Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

STATISTICAL UNCERTAINTY QUANTIFICATION FOR EXPENSIVE BLACK-BOX MODELS: METHODOLOGIES AND INPUT UNCERTAINTY APPLICATIONS

Henry Lam

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

ABSTRACT

This tutorial reviews methodologies for quantifying statistical uncertainty in computationally expensive black-box models, which arise frequently in data-driven simulation analyses under input uncertainty. When facing these models, it can be difficult to run repeated evaluations due to computation cost, and also to obtain auxiliary information such as gradients due to analytical intractability, thus rendering many traditional statistical approaches challenging to apply. We describe several lines of approaches to resolve these challenges, including data-splitting methods based on batching variants, a recent so-called cheap bootstrap approach, and subsampling schemes. We discuss the applications of these approaches to simulation, including problems suffering from both aleatory error exhibited via Monte Carlo noises and epistemic error stemming from the input uncertainty.

1 INTRODUCTION

This tutorial focuses on uncertainty quantification that is broadly speaking computation-aware. By uncertainty quantification, we will refer throughout this tutorial to the construction of statistically valid confidence intervals (CIs) which, along the way, would also require understanding and estimating sampling variabilities and distributions. While the general problem of uncertainty quantification is ubiquitous in statistics, it is also at the core of output analysis in simulation. The latter has implications in assessing the reliability of model predictions and allocating downstream resources. For instance, in a feasibility study, even if the point estimate of a target system performance measure lies in the feasible region, a long confidence interval that reaches part of the infeasible region would alert that the actual value of the performance measure can be infeasible. In this way, it informs the need to collect more data in order to obtain a more affirmative feasibility determination, or otherwise the decision-making is understood to bear a significant level of ambiguity or risk of incorrect determination.

We are especially interested in models that are *computationally expensive* and *black-box*. Here, computational expensiveness means that it is costly to (re-)run the model evaluation, and black-box means that it is difficult or even impossible to obtain auxiliary model information such as gradients. Uncertainty quantification on such models arise commonly in the so-called *input uncertainty* problem, which has gathered fast-growing interests in recent years among the stochastic simulation community (e.g., the surveys Barton et al. 2002; Henderson 2003; Chick 2006; Barton 2012; Song et al. 2014; Lam 2016a; Corlu et al. 2020; Barton et al. 2022). Input uncertainty refers to the errors on simulation outputs that are propagated from the statistical noises in calibrating unknown-in-advance input parameters or distributions via external data. For simulation models of large sophisticated systems, quantifying these output errors, or in other words tackling the input uncertainty problem, entails statistical uncertainty quantification on expensive black-box models with the difficulties that we described above.

979-8-3503-6966-3/23/$31.00 ©2023 IEEE

1.1 Problem Formulation and Motivating Examples

To make our discussion more concrete, let us provide a basic mathematical setup. Denote $\psi(P) \in \mathbb{R}$ as a target quantity of interest, which depends on the underlying true distribution of the randomness P. We consider the data-driven setting where P is unknown but can be observed via data. We aim for a valid CI of $\psi(P)$ at level $1-\alpha$ (e.g., 95%), which is an interval $[L,U]$ such that

$$\mathbb{P}(\psi(P) \in [L,U]) \approx 1-\alpha \tag{1}$$

where the probability $\mathbb{P}$ is with respect to the data, and the approximation "$\approx$" will be made more precise later.

Our premise is that $\psi(Q)$, given any inserted distribution Q, is computationally expensive to evaluate. Moreover, it is a black-box evaluation in that we can only evaluate $\psi(Q)$ given Q, but not other quantities such as the gradient of $\psi(Q)$ with respect to Q (defined in some appropriate sense). We present a simple example to illustrate how our presented setup applies to the input uncertainty problem:

Example 1 (Long-run system simulation under input uncertainty) 'Consider a queueing system with a known Poisson arrival process, but the service time distribution is only historically observed. Our target quantity of interest is the steady-state average waiting time of the true system. In this case, the target quantity, which we denote as $\psi(P)$, depends on the service time distribution which we denote as P (we do not view the interarrival time distribution as part of P here because it is known and hence absorbed into the evaluation of $\psi(\cdot)$). We aim to construct a CI for the true value of $\psi(P)$. Here, the quantity $\psi(Q)$ for a given Q is expensive to evaluate because it requires running the system simulation for a very long time horizon (and we assume this time horizon is long enough that the transient effect is negligible). Moreover, the gradient information of $\psi(Q)$ with respect to Q is captured by the so-called influence function (Hampel et al. 2011) for a steady-state quantity and can be challenging to obtain mathematically and to compute (Lam 2016b).

The above discussion and example take the viewpoint that the evaluation of $\psi(\cdot)$ is deterministic. However, in stochastic simulation, many problems in fact involve $\psi(\cdot)$ that can only be noisily evaluated. This occurs generically when the target quantity of interest is a transient performance measure in the form of a summary statistic such as moment, where evaluating $\psi(Q)$ for a given Q would require repeatedly simulating system trajectories and then taking the sample moment of all the realized outputs. This adds challenges to uncertainty quantification due to the interacting impacts from two noise sources, one from the input data and one from the Monte Carlo computation. Quantifying the overall uncertainty and CI construction requires accounting for both of these errors, i.e., a $(1-\alpha)$-level CI needs to satisfy (1), but now $\mathbb{P}$ is with respect to both the input data and the Monte Carlo computation noises.

In this latter setup, the statistical error in calibrating the input distribution P is sometimes known as the *epistemic* or extrinsic uncertainty, while the Monte Carlo computation noise in evaluating ψ is known as the *aleatory*, intrinsic, or stochastic uncertainty. Here, the "expensiveness" of the model comes from the need to repeatedly run the Monte Carlo, and is a more subtle concept than in deterministic simulation as the level of expensiveness now relates to the Monte Carlo size which in turn affects the noise level. In fact, we will see later that naive approaches to construct valid CIs would require the Monte Carlo size to scale at least linearly with the input data size (Lam and Qian 2022). In other words, the model computation effort intertwines with the external data size, a phenomenon that distinguishes stochastic simulation rather uniquely from deterministic simulation problems. Let us close this discussion with a simple example:

Example 2 (Transient performance analysis under input uncertainty) Consider the same setup as Example 1, except that the target quantity of interest $\psi(P)$ is the expectation of the average waiting time over a prescribed finite time horizon. Here, to evaluate $\psi(Q)$ for any given Q, we would generate say R independent system trajectories, each with a realized average waiting time over the prescribed horizon, say $\hat{\psi}_r(Q)$, for $r = 1,\ldots,R$. Then we would output $\hat{\psi}(Q) = (1/R)\sum_{r=1}^{R} \hat{\psi}_r(Q)$ as an approximation of $\psi(Q)$. The computation cost in evaluating $\psi(\cdot)$ is thus controlled by the Monte Carlo size R.

1.2 Scope and Generalizations

Our goal is to present a range of methods to construct CIs for problems like Examples 1 and 2, with *low computation* and *no auxiliary information*. Specifically, we will present three lines of methods:

Data-Splitting: This approach, under the general umbrella of so-called cancellation methods, divides data into batches and conducts inference by suitably aggregating these batch estimates. The number of batches dictates the number of model evaluation, and so when the number of batches is small, so is the computation effort. This approach encompasses several variants including batching, sectioning, batching-sectioning hybrid, batched jackknife and overlapping batching.

Cheap Bootstrap: This is a recently proposed bootstrap alternative where, instead of using the distributional approximation of the resample to original sample as in conventional bootstraps, it uses sample-resample independence as a driver to derive pivotal statistics. By coupling with asymptotic normality, this approach allows to use very few resamples and hence model evaluations to generate valid CIs, as opposed to a large number of resamples required to estimate distributional summary quantities in the conventional bootstraps.

Subsampling: This generally refers to bootstrap schemes with a resample size that is smaller than the original data size, and constructs CIs by exploiting the scaling relation of corresponding limit theorems. While evaluating model with a smaller data size can already save computation in some problems, our focus here is to use it as a mechanism to save the overall computation budget in nested simulation procedures used to handle the joint effect of aleatory and epistemic uncertainties in input uncertainty problems.

We will detail the first two methodological lines above which apply to both deterministic simulation (Section 3) and problems with aleatory uncertainty (Section 4). Then, towards the end of this tutorial, we will describe how the third methodological line applies to handle aleatory uncertainty (Section 4.3). Before we discuss these approaches, we will first explain the challenges faced by existing statistical methods (Section 2). While in this tutorial we will focus only on the essential concepts and the pedagogical Examples 1 and 2, we comment some possible generalizations in multiple directions: 1) In addition to simulation input uncertainty, expensive models also appear frequently in modern statistical and machine learning problems where model evaluation requires solving large mathematical programs or running extensive iterative procedures, and our methodologies could be relevant to these problems. 2) Our setups and methodologies apply to problems involving multiple input distributions and correspondingly multiple separate data sets. These problems are common in simulation input uncertainty as a stochastic system usually has multiple input sources (e.g., interarrival, service, multi-class customers in a queueing network). 3) While we will confine our discussion to i.i.d. data, we can work with dependent data with proper modifications and additional assumptions. 4) We will focus on a nonparametric framework (i.e., use the empirical distribution from data to estimate P), but the same principles apply parametrically (i.e., P estimated within a parametric family). Lastly, numerical results in support of our reviewed methods can be found in the respective references in the sequel.

We use the following notations throughout this tutorial. For any sequences a and b both depending on parameter n, we say $a = O(b)$ if $|a/b| \leq C$ for some constant $C > 0$ for all sufficiently large n, $a = o(b)$ if $a/b \to 0$ as $n \to \infty$, $a = \Omega(b)$ if $|a/b| \geq C$ for some constant $C > 0$ for all sufficiently large n, and $a = \omega(b)$ if $|a/b| \to \infty$ as $n \to \infty$. We use $A = O_p(b)$ to represent a random variable A that has stochastic order at least b, i.e., for any $\varepsilon > 0$, there exists $M, N > 0$ such that $P(|A/b| \leq M) > 1 - \varepsilon$ for $n > N$. We use $A = o_p(b)$ to represent a random variable A that has stochastic order less than b, i.e., $A/b \xrightarrow{p} 0$.

2 COMMON STATISTICAL APPROACHES AND CHALLENGES

Let us first present some common statistical inference approaches and explain their bottlenecks in handling expensive black-box models. We first focus on the basic case where computing ψ does not involve Monte Carlo noise (but nonetheless is expensive) in Section 2.1, followed by the case with both aleatory and epistemic uncertainties in Section 2.2.

2.1 Deterministic Model Evaluation

Using the notation in the introduction, suppose we have i.i.d. data $\{X_1,\ldots,X_n\} \in \mathscr{X}$ of size n drawn from the distribution P. To estimate the target quantity $\psi(P)$, a natural point estimate is $\psi(\hat{P}_n)$, where $\hat{P}_n(\cdot) := (1/n)\sum_{i=1}^n I(X_i \in \cdot)$ is the empirical distribution and $I(\cdot)$ denotes the indicator function. Under mild assumptions (non-degenerate Hadamard derivative of ψ; Van der Vaart 2000), we have the central limit theorem (CLT)

$$\sqrt{n}(\psi(\hat{P}_n) - \psi(P)) \Rightarrow N(0,\sigma^2) \tag{2}$$

for some $\sigma^2 > 0$. Here $N(0,\sigma^2)$ is, roughly speaking, the Hadamard derivative of ψ evaluated at the limit of the empirical process $\sqrt{n}(\hat{P}_n - P)$. More transparently, if ψ satisfies a "Taylor expansion" in the form

$$\psi(\hat{P}_n) - \psi(P) = \int IF_P(x) d(\hat{P}_n - P)(x) + o_p\left(\frac{1}{\sqrt{n}}\right) \tag{3}$$

for some function $IF_P(\cdot) : \mathscr{X} \to \mathbb{R}$, then (2) holds with $\sigma^2 = Var_P(IF_P(X))$, where $Var_P(\cdot)$ denotes the variance with X generated under P. Here, $IF_P(\cdot)$ is called the influence function (Hampel, Ronchetti, Rousseeuw, and Stahel 2011) and can be viewed as a Gateaux derivative of ψ (Serfling 2009 §6).

In view of (2), a $(1-\alpha)$-level CI for $\psi(P)$ can be obtained as

$$\left[\psi(\hat{P}_n) - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \psi(\hat{P}_n) + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \tag{4}$$

where $z_{1-\alpha/2}$ is the critical value given by the $(1-\alpha/2)$-th quantile of standard normal. It can be shown that the coverage probability, i.e., probability of $\psi(P)$ lying in (4), converges to exactly $1-\alpha$ as the data size n grows to ∞. We call such a CI *asymptotically exact*. Note that despite the clear form of (4), it might not be easy to obtain σ. To this end, there are two major statistical approaches.

Delta method: This approach estimates σ directly by utilizing its analytical expression. For instance, if we know the closed-form of the influence function, then we can use the empirical counterpart of $Var_P(IF_P(X))$ as a plug-in of σ^2 (under additional regularity conditions). This approach is also known as the *infinitesimal jackknife* (Efron 1981). A critical requirement for using this approach is the availability of analytical knowledge and computability of the relevant influence function or Hadamarad derivative, which unfortunately is not always possible especially for sophisticated models.

Resampling: Broadly speaking, this approach conducts inference by using the distribution of model outcomes evaluated at different data sets that are derived from the original data. Two main methods under this umbrella are the bootstrap and the jackknife. Let us overview each.

Bootstrap: The bootstrap hinges on the principle that the resample distribution is a close approximation for the sampling distribution of an original statistic. To explain, given a data set $\{X_1,\ldots,X_n\}$, a resample $\{X_1^*,\ldots,X_n^*\}$ is obtained by sampling with replacement n times from $\{X_1,\ldots,X_n\}$, or in other words each X_i^* is drawn from $\hat{P}_n$ independently conditional on $\hat{P}_n$. Denoting $P_n^*(\cdot) := (1/n)\sum_{i=1}^n I(X_i^* \in \cdot)$ as the resample empirical distribution, we have a resample estimate $\psi(P_n^*)$. Under similar assumptions as for the CLT (2), we have the bootstrap CLT

$$\sqrt{n}(\psi(P_n^*) - \psi(\hat{P}_n)) \Rightarrow N(0,\sigma^2) \tag{5}$$

conditional on the data $X_1, X_2, \ldots$ in probability as $n \to \infty$. Comparing with the limit theorem in (2), (5) signifies that replacing the role of the ground-truth $\psi(P)$ by the estimate $\psi(\hat{P}_n)$, and correspondingly replacing $\psi(\hat{P}_n)$ by the resample counterpart $\psi(P_n^*)$, give rise to the same limiting distribution. Note that the weak convergence in (5) is defined "conditionally" on the data. It more precisely means the random variable $\mathbb{P}(\sqrt{n}(\psi(P_n^*) - \psi(\hat{P}_n) \le x | \hat{P}_n) \xrightarrow{p} \mathbb{P}(N(0,\sigma^2) \le x)$ for any $x \in \mathbb{R}$ as $n \to \infty$, where $\mathbb{P}(\cdot|\hat{P}_n)$ denotes the probability with respect to the resampling randomness conditional on the data.

A key insight is that (2) and (5) have the same asymptotic limit $N(0,\sigma^2)$. Thus, if we can find the quantiles of the LHS variable in (5), then they can be used to approximate the counterparts in (2)

which then allows us to construct the CI of interest. The nice thing about (5) is that, indeed, we could compute the quantiles of the LHS because everything there is extractable from the data. In particular, we can simulate many resample estimates, say $\psi(P_n^{*b})$ for $b = 1,\ldots,B$, to approximate the quantiles of $\psi(P_n^*) - \psi(\hat{P}_n)$ (conditional on the data). More precisely, note that an exact-coverage $(1-\alpha)$-level CI of $\psi(P)$ is readily seen to be $[\psi(\hat{P}_n) - q_{1-\alpha/2}, \psi(\hat{P}_n) - q_{\alpha/2}]$, where $q_{\alpha/2}$ and $q_{1-\alpha/2}$ are the $\alpha/2$ and $(1-\alpha/2)$-th quantiles of $\psi(\hat{P}_n) - \psi(P)$. By the same limit of (2) and (5), we can approximate $q_{\alpha/2}$ and $q_{1-\alpha/2}$ with the $\alpha/2$ and $(1-\alpha/2)$-th quantiles of $\psi(P_n^*) - \psi(\hat{P}_n)$, which can be in turn estimated by the sample quantiles of $\psi(P_n^{*b}) - \psi(\hat{P}_n)$ that we denote as $\hat{q}^*_{\alpha/2}$ and $\hat{q}^*_{1-\alpha/2}$. To sum up, we output the interval $[\psi(\hat{P}_n) - \hat{q}^*_{1-\alpha/2}, \psi(\hat{P}_n) - \hat{q}^*_{\alpha/2}]$. This is called the *basic bootstrap* (Efron and Tibshirani 1994; Davison and Hinkley 1997).

We can also use the so-called *percentile bootstrap* interval $[\hat{q}^*_{\alpha/2}, \hat{q}^*_{1-\alpha/2}]$ as the CI, which has a simpler expression and is justified by the additional symmetry of the limiting normal distribution. Another option is the *studentized bootstrap* interval (Davison and Hinkley 1997 §2.4) that uses the quantiles of a studentized statistic $(\psi(P_n^*) - \psi(P_n))/\sigma_n^*$ where σ_n^* refers to the resample counterpart of $\hat{\sigma}_n$, the empirical estimate of σ, which could give a more accurate CI but at the expense of needing an iterated layer of resampling or analytical knowledge. Finally, we could also estimate σ^2/n in (4) directly via $Var^*(\psi(P_n^*))$, where $Var^*(\cdot)$ denotes the variance with respect to the resampling randomness conditional on data. This is called the *standard error bootstrap* (Efron 1981).

Jackknife: This uses a leave-one-out idea as follows. Call $\psi(P_n^{(-i)})$ the i-th leave-one-out estimate, where $P_n^{(-i)}$ denotes the empirical distribution excluding the i-th observation. In other words, $\psi(P_n^{(-i)})$ is the estimate constructed from the data set except the i-th observation. Call $J_i := n\psi(\hat{P}_n) - (n-1)\psi(P_n^{(-i)})$ the i-th jackknife sample. Then an asymptotically exact CI is given by $[\bar{J} - z_{1-\alpha/2}S_J/\sqrt{n}, \bar{J} + z_{1-\alpha/2}S_J/\sqrt{n}]$ where $\bar{J} = (1/n)\sum_{i=1}^n J_i$ and $S_J^2 = (1/(n-1))\sum_{i=1}^n (J_i - \bar{J})^2$ are the sample mean and variance of the jackknife samples. Intuitively, the jackknife samples act as finite-difference approximations for the influence function which has a Gateaux derivative interpretation, and hence S_J^2 approximates the empirical counterpart of $Var_P(IF_P(X))$ (Efron and Tibshirani 1994). On the other hand, $\bar{J}$ approximates $\psi(\hat{P}_n)$, in fact with the additional benefit of having a reduced bias. These together lead to the validity of the jackknife CI.

Why are the bootstraps and jackknife expensive? All the bootstrap methods described above (basic, percentile, studentized, standard error) and the jackknife requires re-evaluating the model for a large number of times. In the bootstraps, we need to use enough resamples B in order to estimate the resample quantiles or variance. In the jackknife, we need to use n re-evaluations because of its leave-one-out structure. In Example 1 for instance, this means constructing a CI for the steady-state average waiting time requires simulating a large number B of long-horizon trajectories in the bootstraps and n trajectories in the jackknife.

General challenges in delta method and resampling. To summarize, we either need analytical knowledge on the Hadamard derivative or influence function in the delta method, or the need to re-run the model many times in resampling methods. As discussed in the introduction, both of these can raise challenges for expensive black-box models encountered in the simulation input uncertainty problem.

2.2 Joint Aleatory and Epistemic Uncertainties

When computing ψ requires Monte Carlo and incurs stochastic uncertainty like in Example 2, we face the additional intricacy related to the joint impacts of Monte Carlo and data noises. Using the same setting as the beginning of Section 2, but now $\psi(Q)$ for a given Q can only be estimated by running unbiased simulation repetitions, i.e., we have access to the generation of noisy $\psi_r(Q)$ that is unbiased for $\psi(Q)$, and we use $\hat{\psi}_R(Q) = (1/R)\sum_{r=1}^R \psi_r(Q)$ as an estimate of $\psi(Q)$. Given i.i.d. data of size n, we again denote $\hat{P}_n$ as the empirical distribution. A natural point estimate for $\psi(P)$ is $\hat{\psi}_R(\hat{P}_n)$. Under mild assumptions,

including those on the aleatory noise (Glynn and Lam 2018), we have a CLT

$$\hat{\psi}_R(\hat{P}_n) - \psi(P) \overset{d}{\approx} N\left(0, \frac{\sigma^2}{n} + \frac{\tau^2}{R}\right) \tag{6}$$

when n and R are both large, and "$\overset{d}{\approx}$" denotes "approximately equal in distribution". The approximation (6) can be written more rigorously in terms of weak convergence, by letting n and R scaling linearly with a common scaling parameter, but (6) is easy to intuit. Essentially, it stipulates that the point estimate $\hat{\psi}_R(\hat{P}_n)$ now incurs two variance components, σ^2/n coming from the input uncertainty and τ^2/R coming from the Monte Carlo uncertainty. These two variances correspond to the two "hats" in $\hat{\psi}_R(\hat{P}_n)$. Lastly, we note that $\psi(Q)$ does not necessarily need to be the expectation of $\hat{\psi}_r(Q)$ and can be more general summary statistic, in which case $\hat{\psi}_R(Q)$ would be constructed accordingly.

In the following, we explain why the approaches presented in Section 2.1 face additional challenges in the presence of aleatory uncertainty. First, it is reasonable to consider the case where R is at least of the same order as n, so that the aleatory uncertainty is not the dominant error; otherwise, the input uncertainty component becomes negligible and we can simply estimate τ^2 by using the sample variance of all the R simulation runs, which essentially reduces the problem to standard mean estimation. Thus, from now on, we assume $R = \Omega(n)$. Correspondingly, unlike in the deterministic case, we regard a CI construction scheme to be computationally acceptable for aleatory-noise-corrupted problems if it requires only order n model evaluations.

Why is quantile-based bootstrap expensive? Let us consider the basic bootstrap for illustration. The bootstrap principle entails that

$$\hat{\psi}_R(P_n^*) - \psi(\hat{P}_n) \overset{d}{\approx} N\left(0, \frac{\sigma^2}{n} + \frac{\tau^2}{R}\right) \tag{7}$$

conditional on data $X_1, X_2, \ldots$, which again can be argued more rigorously via a conditional weak convergence in probability. Note that here: 1) The computation of $\hat{\psi}_R(P_n^*)$ typically requires newly generated simulation runs different from the ones used to compute $\hat{\psi}_R(\hat{P}_n)$ in the CLT (6); 2) The centering quantity in (7), $\psi(\hat{P}_n)$, is in fact not obtainable. Instead, we can use a finite R number of repetitions to approximate this $\psi(\hat{P}_n)$, which would then add extra variance into the RHS of (7). Moreover, suppose this R repetitions are the same ones used to compute the original point estimate $\hat{\psi}_R(\hat{P}_n)$ in (6), then the quantiles extracted from the resample estimates become dependent with the original estimate. In other words, the approximate equality in distribution, as well as the independence between the resample quantiles and the original estimate, which are essential to applying the bootstrap are both lost in this setting. If we compute the center $\psi(\hat{P}_n)$ in (7) with R simulation repetitions that are independent of those used to compute the original point estimate, or if we use $R = \omega(n)$ in these tasks, then the resample quantiles and original estimate become asymptotically independent, and we only face the issue of distributional mismatch between resample and original estimates which could be resolved by using suitable shrinkage techniques (Barton et al. 2018). Nonetheless, in any case, we need to resample B times for a sufficiently large B, and each resample estimate requires R simulation repetitions. This gives an overall computation demand of at least BR, which is of larger order than n.

Why is standard error bootstrap expensive? Another approach is to use the bootstrap input variance $Var^*(\psi(P_n^*))$ to estimate the input variance σ^2/n directly (note that the Monte Carlo component τ^2, like discussed before, can be estimated accurately by simply taking the sample variance of all simulation runs; Cheng and Holland 1997). However, note that we could only evaluate ψ noisily with a finite number R of simulation repetitions. In other words, the bootstrap input variance $Var^*(\psi(P_n^*))$ can be viewed as the variance of a conditional expectation where the latter can only be unbiasedly estimated. Tackling this requires nested simulation, where in the outer level we resample B times to obtain P_n^{*b}, $b = 1, \ldots, B$, then

given each resample, we generate R inner simulation runs $\psi_r(P_n^{*b})$, $r = 1, \ldots, R$ and take average to obtain $\hat{\psi}_R(P_n^{*b})$. Then, we can estimate $Var^*(\psi(P_n^*))$ via

$$\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\psi}_R(P_n^{*b}) - \overline{\hat{\psi}})^2 - \frac{1}{BR(R-1)}\sum_{b=1}^{B}\sum_{r=1}^{R}(\psi_r(P_n^{*b}) - \hat{\psi}_R(P_n^{*b}))^2 \tag{8}$$

where $\overline{\hat{\psi}} = (1/B)\sum_{b=1}^{B}\hat{\psi}_R(P_n^{*b})$ (e.g., Cheng and Holland 1997; Song and Nelson 2015; Lin et al. 2015). To understand (8), note that a standard conditioning argument gives $Var^*(\hat{\psi}_R(\hat{P}_n^*)) = Var^*(\psi(\hat{P}_n^*)) + E[Var^*(\hat{\psi}_R(\hat{P}_n^*)|\hat{P}_n^*)]$, a quantity that is estimated unbiasedly by the first term in (8). The second term in (8) is a bias correction that removes $E[Var^*(\hat{\psi}_R(\hat{P}_n^*)|\hat{P}_n^*)]$ unbiasedly, so that (8) becomes an unbiased estimator for $Var^*(\psi(P_n^*))$. Connecting to analysis-of-variance, these can be viewed as estimating the between-group variance in a random-effect model.

It turns out that the variance of the estimator (8) is of order (Sun et al. 2011; Lam and Qian 2022)

$$O\left(\frac{1}{Bn^2} + \frac{1}{BR^2}\right) \tag{9}$$

Now, note that a reasonable estimator of the input variance σ^2/n, which is of order n, should have an error $o(1/n)$ since only so would make the error relatively negligible. This translates to the requirement of the variance of (8) to be $o(1/n^2)$. Looking at the first term in (9), this implies that B needs to be $\omega(1)$. With this, we argue that the overall computation need, BR, cannot be $O(n)$ because, if $BR = O(n)$, then $R = o(n)$ and this deems the denominator of the second term in (9) to be $o(n^2)$ so that (9) is $\omega(1/n^2)$. Lam and Qian (2022) calls this $\omega(n)$ overall computation requirement to be a "complexity barrier" in using the standard error bootstrap in the presence of aleatory uncertainty.

Why is delta method expensive? The delta method or infinitesimal jackknife to estimate the input variance σ^2/n runs into the same difficulty as in the deterministic case: It requires analytical knowledge and computability of the Hadamard derivative or the influence function. Alternately, we can consider using finite-difference or zeroth-order gradient estimators to approximate these derivative quantities, but such approaches are known to have slow convergence (Zazanis and Suri 1993; Fox and Glynn 1989) and, moreover, the infinite-dimensional nature of these derivatives (taken with respect to distributions) adds sophistication to the estimation accuracy. More concretely, suppose σ^2 can be expressed as $Var_P(IF_P(X))$ where the influence function IF_P satisfies (3). Then we can use $\frac{1}{n}\sum_{i=1}^{n} IF_{\hat{P}_n}(X_i)^2$ as an estimate of $Var_P(IF_P(X))$ (note that $IF_Q(X)$ can be assumed without loss of generality to have mean zero under Q because otherwise we can replace it by $IF_Q(X) - E_Q[IF_Q(X)]$ which does not affect (3)). Using the Gateaux derivative interpretation, we can estimate $IF_{\hat{P}_n}(X_i)$ via $(\psi_r((1-\varepsilon)\hat{P}_n + \varepsilon\delta_{X_i}) - \psi_r(\hat{P}_n))/\varepsilon$ for some small ε that scales in terms of n, where δ_{X_i} denotes the delta mass at X_i. Like in the standard error bootstrap, a meaningful estimation of σ^2/n, which is of order $1/n$, should have an error $o(1/n)$. This turns out to be achievable by using an order n overall computation budget via choosing $\omega(1/n^{1/4}) \leq \varepsilon \leq o(1)$ (Lam and Qian 2019). This is better than the quantile-based and standard error bootstraps, but it requires careful procedural tuning that could affect practical performances.

3 LOW-COMPUTATION UNCERTAINTY QUANTIFICATION

In view of the challenges faced by common statistical methods, we present several methodologies to construct CIs with lower computation and without auxiliary analytical information. Like in Section 2, we will divide into the deterministic case and the stochastic case. We will focus on the former in this section and discuss two inter-related approaches, data-splitting and the cheap bootstrap, followed by some further aspects and comparisons among these approaches.

3.1 Data-Splitting

This approach divides data into groups and computes individual estimates from different groups. To construct a CI, the key idea is to judiciously create pivotal statistic from these individual estimates that cancels out the nuisance parameter, typically the asymptotic estimation variance (Glynn and Iglehart 1990; Schmeiser 1982; Schruben 1983). For this reason this approach bears the umbrella name of *cancellation methods* and encompasses batching methods and more general schemes such as standardized time series. Historically, this approach is motivated from simulation output analysis involving serially dependent data such as steady-state estimation (Steiger and Wilson 2002; Steiger et al. 2005; Muñoz and Glynn 1997) and Markov chain Monte Carlo in Bayesian statistics (Geyer 1992; Flegal and Jones 2010; Jones et al. 2006). Despite its historical motivation as a powerful tool to handle dependent data, such data-splitting methodologies can be made computationally light by using a small number of data divisions and subsequently model evaluations.

Batching. We divide the data into B batches, called $(X_1,\dots,X_m),(X_{m+1},\dots,X_{2m}),\dots,(X_{(B-1)m+1},\dots,X_n)$ where the data size per batch is $m = n/B$ (for simplicity, assume n is divisible by B). For each batch b, we construct the batch estimate $\psi(\hat{P}_m^{(b)})$ where $\hat{P}_m^{(b)}(\cdot) = (1/m)\sum_{i=(b-1)m+1}^{bm} I(X_i \in \cdot)$ is the empirical distribution of the batch b data. Then we output

$$\left[\bar{\psi} - t_{B-1,1-\alpha/2}\frac{S_{bat}}{\sqrt{B}}, \bar{\psi} + t_{B-1,1-\alpha/2}\frac{S_{bat}}{\sqrt{B}}\right] \tag{10}$$

where $\bar{\psi} = \frac{1}{B}\sum_{b=1}^{B}\psi(\hat{P}_m^{(b)})$ is the sample mean and $S_{bat}^2 = \frac{1}{B-1}\sum_{b=1}^{B}(\psi(\hat{P}_m^{(b)}) - \bar{\psi})^2$ is the sample variance of $\psi(\hat{P}_m^{(b)})$'s. The critical value $t_{B-1,1-\alpha/2}$ is the $(1-\alpha/2)$-th quantile of t_{B-1}, the t-distribution with degree of freedom $B-1$.

Batching gives rise to asymptotically exact CIs as $n \to \infty$, for any fixed $B \geq 2$. In other words, it uses a number of model evaluations as small as 2 in total, and in this sense it is computationally light. Its validity is based on the independent CLT for each batch $\sqrt{m}(\psi(\hat{P}_m^{(b)}) - \psi(P)) \Rightarrow N(0,\sigma^2)$ so that the pivotal statistic

$$\frac{\bar{\psi} - \psi(P)}{S_{bat}/\sqrt{B}} \Rightarrow \frac{\bar{Z}}{S/\sqrt{B}} \tag{11}$$

as $n \to \infty$ via the continuous mapping theorem, where $\bar{Z}$ and S^2 are respectively the sample mean and variance of B i.i.d. $N(0,\sigma^2)$. The RHS of (11) can be seen to follow a t_{B-1}-distribution by using the elementary properties of normal and χ^2-distributions, which gives rise to the asymptotically exact CI (10).

Sectioning. This is a variant of batching that uses

$$\left[\psi(\hat{P}_n) - t_{B-1,1-\alpha/2}\frac{S_{sec}}{\sqrt{B}}, \psi(\hat{P}_n) + t_{B-1,1-\alpha/2}\frac{S_{sec}}{\sqrt{B}}\right] \tag{12}$$

where $S_{sec}^2 = \frac{1}{B-1}\sum_{b=1}^{B}(\psi(\hat{P}_m^{(b)}) - \psi(\hat{P}_n))^2$. Compared to batching, sectioning uses the grand estimate $\psi(\hat{P}_n)$ constructed from the entire data set as the center of the interval and in the formula of S_{sec}^2. The asymptotic exactness of (12), for a fixed $B \geq 2$, follows in the same way as that of (10), by noting that $\psi(\hat{P}_n)$ and $\bar{\psi}$ are equal up to a negligible $o_p(1/\sqrt{n})$ error. Compared to batching, sectioning uses $B+1$ model evaluations because it requires the grand estimate $\psi(\hat{P}_n)$, giving rise to a minimum of 3 model evaluations.

Batching-Sectioning Hybrid. We can combine batching and sectioning together by using

$$\left[\psi(\hat{P}_n) - t_{B-1,1-\alpha/2}\frac{S_{bat}}{\sqrt{B}}, \psi(\hat{P}_n) + t_{B-1,1-\alpha/2}\frac{S_{bat}}{\sqrt{B}}\right]$$

i.e., the interval center uses that of sectioning while the half-width uses batching (Nakayama 2014). Its validity justification is similar to the above and, like sectioning, it uses $B+1$ and hence a minimum of 3 model evaluations.

Batched Jackknife. We can combine batching with the jackknife to substantially reduce the computation demand of the standard jackknife from n to the number of batches. The main idea is to use leave-one-batch-out instead of just leave-one-observation-out (Asmussen and Glynn 2007 §III.5b). More precisely, like in batching, we divide the data into B batches. To construct the b-th leave-one-batch-out estimate, we use all the data except the b-th batch, namely $X_1,\dots,X_{(b-1)m},X_{bm+1},\dots,X_n$, to evaluate $\psi(\hat{P}_{n-m}^{(-b)})$ where $\hat{P}_{n-m}^{(-b)}$ is the empirical distribution of $X_1,\dots,X_{(b-1)m},X_{bm+1},\dots,X_n$. Together with the grand estimate $\psi(\hat{P}_n)$, we define the b-th batched jackknife sample as $J_b = B\psi(\hat{P}_n) - (B-1)\psi(\hat{P}_{n-m}^{(-b)})$. Then we output

$$\left[\bar{J} - t_{B-1,1-\alpha/2}\frac{S_{BJ}}{\sqrt{B}}, \bar{J} + t_{B-1,1-\alpha/2}\frac{S_{BJ}}{\sqrt{B}}\right] \tag{13}$$

where $\bar{J} = \frac{1}{B}\sum_{b=1}^{B} J_b$ and $S_{BJ}^2 = \frac{1}{B-1}\sum_{b=1}^{B}(J_b-\bar{J})^2$ are the sample mean and variance of the batched jackknife samples. The asymptotic exactness of (13) holds under the same assumption as batching, namely the independent CLT per batch. Computationally, it has a similar cost as sectioning and batching-sectioning-hybrid in that it uses $B+1$ model evaluations in total.

Overlapping Batching. This approach allows the data in different batches to overlap (Alexopoulos et al. 2007; Meketon and Schmeiser 1984; Song and Schmeiser 1995; Su et al. 2018). Thus, unlike batching, it does not require $Bm = n$. Denote the starting points of (overlapping) batches as $1, d+1, 2d+1, \dots, (B-1)d+1$, so that $(B-1)d+m = n$. We construct the batch estimate $\psi(\hat{P}_m^{(b)})$ where $\hat{P}_m^{(b)}$ is the empirical distribution of $X_{(b-1)d+1},\dots,X_{(b-1)d+m}$. Following Su et al. (2018) to denote $\beta = m/n$, we output

$$\left[\bar{\psi} - q_{1-\alpha/2}^{OB}(\beta,B)\frac{S_{OB}}{\sqrt{n}}, \bar{\psi} + q_{1-\alpha/2}^{OB}(\beta,B)\frac{S_{OB}}{\sqrt{n}}\right] \tag{14}$$

where $S_{OB}^2 = \frac{m}{B}\sum_{b=1}^{B}(\psi(\hat{P}_m^{(b)}) - \bar{\psi})^2$, $\bar{\psi} = \frac{1}{B}\sum_{b=1}^{B}\psi(\hat{P}_m^{(b)})$ and $q_{1-\alpha/2}^{OB}(\beta,B)$ is the $(1-\alpha/2)$-th quantile of the distribution of $(\sum_{b=1}^{B}\tilde{W}_b)/\sqrt{\beta B\sum_{b=1}^{B}(\tilde{W}_b - (1/B)\sum_{j=1}^{B}\tilde{W}_j)^2}$ where $W_b = W((b-1)(1-\beta)/(B-1)+\beta) - W((b-1)(1-\beta)/(B-1))$, and $\{W(x)\}_{x\in[0,1]}$ is the standard Brownian motion on $[0,1]$. This is the so-called OB-I approach in Su et al. (2018), who also present other variants such as sectioning and more general standardized time series. The asymptotic exactness of (14) and these variants follow a bit more intricately than the previous data-splitting methods, but the general idea is similar in that they all utilize pivotal statistics similar to (11) that exhibit tractable limiting distributions from which one can invert to obtain valid CIs.

3.2 Cheap Bootstrap

This is a recent approach to conduct computationally light inference based on resampling (Lam 2022b; Lam and Liu 2023; Lam 2022a). Unlike the standard bootstrap principle that utilizes the resemblance of the resample distribution to the sampling distribution (i.e., (5)), the cheap bootstrap uses sample-resample independence and the direct exploitation of asymptotic normality. Suppose we compute the point estimate $\psi(\hat{P}_n)$ and resample estimates $\psi(P_n^{*b})$, $b = 1,\dots,B$. The cheap bootstrap interval is given by

$$\left[\psi(\hat{P}_n) - t_{B,1-\alpha/2}S_{CB},\ \psi(\hat{P}_n) + t_{B,1-\alpha/2}S_{CB}\right] \tag{15}$$

where $S_{CB}^2 = (1/B)\sum_{b=1}^{B}\left(\psi(P_n^{*b}) - \psi(\hat{P}_n)\right)^2$. The interval (15) appears similar to the standard error bootstrap when B is large, as S_{CB}^2 corresponds to the sample variance computed from the resample estimates. However,

B can be as small as 1 in the construction. Note that the degree of freedom in the t-distribution is B (instead of the usual $B-1$ in typical t-interval encountered in textbook statistics and the batching-type methods discussed earlier).

The cheap bootstrap interval (15) is asymptotically exact for any fixed $B \geq 1$, under the same standard assumptions as the conventional bootstraps, namely that (2) and (5) hold. Arguing its validity requires a twist to the standard bootstrap principle, by looking at the joint distribution of resample and original estimates via combining (2) and (5). Intuitively, the universality of the limit in (5) regardless of the data realization implies the asymptotic independence of any resample from the original sample as well as among the resamples. In other words, we have

$$\sqrt{n}(\psi(\hat{P}_n) - \psi(P), \psi(P_n^{*1}) - \psi(\hat{P}_n), \ldots, \psi(P_n^{*B}) - \psi(\hat{P}_n)) \Rightarrow (Z_0, Z_1, \ldots, Z_B)$$

where $Z_b, b = 0, \ldots, B$ are i.i.d. $N(0, \sigma^2)$ variables. From this, we show that the pivotal statistic $(\psi(\hat{P}_n) - \psi(P))/S_{CB}$ converges to a t_B-distribution from which we can invert to obtain a valid CI. Computationally, the cheap bootstrap uses $B+1$ model evaluations, which include B resample estimates and one original estimate.

3.3 Further Aspects

While the presented methods above arguably alleviate the issue of expensive model evaluation, our criterion on their success has been confined to large-sample coverage validity, i.e., asymptotic exactness. Evidently, another important criterion is the interval width, where a shorter and less variable interval is preferable. Moreover, with a finite data size n, the choice of B could affect the practical coverage performance, both because the smallness of B should be viewed relative to n, and also because of the impacts of higher-order coverage errors. In this subsection we will provide understanding on the half-width and higher-order coverage error, and delegate other considerations and comparisons to the next subsection.

Half-Width. Let us first dissect the half-width behaviors of most of the presented low-computation methods. The data-splitting methods, with the exception of overlapping batching, all have half-widths in the form $t_{B-1,1-\alpha/2}S/\sqrt{B}$ for some variance estimate S^2. Delving into the limits of the corresponding pivotal statistics would reveal that this half-width is approximately $t_{B-1,1-\alpha/2}(\sigma/\sqrt{n})\sqrt{\chi^2_{B-1}/(B-1)}$ where χ^2_{B-1} stands for a χ^2 variable with degree of freedom $B-1$. Mechanical calculation would show that the expectation and the variance of the half-width both naturally decrease monotonically in B. More importantly, the decrease rate is very fast when B deviates slightly away from 2 and then levels off gradually towards the level of the normality CI (i.e., $z_{1-\alpha/2}\sigma/\sqrt{n}$), which is viewed as the idealized interval with either analytical knowledge of σ or infinite data and batches. For example, when $B = 2$, the expected half-width has a 417% inflation relative to the idealized interval, but when $B = 3, 4, 6$, the inflations become $95\%, 50\%, 25\%$ respectively. That is, a batching interval using 6 batches gives a roughly 25% wider interval than the idealized normality CI. Cheap bootstrap interval follows a similar behavior as above in terms of half-width, except that $B-1$ is now replaced by B which represents the number of resamples. On the other hand, overlapping batching has a quite distinct half-width behavior because of its different limiting distribution of its pivotal statistic.

Higher-Order Coverage Error. The coverage error of an asymptotically exact $(1-\alpha)$-level CI, say $\mathscr{I}$, refers to the remainder term in $\mathbb{P}(\psi(P) \in \mathscr{I}) = 1 - \alpha + \text{remainder}$. It is known that the most basic conventional bootstrap approaches (using infinite or a large enough number of resamples B), including the basic and percentile bootstraps, exhibit a coverage error of order $O(1/n)$ if $\mathscr{I}$ is two-sided (like the ones focused on in this tutorial) and $O(1/\sqrt{n})$ if $\mathscr{I}$ is one-sided (i.e., one of the limits of the interval is $\pm\infty$). These error orders can be improved via studentization or calibration (Hall 2013). Obtaining these error terms requires careful analyses of Edgeworth expansions, under regularity conditions including, most importantly, the restriction to the so-called smooth function model instead of more complicated functionals of ψ for technicality reasons (Hall 2013).

The data-splitting methods and cheap bootstrap exhibit similar coverage error behaviors as the basic conventional bootstraps in that, under regularity conditions similar to the analyses of conventional bootstraps, they have two-sided coverage error $O(1/n)$ and one-sided error $O(1/\sqrt{n})$ (He and Lam 2021b; He and Lam 2021a; Lam 2022b). Nonetheless, the analyses of these schemes arguably go beyond the classical analyses because these schemes have a t-limit instead of normal limit. To bypass this, we need to suitably aggregate the Edgeworth expansions of individual batch or resample estimates. This results in expansion coefficients that are less transparent than in conventional bootstraps but, at least for the first-order terms, they can be expressed as expectations of normal variables amenable to Monte Carlo approximations (He and Lam 2021b; Lam 2022b).

Another perspective in understanding the coverage error is through finite-sample bounds, especially powerful in shedding insights for high-dimensional problems by capturing the dimensionality effects. These bounds have been actively studied in the high-dimensional CLT and Berry-Esseen literature (Lopes 2022; Chernozhukov et al. 2020). For the cheap bootstrap, an error order of $O(1/\sqrt{n})$ has been shown to hold when the problem dimension (in a smooth function model) grows arbitrarily close to n for any fixed $B \geq 1$ (Lam and Liu 2023).

3.4 Comparisons of Low-Computation Uncertainty Quantification Methods

Given all the methods and discussions so far, it may now be natural to ask: *Which of the low-computation method is the best and, furthermore, are there even better methods than what we have presented?* There may not be a simple answer as it imaginably involves multifaceted considerations. Nonetheless, in the following, we will attempt to discuss this question from two angles.

Asymptotic Statistical Optimality. Suppose we focus on asymptotic performances as data size $n \to \infty$, but with a fixed B model evaluation budget. In this regime, all the discussed methods exhibit asymptotic exactness (as discussed in Sections 3.1 and 3.2) and, except overlapping batching, all of them have the same half-width behavior characterized by the χ^2-distribution (as discussed in Section 3.3). Thus, in terms of the two main CI criteria of coverage and half-width, the low-computation methods appear largely the same asymptotically. Previous works that can make distinctive comparisons (based on criteria such as the mean squared error or the variance of the variance estimator; e.g., Schmeiser 1982; Song and Schmeiser 1995; Flegal and Jones 2010; Meketon and Schmeiser 1984) allow varying computation costs and thus are not directly relevant to this considered setting.

The universality of the asymptotic exactness and half-width behavior hints that most of these methods are already optimal in a certain sense. One way to formalize this is to use the notion of asymptotically uniformly most accurate unbiased CIs recently suggested in He and Lam (2023). It considers the class of CIs that use B model evaluations, where each model evaluation $\psi(\cdot)$ can be made on any subset of the size-n data. Then a CI $\mathscr{I}$ is *asymptotically unbiased at level* $1-\alpha$ if $\lim_{n\to\infty}\mathbb{P}(\psi(P) \in \mathscr{I}) \geq 1-\alpha$ and for any $\delta \neq 0$, $\limsup_{n\to\infty}\mathbb{P}(\psi(P)+n^{-1/2}\delta \in \mathscr{I}) \leq 1-\alpha$. Correspondingly, a CI $\mathscr{I}$ is *asymptotically uniformly most accurate unbiased at level* $1-\alpha$ if it minimizes $\limsup_{n\to\infty}\mathbb{P}(\psi(P)+n^{-1/2}\delta \in \mathscr{I}')$ among all asymptotically unbiased level $1-\alpha$ CIs $\mathscr{I}'$ for any $\delta \neq 0$. This definition is a large-sample counterpart of the uniformly most accurate unbiased CI (Lehmann et al. 2005) arising as the dual of a similar notion in hypothesis tests. Under this framework, standard batching, batched jackknife and the cheap bootstrap are statistically optimal. Furthermore, batching using arbitrary non-uniform batch sizes and overlapping batching can also achieve optimality by suitably aggregating the batch estimates. Intriguingly, the resulting optimal overlapping batching bears a different form than previously suggested and uses a t-critical value.

Robustness to B. In practice, the smallness of B should be viewed relative to n, and a preferable method ideally performs well regardless of B. In this sense, batching, sectioning and batching-sectioning hybrid suffers from a constrained tradeoff between the number of batches and the sample size per batch (i.e., $Bm = n$). For a given data set, these methods could have deteriorated performances as B increases since the sample size per batch would become too small and impact the validity of asymptotic normality.

On the other hand, batched jackknife, overlapping batching and the cheap bootstrap face less challenges from this constrained tradeoff and is more robust to the choice of B: As B increases, their batch or resample sizes remain sufficiently large. For instance, the resample size in the cheap bootstrap is unchanged in terms of B. In other words, these methods perform well both when B is small or big. Nonetheless, note that when B is big, the behavior of the batched jackknife reduces to that of standard jackknife, which implies that it cannot be used for quantile estimation or related problems (Ghosh et al. 1984; Martin 1990). In contrast, the cheap bootstrap and overlapping batching continue to apply in these problems.

4 UNCERTAINTY QUANTIFICATION WITH ALEATORY UNCERTAINTY

We now present low-computation uncertainty quantification methodologies for problems with both aleatory and epistemic uncertainties. Like Section 3, we consider data-splitting and the cheap bootstrap in Sections 4.1 and 4.2 and, furthermore, we introduce subsampling as an additional strategy in Section 4.3.

4.1 Data-Splitting

We focus on standard batching for illustration. The main idea is similar to the deterministic model case where we divide data into batches and create a suitable pivotal statistic that cancels out the nusiance parameter. However, now the nuisance parameter consists of not only the input variance but also the Monte Carlo variance. More precisely, like in Section 3.1, we divide the data into B batches, say $(X_1,\ldots,X_m),(X_{m+1},\ldots,X_{2m}),\ldots,(X_{(B-1)m+1},\ldots,X_n)$ where the data size per batch is $m=n/B$ (again for simplicity, assume n is divisible by B). For each batch b, we evaluate the batch estimate $\hat{\psi}_R(\hat{P}_m^{(b)})$ where $\hat{P}_m^{(b)}(\cdot)=(1/m)\sum_{i=(b-1)m+1}^{bm} I(X_i\in\cdot)$ is the empirical distribution of the batch b data, by using R unbiased simulation repetitions $\psi_r(\hat{P}_m^{(b)})$ and taking average, i.e., $\hat{\psi}_R(\hat{P}_m^{(b)})=(1/R)\sum_{r=1}^{R}\psi_r(\hat{P}_m^{(b)})$. Then we output

$$\left[\bar{\hat{\psi}}-t_{B-1,1-\alpha/2}\frac{S_{bat}}{\sqrt{B}},\bar{\hat{\psi}}+t_{B-1,1-\alpha/2}\frac{S_{bat}}{\sqrt{B}}\right] \tag{16}$$

where $\bar{\hat{\psi}}=\frac{1}{B}\sum_{b=1}^{B}\hat{\psi}_R(\hat{P}_m^{(b)})$ is the sample mean and $S_{bat}^2=\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\psi}_R(\hat{P}_m^{(b)})-\bar{\hat{\psi}})^2$ is the sample variance of $\hat{\psi}_R(\hat{P}_m^{(b)})$'s.

The validity of (16) relies on the independent CLT for each batch that accounts for both input and Monte Carlo noises, namely $\hat{\psi}_R(\hat{P}_m^{(b)})-\psi(P)\overset{d}{\approx}N(0,\sigma^2/m+\tau^2/R)$ so that the pivotal statistic similar to (11) holds and leads to the asymptotic exactness of (16). In this approach, the computation effort is BR, where $B\geq 2$ and R is a sufficiently large number that can invoke asymptotic normality. Compared to the conventional bootstraps described in Section 2.2 that requires $\omega(n)$ computation, the effort here scales linearly with n (if R is linear in n which, as we described in Section 2.2, is assumed because otherwise the problem reduces to a simple mean estimation). Other variants including sectioning and hybrid can be similarly reasoned. However, the batched jackknife and overlapping batching require more intricate additional configurations in order to properly cancel out the Monte Carlo variance.

4.2 Cheap Bootstrap

The cheap bootstrap can also be applied to account for both input and Monte Carlo uncertainties. Like before, the main idea is to suitably aggregate the resample estimates $\hat{\psi}_R(P_n^{*b})$, $b=1,\ldots,B$ and the original estimate $\hat{\psi}_R(\hat{P}_n)$ to construct a pivotal statistic. However, in this case the sample-resample independence no longer holds because of the Monte Carlo noise that impacts the same quantity $\hat{\psi}_R(\hat{P}_n)$ in both (6) and (7). To resolve this issue, one idea is to use a variance estimator centered at the average of resample estimates $\hat{\psi}_R(P_n^{*b})$'s instead of just $\hat{\psi}_R(\hat{P}_n)$ as in (15). This cancels out the Monte Carlo impacts, but requires using a number of resamples $B\geq 2$. Another approach, which reduces B to as low as 1, is to use a variance estimator that resembles that in (15), namely $S_{CB}^2=(1/B)\sum_{b=1}^{B}(\hat{\psi}_R(P_n^{*b})-\hat{\psi}_R(\hat{P}_n))^2$. We

define $q_{CB,1-\alpha/2} = \min\left\{q : \min_{\theta\geq 0} F(q;\theta) \geq 1-\frac{\alpha}{2}\right\}$ where $F(q;\theta)$ is the distribution function of $(\theta V_1 + V_2)/\sqrt{\frac{\theta^2+1}{B}\left(Y+V_3^2\right) - 2\sqrt{\frac{\theta^2+1}{B}}V_3V_2 + V_2^2}$ with $V_1, V_2, V_3 \overset{i.i.d.}{\sim} N(0,1)$ and $Y \sim \chi^2_{B-1}$ being all independent (and we set $Y=0$ when $B=1$). Then we output

$$\left[\hat{\psi}_R(\hat{P}_n) - q_{CB,1-\alpha/2}S_{CB},\ \hat{\psi}_R(\hat{P}_n) + q_{CB,1-\alpha/2}S_{CB}\right] \tag{17}$$

Note that $q_{CB,1-\alpha/2}$ can be calibrated by running Monte Carlo on rudimentary distributions including normal and χ^2. The interval (17) is asymptotically valid for any fixed $B \geq 1$ in the sense that the asymptotic coverage is at least $1-\alpha$ (but not necessarily exactly). The reason of asymptotic validity instead of exactness, as well as the appearance of $q_{CB,1-\alpha/2}$ instead of the usual t-critical value, is because of the joint approximation $\left(\hat{\psi}_R(\hat{P}_n) - \psi(P),\ \hat{\psi}_R(P_n^{*1}) - \hat{\psi}_R(\hat{P}_n), \ldots,\ \hat{\psi}_R(P_n^{*B}) - \hat{\psi}_R(\hat{P}_n)\right) \overset{d}{\approx} \left(\frac{\sigma}{\sqrt{n}}Z_0 + \frac{\tau}{\sqrt{R}}W_0,\ \frac{\sigma}{\sqrt{n}}Z_1 + \frac{\tau}{\sqrt{R}}W_1 - \frac{\tau}{\sqrt{R}}W_0, \ldots,\ \frac{\sigma}{\sqrt{n}}Z_B + \frac{\tau}{\sqrt{R}}W_B - \frac{\tau}{\sqrt{R}}W_0\right)$ where $Z_0, Z_1, \ldots, Z_B, W_0, W_1, \ldots, W_B \sim N(0,1)$ signify all the independent input uncertainties and Monte Carlo uncertainties in the original and resample estimates, which is more intricate than the asymptotic independence among $\psi(P_n^{*b})$'s and $\psi(\hat{P}_n)$ in the deterministic model case.

4.3 Subsampling

This approach aims to estimate the input variance σ^2/n directly, and thus has the advantage over data-splitting and the cheap bootstrap that it gives tighter CIs. It uses the standard error bootstrap idea, but instead of a full-size resample, it uses a subsample, i.e., a resample size s that is smaller than n, to create subsample empirical distribution P_s^* and ultimately rescale the variance estimate by the subsample ratio $\theta = s/n$ (Lam and Qian 2022). More precisely, this approach estimates σ^2/n via $\theta Var_*(\psi(P_s^*))$, which can be approximated by $\theta\left(\frac{1}{B-1}\sum_{b=1}^B(\hat{\psi}_R(P_s^{*b}) - \overline{\hat{\psi}})^2 - \frac{1}{BR(R-1)}\sum_{b=1}^B\sum_{r=1}^R(\psi_r(P_s^{*b}) - \hat{\psi}_R(P_s^{*b}))^2\right)$ where $\overline{\hat{\psi}} = (1/B)\sum_{b=1}^B \hat{\psi}_R(P_s^{*b})$, which is formula (8) with n replaced by s and scaled by θ.

The overall computation effort is BR. Unlike the standard error bootstrap discussed in Section 2.2 that requires a total effort $\omega(n)$, subsampling allows us to choose B to be growing arbitrarily in n and R to be of order θn, but θ can now be arbitrarily small such that R is still growing with n. In other words, the total effort can now be made independent of n, which is a substantial improvement over the naive standard error bootstrap and in a sense resolves the complexity barrier in Section 2.2.

Note that while subsampling (where we take a broad meaning here to refer to any schemes using a smaller resample size than the original data size, and include more specific methods like the m-out-of-n bootstrap; Bickel et al. 1997) can reduce computation effort for problems where $\psi(\hat{P}_n)$ is easier to compute for smaller n, our rationale in using subsampling here is different in that it aims to reduce the multiplicative computation effort in a nested procedure to handle the aleatory uncertainty. To explain how it works and how the resolution of the complexity barrier above is attained, note that, intuitively, the reason why we need $\omega(n)$ computation is because of the small magnitude of the input variance σ^2/n itself, so that we need a large number of inner simulation repetitions R to satisfactorily wash away the Monte Carlo noises. Thus, if we can use a smaller data size s than n, then we can correspondingly drive down the inner simulation repetitions and, because of the reciprocal relation of σ^2/n with n, we can scale back our estimate from smaller s to the original n. To summarize mathematically, from the bootstrap approximation $Var_*(\psi(P_n^*)) \approx \sigma^2/n$, we see that $Var_*(\psi(P_s^*)) \approx \sigma^2/s$ and thus $\theta Var_*(\psi(P_s^*)) \approx \sigma^2/n$ which implies $\theta Var_*(\psi(P_s^*))$ is a valid estimator of σ^2/n and with a small s, $\theta Var_*(\psi(P_s^*))$ can be estimated well with lighter computation effort.

ACKNOWLEDGMENTS

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

REFERENCES

Alexopoulos, C., N. T. Argon, D. Goldsman, G. Tokol, and J. R. Wilson. 2007. "Overlapping Variance Estimators for Simulation". *Operations Research* 55(6):1090–1103.

Asmussen, S., and P. W. Glynn. 2007. *Stochastic Simulation: Algorithms and Analysis*, Volume 57. Springer Science & Business Media.

Barton, R. R. 2012. "Tutorial: Input Uncertainty in Output Analysis". In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. Uhrmacher, 1–12. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Barton, R. R., S. E. Chick, R. C. Cheng, S. G. Henderson, A. M. Law, B. W. Schmeiser, L. M. Leemis, L. W. Schruben, and J. R. Wilson. 2002. "Panel Discussion on Current Issues in Input Modeling". In *Proceedings of the 2002 Winter Simulation Conference*, edited by E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, 353–369. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Barton, R. R., H. Lam, and E. Song. 2018. "Revisiting Direct Bootstrap Resampling for Input Model Uncertainty". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 1635–1645. Institute of Electrical and Electronics Engineers, Inc.

Barton, R. R., H. Lam, and E. Song. 2022. "Input Uncertainty in Stochastic Simulation". In *The Palgrave Handbook of Operations Research*, 573–620. Springer.

Bickel, P. J., F. Götze, and W. R. van Zwet. 1997. "Resampling Fewer than n Observations: Gains, Losses, and Remedies for Losses". *Statistica Sinica* 7:1–31.

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

Chernozhukov, V., D. Chetverikov, and Y. Koike. 2020. "Nearly Optimal Central Limit Theorem and Bootstrap Approximations in High Dimensions". *arXiv preprint arXiv:2012.09513*.

Chick, S. E. 2006. "Bayesian Ideas and Discrete Event Simulation: Why, What and How". In *Proceedings of the 2006 Winter Simulation Conference*, edited by L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, 96–106. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Corlu, C. G., A. Akcay, and W. Xie. 2020. "Stochastic Simulation under Input Uncertainty: A Review". *Operations Research Perspectives* 7:100162.

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and Their Application*. Number 1. Cambridge University Press.

Efron, B. 1981. "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods". *Biometrika* 68(3):589–599.

Efron, B., and R. J. Tibshirani. 1994. *An Introduction to the Bootstrap*. CRC Press.

Flegal, J. M., and G. L. Jones. 2010. "Batch Means and Spectral Variance Estimators in Markov Chain Monte Carlo". *The Annals of Statistics* 38(2):1034–1070.

Fox, B. L., and P. W. Glynn. 1989. "Replication Schemes for Limiting Expectations". *Probability in the Engineering and Informational Sciences* 3(3):299–318.

Geyer, C. J. 1992. "Practical Markov Chain Monte Carlo". *Statistical Science* 7(4):473–483.

Ghosh, M., W. C. Parr, K. Singh, and G. J. Babu. 1984. "A Note on Bootstrapping the Sample Median". *The Annals of Statistics* 12(3):1130–1135.

Glynn, P. W., and D. L. Iglehart. 1990. "Simulation Output Analysis using Standardized Time Series". *Mathematics of Operations Research* 15(1):1–16.

Glynn, P. W., and H. Lam. 2018. "Constructing Simulation Output Intervals under Input Uncertainty via Data Sectioning". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 1551–1562. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Hall, P. 2013. *The Bootstrap and Edgeworth Expansion*. Springer Science & Business Media.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. 2011. *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons.

He, S., and H. Lam. 2021a. "Higher-Order Coverage Error Analysis for Batching and Sectioning". In *Proceedings of the 2021 Winter Simulation Conference*, edited by S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper, 1–12. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

He, S., and H. Lam. 2021b. "Higher-Order Coverage Errors of Batching Methods via Edgeworth Expansions on t-Statistics". *arXiv preprint arXiv:2111.06859*.

He, S., and H. Lam. 2023. "Optimal Batching under Computation Budget". In *Proceedings of the 2023 Winter Simulation Conference*, edited by C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, 1–12. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Henderson, S. G. 2003. "Input Modeling: Input Model Uncertainty: Why Do We Care and What Should We Do about It?". In *Proceedings of the 2003 Winter Simulation Conference*, edited by S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, 90–100. Piscataway, New Jersey: IEEE.

Jones, G. L., M. Haran, B. S. Caffo, and R. Neath. 2006. "Fixed-Width Output Analysis for Markov Chain Monte Carlo". *Journal of the American Statistical Association* 101(476):1537–1547.

Lam, H. 2016a. "Advanced Tutorial: Input Uncertainty and Robust Analysis in Stochastic Simulation". In *Winter Simulation Conference (WSC), 2016*, edited by T. M. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, 178–192. Institute of Electrical and Electronics Engineers, Inc.

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

Lam, H. 2022a. "Cheap Bootstrap for Input Uncertainty Quantification". In *Proceedings of the 2022 Winter Simulation Conference*, edited by B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C. G. Corlu, L. H. Lee, E. P. Chew, T. Roeder, and P. Lendermann, 2318–2329. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Lam, H. 2022b. "A Cheap Bootstrap Method for Fast Inference". *arXiv:2202.00090*.

Lam, H., and Z. Liu. 2023. "Bootstrap in High Dimension with Low Computation". In *Proceedings of the 40th International Conference on Machine Learning*, edited by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Volume 202 of *Proceedings of Machine Learning Research*, 18419–18453: PMLR.

Lam, H., and H. Qian. 2019. "Random Perturbation and Bagging to Quantify Input Uncertainty". In *Proceedings of the 2019 Winter Simulation Conference*, edited by N. Mustafee, K.-H. G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, 320–331. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Lam, H., and H. Qian. 2022. "Subsampling to Enhance Efficiency in Input Uncertainty Quantification". *Operations Research* 70(3):1891–1913.

Lehmann, E. L., J. P. Romano, and G. Casella. 2005. *Testing Statistical Hypotheses*, Volume 3. Springer.

Lin, Y., E. Song, and B. L. Nelson. 2015. "Single-Experiment Input Uncertainty". *Journal of Simulation* 9:249–259.

Lopes, M. E. 2022. "Central Limit Theorem and Bootstrap Approximation in High Dimensions: Near $1/\sqrt{n}$ Rates via Implicit Smoothing". *Annals of Statistics* 50(5):2492–2513.

Martin, M. A. 1990. "On Using the Jackknife to Estimate Quantile Variance". *Canadian Journal of Statistics* 18(2):149–153.

Meketon, M. S., and B. Schmeiser. 1984. "Overlapping Batch Means: Something for Nothing?". In *Proceedings of the 1984 Winter simulation Conference*, edited by S. Sheppard, U. Pooch, and D. Pegden, 227–230. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Muñoz, D. F., and P. W. Glynn. 1997. "A Batch Means Methodology for Estimation of a Nonlinear Function of a Steady-State Mean". *Management Science* 43(8):1121–1135.

Nakayama, M. K. 2014, November. "Confidence Intervals for Quantiles Using Sectioning When Applying Variance-Reduction Techniques". *ACM Transactions on Modeling and Computer Simulation* 24(4):1–21.

Schmeiser, B. 1982. "Batch Size Effects in the Analysis of Simulation Output". *Operations Research* 30(3):556–568.

Schruben, L. 1983. "Confidence Interval Estimation using Standardized Time Series". *Operations Research* 31(6):1090–1108.

Serfling, R. J. 2009. *Approximation Theorems of Mathematical Statistics*, Volume 162. John Wiley & Sons.

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

Song, E., B. L. Nelson, and C. D. Pegden. 2014. "Advanced Tutorial: Input Uncertainty Quantification". In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk, S. Diallo, I. Ryzhov, L. Yilmaz, S. Buckley, and J. Miller, 162–176. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Song, W. T., and B. W. Schmeiser. 1995. "Optimal Mean-Squared-Error Batch Sizes". *Management Science* 41(1):110–123.

Steiger, N. M., E. K. Lada, J. R. Wilson, J. A. Joines, C. Alexopoulos, and D. Goldsman. 2005. "ASAP3: A Batch Means Procedure for Steady-State Simulation Analysis". *ACM Transactions on Modeling and Computer Simulation* 15(1):39–73.

Steiger, N. M., and J. R. Wilson. 2002. "An Improved Batch Means Procedure for Simulation Output Analysis". *Management Science* 48(12):1569–1586.

Su, Z., R. Pasupathy, Y. Yeh, and P. W. Glynn. 2018. "Overlapping Batch Confidence Intervals on Statistical Functionals Constructed from Time Series: Application to Quantiles, Optimization, and Estimation". https://web.ics.purdue.edu/~pasupath/PAPERS/2022suetal.pdf, *accessed 29th September 2023*.

Sun, Y., D. W. Apley, and J. Staum. 2011. "Efficient Nested Simulation for Estimating the Variance of a Conditional Expectation". *Operations Research* 59(4):998–1007.

Van der Vaart, A. W. 2000. *Asymptotic Statistics*, Volume 3. Cambridge University Press.

Zazanis, M. A., and R. Suri. 1993. "Convergence Rates of Finite-Difference Sensitivity Estimates for Stochastic Systems". *Operations Research* 41(4):694–703.

AUTHOR BIOGRAPHIES

HENRY LAM is an Associate Professor in the IEOR Department at Columbia University. His email address is henry.lam@columbia.edu.