1):2–11.

Plumlee, M., and H. Lam. 2017. "An Uncertainty Quantification Method for Inexact Simulation Models". *arXiv preprint arXiv:1707.06544*.

Sargent, R. G. 2010. "Verification and Validation of Simulation Models". In *Proceedings of the 2010 winter simulation conference*, 166–183. IEEE.

Schruben, L. W. 1980, mar. "Establishing the Credibility of Simulations". *Simulation* 34(3):101–105.

Song, E., B. L. Nelson, and C. D. Pegden. 2014. "Advanced Tutorial: Input Uncertainty Quantification". In *Proceedings of the Winter Simulation Conference 2014*, 162–176. IEEE.

Tarantola, A. 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation*, Volume 89. siam.

Wiesemann, W., D. Kuhn, and M. Sim. 2014. "Distributionally Robust Convex Optimization". *Operations Research* 62(6):1358–1376.

Zhang, X., and L. Zou. 2016. "Simulation Metamodeling in the Presence of Model Inadequacy". In *2016 Winter Simulation Conference (WSC)*, 566–577. IEEE.

AUTHOR BIOGRAPHIES

YUANLU BAI is a PhD student in the Department of Industrial Engineering and Operations Research at Columbia University. She received an M.S. in Operations Research from Columbia University, and a B.S. in Statistics as well as a B. Ec in Economics from Peking University. Her research interest lies in stochastic simulation, such as uncertainty quantification and extreme risk analysis. Her email address is yb2436@columbia.edu.

HENRY LAM is an Associate Professor in the Department of Industrial Engineering and Operations Research at Columbia University. He received his Ph.D. degree in statistics from Harvard University in 2011, and was on the faculty of Boston University and the University of Michigan before joining Columbia in 2017. His research focuses on Monte Carlo simulation, uncertainty quantification, risk analysis, and stochastic and robust optimization. His email address is henry.lam@columbia.edu.