Proceedings of the 2019 Winter Simulation Conference
N. Mustafee, K.-H.G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, eds.

RANDOM PERTURBATION AND BAGGING TO QUANTIFY INPUT UNCERTAINTY

Henry Lam
Huajie Qian

Department of Industrial Engineering and Operations Research
Columbia University
500 W. 120th Street
New York, NY 10027

ABSTRACT

We consider the problem of estimating the output variance in simulation analysis that is contributed from the statistical errors in fitting the input models, the latter often known as the input uncertainty. This variance contribution can be written in terms of the sensitivity estimate of the output and the variance of the input distributions or parameters, via the delta method. We study the direct use of this representation in obtaining efficient estimators for the input-contributed variance, by using finite-difference and random perturbation to approximate the gradient, focusing especially in the nonparametric case. In particular, we analyze a particular type of random perturbation motivated from resampling that connects to an infinitesimal jackknife estimator used in bagging. We illustrate the optimal simulation allocation and the simulation effort complexity of this scheme, and show some supporting numerical results.

1 INTRODUCTION

Simulation analysis consists of the repeated generation of random variates from pre-specified input models which, fed through the system logic, produces outputs that are used for prediction and decision-making. When the input models are misspecified or noisily estimated from data, these input errors can propagate to the outputs and lead to an incorrect conclusion. This calls for a need to quantify the impact of these errors, a problem often known as input uncertainty or extrinsic uncertainty, among other names in the literature. For general overview, see, e.g., Barton et al. (2002), Henderson (2003), Chick (2006), Barton (2012), Song et al. (2014), Lam (2016), and Nelson (2013) Chapter 7.

A common way in quantifying input uncertainty is to estimate the variance of the output that is contributed from the input noise. These variances carry information on the amount of variability coming from each individual input model or relative to the Monte Carlo noise in running the simulation replications. Moreover, they can be used as constituent ingredients to construct confidence intervals that account for both the noises from the inputs and the Monte Carlo runs. Established approaches to estimate these input-contributed variances (or "input variance" for short) roughly fall into two categories. First is to use bootstrap resampling (e.g., Cheng and Holland 1997; Song and Nelson 2015). This consists of resampling the input data to construct resampled input models, which are then used to drive simulation replications. The input variance can be approximated using the bootstrapped variance and a removal of the bias induced from the Monte Carlo noise via an analysis-of-variance. In the Bayesian contexts, this also leads to related approaches utilizing posterior sampling (Zouaoui and Wilson 2003; Zouaoui and Wilson 2004). The second line of approach uses delta-method approximation, which involves estimating the sensitivity or gradient of the outputs with respect to the input distributions or parameters, and combining with the estimation variance of these input quantities. These approaches include the two-point method (Cheng and Holland 1998) and regression-based approaches (Wieland and Schmeiser (2006); Lin et al. 2015; Song and Nelson

978-1-7281-3283-9/19/$31.00 ©2019 IEEE

2019). The first line of methods has been studied under both parametric and nonparametric frameworks, whereas the second line of methods has thus far focused on the parametric case.

A recurrent challenge in estimating input variance is the high computational cost, and relatedly the complexity in the simulation plan, in obtaining accurate estimates. On a high level, this is due to the convolution of the input and Monte Carlo noises for which, to obtain accurate estimation, one would need to disintegrate these two effects or wash away the Monte Carlo noise by abundant simulation runs. In bootstrap resampling, this challenge manifests in a nested sampling where one first resamples the input data (capturing the input noise) and then runs a number of replications per resample (capturing the Monte Carlo noise). Both layers need adequate sample sizes and lead to a multiplicatively large overall required simulation size. In delta-method approximation, the corresponding challenge comes from the gradient estimation. In general, without structural information, this involves a finite-difference estimator that requires a perturbation size selection to control both bias and variance, and maintaining a small overall error requires again a large enough Monte Carlo size.

In this paper, we consider the delta-method approximation to estimate input variance. Our focus is the investigation of the required simulation replication size needed for the involved gradient estimation, and in turn consistent and accurate estimation of the variance. We consider especially the nonparametric case, which to our best knowledge has not been studied and also exhibits additional complications compared to the parametric counterpart. In the basic form, our approach to estimate gradient is to use random perturbation, namely by first generating a random vector in the dimension of the parameters, and suitably projecting to each dimension to obtain a finite-difference estimate for the derivative with respect to each parameter. On a high level, this idea follows from the two-point method in Cheng and Holland (1998) and especially the regression approaches in Wieland and Schmeiser (2006) and Song and Nelson (2019). On the other hand, it also bears some differences with these works. First, by choosing the distribution of the perturbation vector properly, and using a "score function" multiplier like in simultaneous perturbation or smoothing used in zeroth-order stochastic optimization (e.g., Spall 1992; Nesterov and Spokoiny 2017; Flaxman et al. 2005), the projection of each dimension in our scheme automatically gives rise to small-bias gradient estimate thanks to the cancellation of higher-order errors. Second, in the nonparametric case, the input parameter is the entire input distribution function (potentially an infinite-dimensional object), and correspondingly the gradient is measured via the so-called influence function (Hampel 1974) in nonparametric statistics theory, which arises as a functional derivative. The length of the random perturbation vector or the "effective" dimension of the problem then grows with the input data size, and leads to a different analysis and ultimate simulation replication size requirement than the parametric case. Lastly, in this paper, we focus on a particular type of random perturbation using a multinomial distribution that is motivated from resampling. We demonstrate some connections of this scheme with the infinitesimal jackknife estimator for variance estimation in bagging (Wager et al. 2014). We also show some results on the optimal simulation allocation, and demonstrate how a small overall simulation effort of this scheme can be attained.

2 SETTING AND OVERVIEW OF THE APPROACH

Consider a target performance measure $\psi(F_1,\ldots,F_m)$ that depends on independent input probability distributions $F_1,\ldots,F_m$ (As an example, think of them as the interarrival and service times in a queueing system). Suppose that, given these distributions, the simulator can generate unbiased copies for ψ.

We consider the situation where $F_1,\ldots,F_m$ are not known, but instead i.i.d. data for each F_i are available. Call them $X_1^{(i)},\ldots,X_{n_i}^{(i)}$ where n_i is the sample size for model i. Then one can form the empirical distribution $\hat{F}_i(\cdot)=(1/n_i)\sum_{j=1}^{n_i}\delta_{X_j^{(i)}}(\cdot)$ (where $\delta_x(\cdot)$ is the delta measure at x), and a point estimate of $\psi(F_1,\ldots,F_m)$ from R simulation runs is given by

$$\hat{\psi}(\hat{F}_1,\ldots,\hat{F}_m)=\frac{1}{R}\sum_{r=1}^{R}\hat{\psi}_r(\hat{F}_1,\ldots,\hat{F}_m)$$

where $\hat{\psi}_r(\hat{F}_1,\ldots,\hat{F}_m)$ is an individual simulation run. The variance of $\hat{\psi}(\hat{F}_1,\ldots,\hat{F}_m)$ is

$$Var(\hat{\psi}(\hat{F}_1,\ldots,\hat{F}_m)) \approx \sum_{i=1}^{m} \frac{Var_{F_i}(IF_i(X^{(i)}))}{n_i} + \frac{\tau^2}{R}$$

where each term in the summation is the variance contributed from input model i (i.e., the input variance) and the last term is the variance contributed from the simulation noise (see Proposition 8 in Glynn and Lam 2018, and a parametric version in Theorem 1 in Cheng and Holland 1997). Here $Var_{F_i}(\cdot)$ refers to the variance under F_i and $X^{(i)}$ is a generic variable generated from F_i. $IF_i(\cdot)$ is the influence function for input model i, which is defined as the directional derivative

$$IF_i(x) = IF_i(x;F_1,\ldots,F_m) = \lim_{\varepsilon\to 0^+} \frac{\psi(F_1,\cdots,F_{i-1},(1-\varepsilon)F_i+\varepsilon\delta_x,F_{i+1},\ldots,F_m)-\psi(F_1,\cdots,F_m)}{\varepsilon}.$$

In other words, $IF_i(x)$ represents the infinitesimal effect in mixing F_i with the delta mass at x, δ_x. This function has mean zero under F_i, i.e., $E_{F_i}[IF_i(X^{(i)};F_1,\ldots,F_m)] = 0$. The constant $\tau^2 = Var(\psi_j(F_1,\ldots,F_m))$ is the variance of one simulation replication, under the true input distributions (or estimated distributions, which give rise to the same variance asymptotically under suitable smoothness conditions).

Our goal is to estimate the input variances $Var_{F_i}(IF_i(X^{(i)}))$. First of all, an asymptotically accurate estimate based on the input data is the empirical influence function variance given by

$$\frac{1}{n_i}\sum_{j=1}^{n_i} IF_i(X_j^{(i)};\hat{F}_1,\ldots,\hat{F}_m)^2. \tag{1}$$

This estimate generally has an error of order $O_p(1/\sqrt{\min_{i=1,\ldots,m} n_i})$ in estimating $Var_{F_i}(IF_i(X^{(i)}))$. The function $IF_i(x;\hat{F}_1,\ldots,\hat{F}_m)$ needs simulation approximation, i.e., we use the estimate

$$\frac{1}{n_i}\sum_{j=1}^{n_i} \widehat{IF}_i(X_j^{(i)};\hat{F}_1,\ldots,\hat{F}_m)^2 \tag{2}$$

where we use a random perturbation approach to estimate IF_i. In other words, letting $F = (F_1,\ldots,F_m)$ and $\hat{F} = (\hat{F}_1,\ldots,\hat{F}_m)$, we consider

$$\widehat{IF}_i(X_j^{(i)};\hat{F}) = \frac{1}{R}\sum_{r=1}^{R} \frac{\hat{\psi}_r(\hat{F}_1,\cdots,\hat{F}_{i-1},(1-h)\hat{F}_i+hG,\hat{F}_{i+1},\ldots,\hat{F}_m)}{h} S_j(G) \text{ for all } i=1,\ldots,m, j=1,\ldots,n_i$$

where G is a discrete distribution supported on the data $\{X_1^{(i)},\ldots,X_{n_i}^{(i)}\}$ that is randomized, and $S_j(G)$ for each j is some "score function" of the perturbation such that $E[\psi(\hat{F}_1,\cdots,\hat{F}_{i-1},(1-h)\hat{F}_i+hG,\hat{F}_{i+1},\ldots,\hat{F}_m)S_j(G)/h|\hat{F}] \approx IF_i(X_j^{(i)};\hat{F})$ for all $i=1,\ldots,m$ and $j=1,\ldots,n_i$. Note that, given the input data, every realization of G can be identified with a vector of probability weights $(G_1,\ldots,G_{n_i})$ on the n_i observations through the relation $P_G(\{X_j^{(i)}\}) = G_j$ for each $j=1,\ldots,n_i$ where P_G denotes the probability under the distribution G, therefore for convenience G will be referred to as a random discrete distribution or a random vector on the probability simplex $\{(G_1,\ldots,G_{n_i})\in\mathbb{R}^{n_i} : \sum_{j=1}^{n_i} G_j = 1, G_j \geq 0 \text{ for all } j\}$ interchangeably. To get a meaningful estimate for $Var_{F_i}(IF_i(X^{(i)}))$, we would like to control the approximation error of (2) for (1) to be $o_p(1)$.

Since one can always estimate each input variance $Var_{F_i}(IF_i(X^{(i)}))$ separately by fixing all others at the empirical distributions, for simplicity we focus on the case of a single input model F ($m=1$) with n i.i.d. observations $X_1,\ldots,X_n$ and our goal is to estimate the input variance $Var_F(IF(X))$. A convenient formula for $S_i(G)$ can be derived for exchangeable random perturbations, i.e., G such that $(G_1,\ldots,G_n)$ and

$(G_{\pi(1)},\dots,G_{\pi(n)})$ have the same distribution for all permutations $\pi:\{1,2,\dots,n\}\to\{1,2,\dots,n\}$. Hence we have $E[G_1]=\frac{1}{n}$ and $\sum_{i=1}^n(G_i-\frac{1}{n})=0$, which implies $\sum_{i=1}^n E[(G_1-\frac{1}{n})(G_i-\frac{1}{n})]=0=Var(G_1)+(n-1)Cov(G_1,G_2)$, i.e., $Cov(G_1,G_2)=-\frac{1}{n-1}Var(G_1)$. Therefore, we let

$$S_i(G)=\frac{n-1}{nVar(G_1)}(G_i-\frac{1}{n}).$$

For sufficiently smooth $\psi(\cdot)$, the perturbed performance measure admits a first order Taylor expansion $\psi((1-h)\widehat{F}+hG)\approx\psi(\widehat{F})+hE_{X\sim G}[IF(X;\widehat{F})|\widehat{F}]=\psi(\widehat{F})+h\sum_{i=1}^n IF(X_i;\widehat{F})G_i$ (see, e.g., Van der Vaart 2000 Chapter 20) around the empirical distribution $\widehat{F}$. Therefore

$$\begin{aligned}E\big[\frac{\psi((1-h)\widehat{F}+hG)}{h}S_i(G)|\hat{F}\big]&\approx E\big[\frac{\psi(\widehat{F})}{h}S_i(G)+\sum_{i'=1}^n IF(X_{i'};\widehat{F})(G_{i'}-\frac{1}{n})S_i(G)|\hat{F}\big]\\&=\frac{n-1}{nVar(G_1)}\sum_{i'\neq i}IF(X_{i'};\widehat{F})Cov(G_{i'},G_i)+\frac{n-1}{n}IF(X_i;\widehat{F})\\&=-\frac{1}{n}\sum_{i'\neq i}IF(X_{i'};\widehat{F})+\frac{n-1}{n}IF(X_i;\widehat{F})=IF(X_i;\widehat{F})\end{aligned}$$

recovering the influence function, where the last equality follows because $\sum_{i=1}^n IF(X_i;\widehat{F})=0$.

3 PERTURBATION MOTIVATED FROM RESAMPLING

In this section we analyze a special yet flexible class of random perturbations that are motivated from resampling of the input data. Consider the multinomial distribution $(s,\frac{1}{n}\mathbf{1})$ with $\mathbf{1}=(1,1,\dots,1)\in\mathbb{R}^n$ that counts s independent and uniform draws from n objects. Letting s_i to be the count of the i-th object drawn ($\sum_{i=1}^n s_i=s$), we set the i-th component of perturbation to be $G_i=s_i/s$, and h to be 1 so that $\psi((1-h)\widehat{F}+hG)=\psi(G)$. It is straightforward to see that $Var(G_1)=\frac{1}{sn}(1-\frac{1}{n})$, hence the appropriate multiplier $S_i(G)=sn(G_i-\frac{1}{n})$ for $i=1,\dots,n$. This leads to the following estimator for $Var_F(IF(X))$ (assuming infinite simulation replications for now)

$$\sigma_s^2=\frac{1}{n}\sum_{i=1}^n\big[Cov_*(\psi(G),sn(G_i-\frac{1}{n}))\big]^2 \tag{3}$$

where Cov_* denotes the covariance with respect to the perturbation G given the input data $X_1,\dots,X_n$ (the data serve as the support of the distribution G), and the subscript s indicates the dependence on s. Adaptations of the estimator (3) to the case of multiple input models shall be discussed briefly at the end of this section.

3.1 Connection to Bagging and Infinitesimal Jackknife

Note that such a random perturbation G coincides with the empirical distribution induced by the resample $\{X_1^*,\dots,X_s^*\}$ of size s drawn uniformly from the data with replacement. In particular, when $s=n$ it coincides with the standard bootstrap resample. This connects the multinomial perturbation to bagging or bootstrap aggregating, a popular ensemble technique (i.e., technique that combines multiple base learners) in machine learning, and interprets the estimator (3) as a so-called Infinitesimal Jackknife (IJ) variance estimator (Efron 2014; Wager et al. 2014).

Bagging has been used to reduce prediction error of base learners especially tree-based methods, such as random forests (Breiman 2001). It possesses variance reduction properties (Bühlmann and Yu 2002) and is capable of smoothing statistical estimators that involve model selection (Efron 2014). Given a training set for a prediction problem, bagging repeatedly draws bootstrap resamples from the data, trains a base learner from each resample, and at the end averages all individual base learners as the bagged learner. For example, a random forest consists of bagged decision trees. The IJ variance estimator estimates the variance of the bagged learner, and can be readily computed from the bagging procedure as a by-product with negligible computation overhead.

The estimator (3) is connected to the IJ estimator as follows. Consider the "plug-in" estimator $\psi(\widehat{F})$ of the performance measure as a base learner. With $\widehat{F}^*$ representing a bootstrap resample, i.e., the perturbation G when $s=n$, a bagged estimator of the performance measure is

$$\psi_{bag}(\widehat{F}) = E_*[\psi(\widehat{F}^*)].$$

Supposing that $\psi(\cdot)$ is sufficiently smooth, then, by linearizing $\psi(\widehat{F}^*)$ around $\widehat{F}$, one can argue that the bias of $\psi_{bag}(\widehat{F})$ relative to $\psi(\widehat{F})$ is of order $1/n$, i.e., $\psi_{bag}(\widehat{F}) = \psi(\widehat{F}) + O_p(1/n)$. This observation also reveals that the variance of the bagged estimator is close to that of the plug-in estimator in the sense

$$\begin{aligned} Var(\psi_{bag}(\widehat{F})) &= Var(\psi(\widehat{F}) + O_p(1/n)) \\ &= Var(\psi(\widehat{F})) + 2Cov(\psi(\widehat{F}), O_p(1/n)) + Var(O_p(1/n)) \\ &\approx Var(\psi(\widehat{F})) + O\big(1/n^{3/2}\big) \end{aligned}$$

where $O(1/n^{3/2})$ is the correct order because $Cov(\psi(\widehat{F}), O_p(1/n)) = O(1/n^{3/2})$ by Cauchy Schwartz inequality and $Var(O_p(1/n)) = O(1/n^2)$. Note that the bagged estimator $\psi_{bag}(\widehat{F})$ is in the form of an expectation over the resamples, therefore is smooth in $\widehat{F}$. This IJ estimator for its variance is derived from applying the delta method on $\psi_{bag}(\widehat{F})$. The following shows the form of the IJ variance estimator (Efron 2014):

Lemma 1 The IJ estimator of Var$(\psi_{bag}(\widehat{F}))$ is

$$\sigma_{IJ}^2 = \sum_{i=1}^{n} \text{Cov}_*(\psi(\widehat{F}^*), N_i^*)^2$$

where N_i^* counts the number of replications of data point X_i in a bootstrap resample. Moreover, letting δ_x be the delta measure at x, we have

$$n\text{Cov}_*(\psi(\widehat{F}^*), N_i^*) = \lim_{\varepsilon \to 0+} \frac{1}{\varepsilon}(\psi_{bag}((1-\varepsilon)\widehat{F} + \varepsilon\delta_{X_i}) - \psi_{bag}(\widehat{F}).$$

To link our random-perturbation variance estimator (3) to the IJ estimator, consider the case $s=n$. Note that $G = \widehat{F}^*$ and that $(nG_i - 1)$ can be identified with N_i^*, therefore $nCov_*(\psi(\widehat{F}^*), N_i^*) = Cov_*(\psi(G), n^2(G_i - \frac{1}{n}))$ for each i and our estimator $\sigma_n^2 = n\sigma_{IJ}^2$ which reconfirms the relation $Var(\psi_{bag}(\widehat{F})) \approx Var(\psi(\widehat{F})) \approx Var_F(IF(X))/n$.

3.2 Estimation Procedure

The estimator (3) assumes a hypothetical infinite number of bootstrap samples. In reality, one needs to use a finite number of perturbations and simulation size per perturbation. Algorithm 1 shows an estimator that incorporates these simulation runs for σ_s^2. By associating the multinomial perturbation with resampling, one can see that each $\widehat{F}^b = G$, each N_i^b/s corresponds to the i-th component G_i, and $nN_i^b - s$ is the multiplier $sn(G_i - \frac{1}{n})$ in (3).

Next, we explain briefly why $\hat{\sigma}_s^2$ should possess a central limit theorem, and in particular implies a consistent estimator for σ_s^2. Our explanation is based on the observation that $\hat{\sigma}_s^2$ is approximately a

U-statistic (Lee 1990; Serfling 2009; Van der Vaart 2000), a type of symmetric statistic that is constructed from aggregating all combinations (without replacement) of the data. To this end, we ignore the error of the gross average $\bar{\bar{\psi}}$ in estimating the true mean $E_*[\psi(G)]$, so that the output of Algorithm 1 can be written in the form

$$\hat{\sigma}_s^2 \approx \frac{1}{B(B-1)}\sum_{b\neq b'}\left[(\bar{\psi}^b - E_*[\psi(G)])(\bar{\psi}^{b'} - E_*[\psi(G)])\frac{1}{n}\sum_{i=1}^{n}(nN_i^b - s)(nN_i^{b'} - s)\right] = \frac{1}{B(B-1)}\sum_{b\neq b'} h(\xi_b,\xi_{b'})$$

where each ξ_b represents the random variates generated in the resampling and the simulation for the b-th perturbation, and h is an appropriate symmetric kernel that maps the pair $\xi_b,\xi_{b'}$ to $(\bar{\psi}^b - E_*[\psi(G)])(\bar{\psi}^{b'} - E_*[\psi(G)])\frac{1}{n}\sum_{i=1}^{n}(nN_i^b - s)(nN_i^{b'} - s)$. This is precisely a U-statistic estimator for σ_s^2 by noting that $E_*[h(\xi_b,\xi_{b'})] = \sigma_s^2$ for all $b \neq b'$. Since U-statistics admit central limit theorems (CLTs), a CLT of the form $\sqrt{B}(\hat{\sigma}_s^2 - \sigma_s^2) \to \mathcal{N}(0,\rho^2)$ is expected to hold for $\hat{\sigma}_s^2$ when B is chosen large enough.

Algorithm 1 Multinomial Perturbation

```
Parameters: $B \geq 2, R, s \leq n$
for $b = 1$ to $B$ do
    Draw a resample $\{X_1^b,\ldots,X_s^b\}$ of size $s$ from $\{X_1,\ldots,X_n\}$, inducing an empirical distribution $\widehat{F}^b$
    Compute $N_i^b =$ number of $X_i$ in $\{X_1^b,\ldots,X_s^b\}$ for all $i$
    for $r = 1$ to $R$ do
        Draw a $\hat{\psi}_r(\widehat{F}^b)$
    end for
    Compute $\bar{\psi}^b = \frac{1}{R}\sum_{r=1}^{R}\hat{\psi}_r(\widehat{F}^b)$
end for
Compute $\bar{\bar{\psi}} = \frac{1}{B}\sum_{b=1}^{B}\bar{\psi}^b$
Output
    $\hat{\sigma}_s^2 = \frac{1}{B(B-1)}\sum_{b\neq b'}\left[(\bar{\psi}^b - \bar{\bar{\psi}})(\bar{\psi}^{b'} - \bar{\bar{\psi}})\frac{1}{n}\sum_{i=1}^{n}(nN_i^b - s)(nN_i^{b'} - s)\right]$
```

We note that the subsampling scheme in Algorithm 1 resembles the subbagging proposal in Mentch and Hooker (2016). In the latter, subsampling is introduced to ensure central limit convergence, viewing a random forest as an infinite-order U-statistic. This is contrary to our use of subsampling here as a tool to alleviate the total simulation requirement as the next section shows.

3.3 Optimal Simulation Budget Allocation and Overall Effort

In this section we present theories on the budget allocation and the overall required simulation effort in using Algorithm 1. We start by showing validity of the variance estimator (3) assuming absence of Monte Carlo errors, and then examine the simulation effort required to achieve consistency in the presence of Monte Carlo errors, as well as optimal allocation of a given simulation budget.

We shall frequently use the notations $\omega(\cdot)$ and $\Theta(\cdot)$. $a = \omega(b)$ means $a/b \to \infty$ as $n,s \to \infty$, and $a = \Theta(b)$ means that there exist constants $\underline{c},\overline{c} > 0$ such that $\underline{c}b \leq a \leq \overline{c}b$ for all n,s.

We assume the following smoothness conditions for ψ:

Assumption 1 (Smoothness at true input model)

$$\psi(\widehat{F}) = \psi(F) + \sum_{i=1}^{m}\int IF(x;F)d(\widehat{F}_i - F_i)(x) + \varepsilon$$

where the remainder satisfies $E[\varepsilon^2] = o(n^{-1})$, and the influence functions IF satisfies $Var_F(IF(X;F)) > 0$ and $E_F[IF^4(X;F)] < \infty$.

Assumption 2 (Smoothness at empirical input model) Assume the empirical influence function converges to the truth in the sense that $E[(IF(X_1;\widehat{F})-IF(X_1;F))^4]\to 0$. Let $\widehat{F}_s^*$ be the resample of size s. Assume the remainder in the Taylor expansion

$$\psi(\widehat{F}_s^*)=\psi(\widehat{F})+\sum_{i=1}^{m}\int IF(x;\widehat{F})d(\widehat{F}_s^*-\widehat{F}_i)(x)+\varepsilon^*$$

satisfies $E_*[(\varepsilon^*-E_*[\varepsilon^*])^4]=o_p\big(s^{-3}\big)$.

We have the following consistency result of the estimator σ_s^2:

Theorem 1 Under Assumptions 1 and 2, it holds

$$\sigma_s^2\to\sigma^2 \text{ in probability as } n\to\infty, s\to\infty$$

where $\sigma^2:=Var_F(IF(X))$ and s is allowed to grow independently of n.

The important message from Theorem 1 is that the resample size s is allowed to grow at an arbitrary rate relative to the data size n.

Now we examine the simulation effort needed for Algorithm 1. We first need to figure out the mean square error (MSE) of the variance estimate $\hat{\sigma}_s^2$ output from Algorithm 1. To proceed, denote by $\tau^2=Var(\hat{\psi}(F))$ the variance of a single simulation replication under the true input model. Correspondingly let $\hat{\tau}^2=Var(\hat{\psi}(\widehat{F}))$ and $\hat{\tau}^{*2}=Var_*(\hat{\psi}(\widehat{F}_s^*))$ be the variances when driven by the empirical input model and the resampled model.

Assumption 3 (Convergence of simulation variance) $\hat{\tau}^2\to\tau^2$ in probability as $n\to\infty$, and $E_*[(\hat{\tau}^{*2}-\hat{\tau}^2)^2]=o_p(1)$ as both $n,s\to\infty$.

Assumptions 1-3 may seem complicated at first glance. However, they can be readily checked to hold for performance measures with finite time horizons under mild conditions.

Theorem 2 (Finite horizon measure) If $\psi(F)=E_F[h(\mathbf{X})]$, where $\mathbf{X}=(X(1),\dots,X(T))$ consists of T i.i.d. variables distributed under F, h is some performance function, and

1. $0<Var_F(\sum_{t=1}^{T}E_F[h(\mathbf{X})|X(t)=X)<\infty$
2. Let $I=(I(1),\dots,I(T))$ be a sequence of indices such that each $1\leq I(t)\leq T$, and denote by $\mathbf{X}_I=(X(I(1)),\dots,X(I(T)))$. It holds $\max_I E_F[|h(\mathbf{X}_I)|^4]<\infty$

then Assumptions 1-3 hold.

The MSE of $\hat{\sigma}_s^2$ can be characterized as:

Lemma 2 Under Assumptions 1-3, if the outer size B and resample size s are chosen to be $\omega(1)$, then the variance estimate $\hat{\sigma}_s^2$ output by Algorithm 1 has the following decomposition

$$\hat{\sigma}_s^2=\sigma_s^2+\mathscr{E}+o_p\big(\sqrt{E[\mathscr{E}^2]}\big)$$

where the leading error $\mathscr{E}$ satisfies $E[\mathscr{E}]=0$ and

$$E[\mathscr{E}^2]=\frac{2s^2n}{B^2}\Big(\frac{\sigma^2}{s}+\frac{\tau^2}{R}\Big)^2+\frac{4s\sigma^2}{B}\Big(\frac{2\sigma^2}{s}+\frac{\tau^2}{R}\Big). \tag{4}$$

Note that, given a total simulation budget $N=BR$, the error (4) takes the form

$$\frac{c_1}{B^2}+\frac{c_2}{BN}+\frac{c_3}{N^2}+\frac{c_4}{B}+\frac{c_5}{N}$$

where $c_i, i=1,\dots,5$ do not depend on B or R. Therefore, in order to minimize the error by adjusting B,R, it is optimal to set B as large as possible, i.e., $B=N$. This is summarized below:

Theorem 3 (Optimal allocation) Under Assumptions 1-3, as the input data size $n \to \infty$, the total simulation budget $N \to \infty$ and resample size $s \to \infty$, the optimal B, R that minimize the leading term in the MSE of $\hat{\sigma}_s^2$ in estimating σ_s^2 are

$$B^* = N, R^* = 1$$

which gives an error $\hat{\sigma}_s^2 - \sigma_s^2 = \mathscr{E} + o_p(\sqrt{E[\mathscr{E}^2]})$ where

$$E[\mathscr{E}^2] = \frac{2s^2 n}{N^2}\big(\frac{\sigma^2}{s} + \tau^2\big)^2 + \frac{4s\sigma^2}{N}\Big(\frac{2\sigma^2}{s} + \tau^2\Big). \tag{5}$$

Theorem 3 is appealing in terms of the transparency of allocation rule. The allocation essentially removes the nestedness of the simulation procedure by requiring only one simulation run per resample. This property comes intuitively from the observation that the ψ in the covariance term in the estimator (3) is unchanged by replacing with merely one simulation run $\hat{\psi}_r$.

By forcing the MSE (4) to $o(1)$, we identify the range of B, R, s that produces a statistically valid variance estimate, giving rise to the overall required simulation effort:

Theorem 4 (Overall simulation effort requirement) Under Assumptions 1-3, if the parameters B, R and s of Algorithm 1 are chosen such that

$$B = \omega\big(n^{1/2}\big),\ BR = \omega\big(sn^{1/2}\big),\ s = \omega(1) \tag{6}$$

then $\hat{\sigma}_s^2$ is consistent, i.e., $\hat{\sigma}_s^2 \to \sigma^2$ in probability. In particular, the minimum required total simulation budget is $N := BR = \omega\big(n^{1/2}\big)$ when the resample size is set to $s = \omega(1)$.

The second condition in (6) forces the total simulation effort to be $N = \omega(sn^{1/2})$. On the other hand, s can grow at an arbitrary rate according to Theorem 1, thus leading to the minimum required effort $N = \omega\big(n^{1/2}\big)$. This shows the advantage, in terms of the required simulation effort, of random perturbation over individualized estimation, i.e., estimating each $IF(X_i;\widehat{F})$ by $(\hat{\psi}((1-h)\widehat{F} + h\delta_{X_i}) - \hat{\psi}(\widehat{F}))/h$. With individualized estimation, obtaining an estimate for each $IF(X_i;\widehat{F})$ costs at least one simulation run, hence the required simulation effort has to scale at least linearly in n, whereas with random perturbation an effort of order $\sqrt{n}$ would suffice.

Finally, we present the optimal resample size s given data size n and budget N to minimize the gross error of $\hat{\sigma}_s^2$ in estimating σ^2. Note that $\hat{\sigma}_s^2 - \sigma^2 = (\hat{\sigma}_s^2 - \sigma_s^2) + (\sigma_s^2 - \sigma^2)$. The first error is the Monte Carlo error with a MSE leading term (5) if single-run allocation is used, and is increasing in s. The second error however is decreasing in s because a larger s corresponds to a smaller perturbation hence better approximation to the influence function. The optimal s is then found by balancing the two errors. We have the following result concerning the optimal resample size (note that the optimal allocation is always non-nested in this case thanks to Theorem 3):

Theorem 5 (Optimal resample size) Let ψ be the finite horizon performance measure from Theorem 2. For a given simulation budget $N = \omega(n^{1/2})$ and data sizes n, if optimal resample size s and B, R of Algorithm 1 are

$$\begin{cases} s^* = \Theta(N^{1/2}n^{-1/4}) & \text{if } N \le n^{3/2} \\ \Theta(n^{1/2}) \le s^* \le \Theta((Nn^{-1}) \wedge n) & \text{if } N > n^{3/2} \end{cases}$$
$$B^* = N, R^* = 1$$

then we have a gross error $\hat{\sigma}_s^2 - \sigma^2 = \mathscr{E} + o_p(N^{-1/2}n^{1/4} + n^{-1/2})$, where the leading term has an MSE

$$E[\mathscr{E}^2] = \Theta\big(\frac{\sqrt{n}}{N} + \frac{1}{n}\big).$$

We comment that in the case of multiple input models $m > 1$, one can estimate the individual variance $Var_{F_i}(IF_i(X^{(i)}))$ by perturbing the i-th model only and using Algorithm 1, as mentioned before, arriving at the following estimator that is similar to (3)

$$\widehat{Var_{F_i}(IF_i(X^{(i)}))} = \frac{1}{n_i}\sum_{j=1}^{n_i}\left[Cov_*(\psi(\hat{F}_1,\ldots,\hat{F}_{i-1},G,\hat{F}_{i+1},\ldots,\hat{F}_m), sn_i(G_i - \frac{1}{n_i}))\right]^2 \text{ for all } i = 1,\ldots,m.$$

An alternative is to perturb all the models simultaneously, i.e., using $\psi(G^{(1)},\ldots,G^{(m)})$ where each $G^{(i)}$ is a random perturbation for the i-th model, and estimate all the m influence functions simultaneously. This leads to the estimator

$$\widehat{Var_{F_i}(IF_i(X^{(i)}))} = \frac{1}{n_i}\sum_{j=1}^{n_i}\left[Cov_*(\psi(G^{(1)},\ldots,G^{(m)}), s_i n_i(G_j^{(i)} - \frac{1}{n_i}))\right]^2 \text{ for all } i = 1,\ldots,m$$

where s_i is the resample size used for the i-th model. This simultaneous perturbation is achieved by drawing resamples from each of the m input models with possibly different resample sizes.

4 NUMERICAL RESULTS

We consider an M/M/1 queuing system with arrival rate 0.5 and service rate 1. Suppose the system is empty at time zero. The performance measure of interest is the probability that the waiting time of the 20-th arrival exceeds 2 units of time, whose true value is approximately 0.182. Specifically, the system has two input distributions, i.e. the inter-arrival time distribution $F_1 = \text{Exp}(0.5)$ and the service time distribution $F_2 = \text{Exp}(1)$, for which we have n_1 and n_2 i.i.d. data available respectively. If A_t is the inter-arrival time between the t-th and $(t+1)$-th arrivals, and S_t is the service time for the t-th arrival, then the system output

$$\psi(F_1,F_2) = \mathbb{E}_{F_1,F_2}[\mathbf{1}\{W_{20} > 2\}]$$

where $\mathbf{1}\{W_{20} > 2\} = 1$ if $W_{20} > 2$ and 0 otherwise, and the waiting time W_{20} is calculated by the Lindley recursion $W_1 = 0, W_{t+1} = \max\{W_t + S_t - A_t, 0\}$ for $t = 1,2,\ldots,19$.

We test Algorithm 1 under different input data sizes. For each input data size, 1,000 95%-level CIs are constructed for the true performance measure $\psi(F)$, each from an independently generated input data set. We allocate 1,000 simulation runs to compute $\hat{\sigma}_1^2, \hat{\sigma}_2^2$ using Algorithm 1, and another 200 simulation runs driven by the empirical input distributions to compute the point estimator $\bar{\psi}(\widehat{F}_1,\widehat{F}_2)$. Let $\hat{\tau}^2$ be the sample variance of the 200 simulation replications. Then the 95%-level CI is computed as $\bar{\psi}(\widehat{F}_1,\widehat{F}_2) \pm 1.96\sqrt{\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2 + \hat{\tau}^2/200}$. Tables 1, 2 and 3 report the coverage probability, mean and standard deviation of the CI lengths, the estimated ratio between input and simulation standard deviations, and the number of times that the algorithm outputs a negative variance estimate (the estimate is not guaranteed to be always positive, and reset to 0 if it turns out negative).

Table 1: Algorithm 1 with $B = 1{,}000, R = 1, s_1 = n_1, s_2 = n_2$.

input data sizes	coverage estimate	mean CI length	std. CI length	$\frac{\text{input se.}}{\text{sim. se.}}$	# neg var
$n_1 = 60, n_2 = 30$	85.1%	0.427	0.179	4.25	6
$n_1 = 200, n_2 = 100$	83.7%	0.252	0.114	2.42	176
$n_1 = 600, n_2 = 300$	89.4%	0.212	0.118	2.04	419
$n_1 = 2{,}000, n_2 = 1{,}000$	95.1%	0.242	0.168	2.56	479

The results show that, compared to standard bootstrap for which $s = n$ (Table 1), using a relatively small resample size s (Tables 2 and 3) can help reduce the variability of the variance estimate and generate

Table 2: Algorithm 1 with $B = 1{,}000, R = 1, s_1 = s_2 = 30$.

input data sizes	coverage estimate	mean CI length	std. CI length	$\frac{\text{input se.}}{\text{sim. se.}}$	# neg var
$n_1 = 60, n_2 = 30$	84.4%	0.420	0.169	4.20	2
$n_1 = 200, n_2 = 100$	92.8%	0.264	0.064	2.30	2
$n_1 = 600, n_2 = 300$	94.9%	0.179	0.031	1.34	9
$n_1 = 2{,}000, n_2 = 1{,}000$	95.2%	0.134	0.016	0.72	82

Table 3: Algorithm 1 with $B = 1{,}000, R = 1$ and $s_i = \min(\sqrt{1{,}000}n_i^{-1/4}, n_i^{1/2})$, the optimal resample size suggested by Theorem 5.

input data sizes	coverage estimate	mean CI length	std. CI length	$\frac{\text{input se.}}{\text{sim. se.}}$	# neg var
$n_1 = 60, n_2 = 30$	87.2%	0.392	0.103	3.59	0
$n_1 = 200, n_2 = 100$	93.1%	0.255	0.043	2.11	0
$n_1 = 600, n_2 = 300$	93.5%	0.174	0.016	1.18	0
$n_1 = 2{,}000, n_2 = 1{,}000$	94.7%	0.133	0.007	0.61	0

statistically more accurate CIs. More negative variance estimates suggests more variability. Table 1 has much more negative variance estimates than the other two tables, therefore using $s = n$ leads to much larger variability in the estimate. Another sign of estimation variability is the much larger standard deviations of the CI lengths in Table 1, especially for large data sizes. This is also reflected through the coverage probability. CIs in Table 1 tend to undercover the truth, except for the case $n_1 = 2{,}000, n_2 = 1{,}000$ which however have much wider CIs. This misleading phenomenon is probably because that manually resetting negative variance estimates to zero makes the estimator upward biased which in turn gives rise to wider confidence intervals. Actually the mean CI length for the case $n_1 = 2{,}000, n_2 = 1{,}000$ in Table 1 is 0.242, almost twice of those in Tables 2 and 3.

Based on the trend of the coverage probability as the data sizes grow, our variance estimate for input uncertainty appears statistically consistent. In both Tables 2 and 3, as the data size increases from 30 to 2,000, the coverage probability approaches the nominal value 95%. Note that this happens under a fixed the simulation budget 1,000 for estimating the input variance, which demonstrates that our algorithm is generating statistically valid variance estimate even when the data size is of the same order as the simulation budget, verifying Theorem 4.

Lastly, compared to the optimal resample size (Table 3) suggested by Theorem 5, the simple strategy of using $s_1 = s_2 = 30$ (Table 2) seems to work equally well in the considered cases. Since the optimal size from Theorem 5 contains hidden constants, in practice it could be useful to have simple and workable strategies like this one.

5 CONCLUSION

We have studied a direct use of the delta method in estimating nonparametric input variance in simulation estimation under input uncertainty. Our approach is based on a random perturbation to obtain a finite-difference estimator that approximates the gradient information needed in the delta method, exhibited by the influence function in the nonparametric setting. By multiplying with a proper score function as in simultaneous perturbation and other smoothing methods, the projection of our random perturbation automatically cancels out higher-order approximation terms and leads to a small bias in estimating the gradient. We have specialized our perturbation distribution choice to a multinomial distribution over the input data set, which connects our approach to the infinitesimal jackknife estimator used in approximating the variance of bagging. We have presented the consistency, sample size requirement and optimal sampling allocation strategy for our estimator, and provided numerical examples to support our theoretical findings. Future work includes the full development and expansion of our analyses to more general classes of random perturbation estimators and more elaborate numerical comparisons with other approaches in both parametric and nonparametric settings.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the National Science Foundation under grants CAREER CMMI-1653339/1834710 and IIS-1849280.

REFERENCES

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.

Breiman, L. 2001. “Random Forests”. *Machine Learning* 45(1):5–32.

Bühlmann, P., and B. Yu. 2002. “Analyzing Bagging”. *The Annals of Statistics* 30(4):927–961.

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.

Cheng, R. C., and W. Holland. 1998. “Two-Point Methods for Assessing Variability in Simulation Output”. *Journal of Statistical Computation Simulation* 60(3):183–205.

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.

Efron, B. 2014. “Estimation and Accuracy After Model Selection”. *Journal of the American Statistical Association* 109(507):991–1007.

Flaxman, A. D., A. T. Kalai, A. T. Kalai, and H. B. McMahan. 2005. “Online Convex Optimization in the Bandit Setting: Gradient Descent Without a Gradient”. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA ’05, 385–394. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics.

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.

Hampel, F. R. 1974. “The Influence Curve and Its Role in Robust Estimation”. *Journal of the American Statistical Association* 69(346):383–393.

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: Institute of Electrical and Electronics Engineers, Inc.

Lam, H. 2016. “Advanced Tutorial: Input Uncertainty and Robust Analysis in Stochastic Simulation”. In *Proceedings of the 2016 Winter Simulation Conference*, edited by T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, 178–192. Piscataway, New Jersy: Institute of Electrical and Electronics Engineers, Inc.

Lee, A. J. 1990. *U-Statistics: Theory and Practice*. New York: Taylor & Francis Group.

Lin, Y., E. Song, and B. Nelson. 2015. “Single-Experiment Input Uncertainty”. *Journal of Simulation* 9(3):249–259.

Mentch, L., and G. Hooker. 2016. “Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests”. *The Journal of Machine Learning Research* 17(1):841–881.

Nelson, B. 2013. *Foundations and Methods of Stochastic Simulation: A First Course*. New York: Springer Science & Business Media.

Nesterov, Y., and V. Spokoiny. 2017. “Random Gradient-Free Minimization of Convex Functions”. *Foundations of Computational Mathematics* 17(2):527–566.

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

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

Song, E., and B. L. Nelson. 2019. “Input-Output Uncertainty Comparisons for Discrete Optimization via Simulation”. *Operations Research* 67(2):562–576.

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. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, 162–176. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Spall, J. C. 1992. “Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation”. *IEEE Transactions on Automatic Control* 37(3):332–341.

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

Wager, S., T. Hastie, and B. Efron. 2014. “Confidence Intervals for Random Forests: the Jackknife and the Infinitesimal Jackknife.”. *Journal of Machine Learning Research* 15(1):1625–1651.

Wieland, J. R., and B. W. Schmeiser. 2006. “Stochastic Gradient Estimation Using a Single Design Point”. 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, 390–397. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Zouaoui, F., and J. R. Wilson. 2003. “Accounting for Parameter Uncertainty in Simulation Input Modeling”. *IIE Transactions* 35(9):781–792.

Zouaoui, F., and J. R. Wilson. 2004. “Accounting for Input-Model and Input-Parameter Uncertainties in Simulation”. *IIE Transactions* 36(11):1135–1151.

AUTHOR BIOGRAPHIES

HENRY LAM is an Associate Professor in the Department of Industrial Engineering and Operations Research at Columbia University. His research focuses on Monte Carlo simulation, uncertainty quantification, risk analysis, and stochastic and robust optimization. His email address is khl2114@columbia.edu.

HUAJIE QIAN is a Ph.D. student in the Department of Industrial Engineering and Operations Research at Columbia University. His research interest lies in simulation uncertainty quantification, data-driven simulation analysis, and stochastic optimization. His email address is h.qian@columbia.edu.