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

EFFICIENT INPUT UNCERTAINTY QUANTIFICATION FOR REGENERATIVE SIMULATION

Linyun He
Eunhye Song

School of Industrial and Systems Engineering
Georgia Institute of Technology
755 Ferst Dr. NW
Atlanta, GA 30332, USA

Ben Feng

Department of Statistics and Actuarial Science
University of Waterloo
200 University Ave W
Waterloo, ON N2L 3G1, CANADA

ABSTRACT

The initial bias in steady-state simulation can be characterized as the bias of a ratio estimator if the simulation model has a regenerative structure. This work tackles input uncertainty quantification for a regenerative simulation model when its input distributions are estimated from finite data. Our aim is to construct a bootstrap-based confidence interval (CI) for the true simulation output mean performance that provides a correct coverage with significantly less computational cost than the traditional methods. Exploiting the regenerative structure, we propose a k-nearest neighbor (kNN) ratio estimator for the steady-state performance measure at each set of bootstrapped input models and construct a bootstrap CI from the computed estimators. Asymptotically optimal choices for k and bootstrap sample size are discussed. We further improve the CI by combining the kNN and likelihood ratio methods. We empirically compare the efficiency of the proposed estimators with the standard estimator using queueing examples.

1 INTRODUCTION

In stochastic simulation, input models refer to the distributions that generate random inputs fed into the simulation logic. These input models typically represent random processes observed in an existing system. In this case, the input models can be estimated by collecting observations from the real-world random processes and fitting distribution to the data. Since data is always finite, the input models are subject to estimation errors. Input uncertainty refers to the variability in the stochastic simulation output caused by such estimation error in the input models. Quantifying input uncertainty helps us make a correct statistical inference about the performance measure under the true input distributions.

In this paper, we focus on input uncertainty quantification (IUQ) for a *steady-state* simulation model. We assume that the distribution families of the input models are known, but the parameters are unknown and estimated from the data. We apply the parametric bootstrap method to find a confidence interval (CI) that covers the true steady-state performance measure. Below, we briefly review IUQ methods most closely related to this paper. For a comprehensive review of the literature, see Barton et al. (2022).

Barton and Schruben (2001) apply bootstrapping to approximate the sampling distribution of the estimated input models and run simulations at each bootstrap sample to estimate the mean, then construct a quantile-based bootstrap CI from the sample means. However, the resulting CI suffers from overcoverage as the simulation error of the bootstrap sample mean inflates the CI width. Brute-forcely reducing overcoverage by increasing the number of replications run at each bootstrap sample can be too computationally demanding if the simulation runtime in nonnegligible. To improve simulation efficiency of the quantile-based bootstrap CI, Glynn and Lam (2018) apply sectioning, Barton et al. (2018) consider bootstrap shrinkage methods while Lam and Qian (2022) propose subsampling.

To the best of our knowledge, the IUQ literature focuses exclusively on unbiased simulation models. However, in steady-state simulation, the simulation output often has initial bias due to finite run time even if a warm-up period is implemented. Our goal in this paper is to *directly tackle the issue of the initial bias of a steady-state simulation in IUQ experiment design* when the simulation model has a regenerative structure and thus the steady-state output mean can be written as a ratio estimator (Henderson and Glynn 2001). In this case, the initial bias of the steady-state performance measure estimator manifests as the bias in the ratio estimator in which the expected values in the numerator and denominator are replaced with the respective sample means (Glynn 2006). To reduce the bias and improve simulation efficiency, we first divide up a single replication run at each bootstrapped parameter into renewal cycles. Then, we construct the numberator and the denominator of the ratio estimator at each bootstrapped parameter by pooling the sample paths generated within the renewal cycles from its k nearest neighboring parameters. Pooling the neighbors increases the biases in the numerator and denominator, however, can significantly reduce their variances, which in turn reduces the bias in the ratio estimator. By analyzing the asymptotic convergence rate of the k-nearest neighbor (kNN) ratio estimator as the real-world data size increases, we propose the asymptotically optimal choice for k as well as the bootstrap sample size.

However, the kNN estimation does not scale well in the parameter dimension and may be less efficient than the nominal estimation for a high-dimensional case. To address this issue, we propose a second ratio estimator that combines the kNN estimator with the likelihood ratio method to reduce the effect of the dimensionality. We empirically compare the two proposed estimators' performances against the standard ratio estimator to demonstrate their finite sample efficiencies.

The remainder of the paper is organized as follows. We mathematically formulate the IUQ problem for a regenerative simulation model in Section 2 and propose two estimators in Section 3. We analyze the the mean square error (MSE) of the estimators and establish a central limit theorem in Section 4. Section 5 studies the empirical performance of the estimators.

2 PROBLEM STATEMENT

Let $\theta^c = (\vartheta_1^c, \ldots, \vartheta_L^c) \in \mathbb{R}^d$ denote the unknown true parameter vector of the $L \geq 1$ independent input models. Within a simulation run, each input model generates an independent and identically distributed (i.i.d.) random variates. This definition can be applied even if there are correlated inputs as long as they are generated as vectors and the correlation structure can be parameterized.

For $l = 1, \ldots, L$, let m_l be the number of observations collected from the lth model and $\widehat{\theta} = (\widehat{\vartheta}_1, \ldots, \widehat{\vartheta}_L)$ be the maximum likelihood estimator (MLE) of θ^c. We assume that m_l/m converges to a nonzero constant for each l, where $m = \sum_{l=1}^{L} m_l$. Then, under some regularity conditions (Van der Vaart 2000), $\mathbb{E}[\|\vartheta_l^c - \widehat{\vartheta}_l\|^2] = \mathscr{O}\left(m_l^{-1}\right)$ and thus $\mathbb{E}[\|\theta^c - \widehat{\theta}\|^2] = \mathscr{O}\left(m^{-1}\right)$, where $\|\cdot\|$ denotes the Euclidean norm. We further assume that each $\widehat{\vartheta}_l$ has a continuous sampling distribution whose probability density function (pdf) is $\tilde{f}_l(\widehat{\vartheta}_l|\vartheta_l^c)$. Thus, $\tilde{f}(\widehat{\theta}|\theta^c) = \prod_{l=1}^{L} \tilde{f}_l(\widehat{\vartheta}_l|\vartheta_l^c)$ represents the sampling distribution of $\widehat{\theta}$ and is defined on the support, $\tilde{\Theta} \subset \mathbb{R}^d$. Without loss of generality, we assume $L = 1$ in the remainder of the paper.

We focus on the *regenerative simulation* in which the simulated system's state periodically returns to a regenerative state. The *regenerative cycle* is defined as the period between two consecutive returns. As its name suggests, a regenerative simulator "restarts" at the beginning of each cycle and progresses independently of the past. Taking a Markovian queueing model as an example, the regenerative state can be selected as the point when all servers become idle. Once the system reaches the regenerative state, new arrivals and services occur until all servers become idle again, completing a regenerative cycle.

Given a generic input parameter vector θ, let $X(\theta, t)$ denote the simulation sample path. In the queueing example, θ includes the parameters of the inter-arrival and service time distributions. Under our assumption, $X(\theta, t)$ is a regenerative stochastic process defined on state space $\mathbb{S}$ at time $t \geq 0$ and $w : \mathbb{S} \to \mathbb{R}$

is a real-valued reward function. The long-run reward rate η is defined as

$$\eta(\theta) \triangleq \lim_{t\to\infty} \frac{1}{t}\int_0^t w(X(\theta,s))\mathrm{d}s,$$

if the limit exists. Let $\mathbf{Z}_j(\theta)$ denote the set of i.i.d. input random vectors generated from the input models parameterized by θ in the jth regenerative cycle. Namely, $\mathbf{Z}_j(\theta) = \{Z_{j,1}(\theta), Z_{j,2}(\theta), \ldots, Z_{j,S_j(\theta)}(\theta)\}$, where $Z_{j,\ell}(\theta)$ is the ℓth input vector and $S_j(\theta)$ is the number of input vectors generated within the cycle. In particular, $S_j(\theta)$ is random and i.i.d. across j conditional on θ. Similarly, $\mathbf{Z}_j(\theta), j = 1,2,\ldots,$ are also i.i.d. given θ. Let $Y_j(\theta) = Y(\mathbf{Z}_j(\theta)) \in \mathbb{R}$ and $A_j(\theta) = A(\mathbf{Z}_j(\theta)) \in \mathbb{R}$ denote the cumulative reward and the length of the jth cycle, respectively. The paired sequence $\{(Y_j(\theta), A_j(\theta))\}_{j\geq 1}$ is i.i.d. conditional on θ. Then, the renewal reward theorem (Ross 1995) stipulates that

$$\eta(\theta) = \frac{\mathbb{E}[Y(\theta)|\theta]}{\mathbb{E}[A(\theta)|\theta]}, \tag{1}$$

where $\mathbb{E}[\cdot|\theta]$ denotes the expectation taken with respect to the inner-level simulation error run with input parameter θ. For instance, $\mathbf{Z}_j(\theta), Y_j(\theta)$, and $A_j(\theta)$ may represent the set of service and inter-arrival times, the integrated number of jobs in the system over time, and the regenerative cycle length, respectively, within the jth cycle. Then, $\eta(\theta)$ is the average number in system.

The standard regenerative simulation estimator of $\eta(\theta)$ replaces the expectations in (1) with their respective sample averages (Glynn 2006):

$$\widehat{\eta}_{std}(\theta) = \frac{\sum_{j=1}^r Y_j(\theta)}{\sum_{j=1}^r A_j(\theta)}, \tag{2}$$

where r denotes the number of regenerative cycles run at θ.

One way to quantify input uncertainty in simulation output is to construct a CI for $\eta(\theta^c)$ that incorporates the sampling error of $\widehat{\theta}$ as well as the simulation error. For exposition, suppose the distribution of $\eta(\widehat{\theta}) - \eta(\theta^c)$ is known and let $q_\alpha(\cdot)$ denote the α-quantile function of a random variable. Then, we expect the following CI for $\eta(\theta^c)$ to have $1-\alpha$ coverage as r tends to infinity:

$$[\widehat{\eta}_{std}(\widehat{\theta}) - q_{1-\alpha/2}(\eta(\widehat{\theta}) - \eta(\theta^c)), \widehat{\eta}_{std}(\widehat{\theta}) - q_{\alpha/2}(\eta(\widehat{\theta}) - \eta(\theta^c))]. \tag{3}$$

When $\eta(\cdot)$ is estimated by $\widehat{\eta}_{std}(\cdot)$, the distribution of $\widehat{\eta}_{std}(\widehat{\theta}) - \widehat{\eta}_{std}(\theta^c)$ is still unknown since θ^c and the sampling distribution of $\widehat{\theta}$, $\tilde{f}$, are unknown. In the IUQ literature, $\tilde{f}$ is typically approximated by the asymptotic distribution of $\widehat{\theta}$ as $m \to \infty$ or by bootstrapping the input data. In this work, we adopt the parametric bootstrap method, where each bootstrapped parameter is an MLE computed from m random inputs generated from the parametric input model given $\widehat{\theta}$. Since we assume that the input distribution family is known, this makes the bootstrapped parameter have the pdf, $\tilde{f}(\cdot|\widehat{\theta})$. Below, we describe a simulation experiment design that adopts bootstrap to approximate (3).

Suppose we first bootstrap size-n parameter set $\{\theta_1, \ldots, \theta_n\}$ using $\widehat{\theta}$. By running r regenerative cycles at each θ_i, we can obtain the ratio estimators $\widehat{\eta}_{std}(\theta_i)$ for all $1 \leq i \leq n$. Throughout the paper, we denote the empirical α-quantile computed from a size-n sample by $\widehat{q}_{\alpha,n}(\cdot)$. For instance, $\widehat{q}_{\alpha,n}(\widehat{\eta}_{std}(\theta) - \widehat{\eta}_{std}(\widehat{\theta}))$ denotes the empirical α-quantile of $\widehat{\eta}_{std}(\theta_1) - \widehat{\eta}_{std}(\widehat{\theta}), \widehat{\eta}_{std}(\theta_2) - \widehat{\eta}_{std}(\widehat{\theta}), \ldots, \widehat{\eta}_{std}(\theta_n) - \widehat{\eta}_{std}(\widehat{\theta})$. Then, (3) can be approximated by

$$[\widehat{\eta}_{std}(\widehat{\theta}) - \widehat{q}_{1-\alpha/2,n}(\widehat{\eta}_{std}(\theta) - \widehat{\eta}_{std}(\widehat{\theta})), \widehat{\eta}_{std}(\widehat{\theta}) - \widehat{q}_{\alpha/2,n}(\widehat{\eta}_{std}(\theta) - \widehat{\eta}_{std}(\widehat{\theta}))]. \tag{4}$$

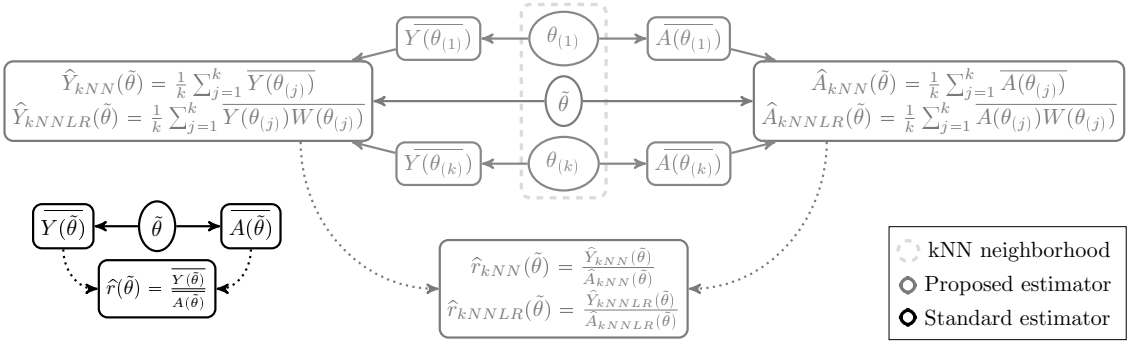

Figure 1: Schematic illustration of standard ratio estimator and the proposed kNN-based estimator.

In the literature, (4) is referred to as the basic bootstrap CI (Barton et al. 2018). If we assume there exists monotonic transformation ρ that makes the distribution of $\rho(\widehat{\eta}_{std}(\theta) - \widehat{\eta}_{std}(\widehat{\theta}))$ symmetric around 0, then (4) can be converted to the following percentile bootstrap CI (Davison and Hinkley 1997)

$$[\widehat{q}_{\alpha/2,n}(\widehat{\eta}_{std}(\theta)), \widehat{q}_{1-\alpha/2,n}(\widehat{\eta}_{std}(\theta))]. \tag{5}$$

There is some empirical evidence in the literature that (5) outperforms (4) for finite n (Barton et al. 2018), thus we adopt (5) in this paper.

3 PROPOSED ESTIMATORS

The standard estimator in (2), $\widehat{\eta}_{std}(\theta)$, is computed from the regenerative cycles run at θ only. In this section, we propose two ratio estimators based on the k nearest neighbors (kNN) method to enhance computational efficiency of $\widehat{\eta}_{std}(\theta)$. We begin by describing the new experiment design in the following.

We bootstrap size-$\tilde{n}$ parameter set $\{\tilde{\theta}_i\}_{1\le i\le\tilde{n}}$, where $\tilde{\theta}_i \overset{i.i.d.}{\sim} \tilde{f}(\tilde{\theta}|\widehat{\theta})$, however, we do not run regenerative simulations at these parameters. Instead, we generate a second size-n set of parameters $\{\theta_j\}_{1\le j\le n}$ to run the simulations at. The two sample sizes, n and $\tilde{n}$, can differ. To distinguish the two sets, we coin the terms, *bootstrap parameter set* and *simulation parameter set*, to refer to $\{\tilde{\theta}_i\}_{1\le i\le\tilde{n}}$ and $\{\theta_j\}_{1\le j\le n}$, respectively.

Let $f(\theta|\widehat{\theta})$ be the sampling pdf of the simulation parameters, which may or may not be the same as the bootstrap pdf, $\tilde{f}(\tilde{\theta}|\widehat{\theta})$. Moreover, f can depend on $\{\tilde{\theta}_i\}_{1\le i\le\tilde{n}}$, which we further elaborate with an example in Section 5. For simplicity, we keep the notation, $f(\theta|\widehat{\theta})$, throughout the paper, while it is easy to extend the results if f depends on $\{\tilde{\theta}_i\}_{1\le i\le\tilde{n}}$. Let Θ denote the support of $f(\theta|\widehat{\theta})$. We require f to be chosen such that all $\theta\in\Theta$ are feasible parameters for the simulation input model. Once we collect the simulation inputs generated with $\{\theta_j\}_{1\le j\le n}$, we adopt a pooling method to compute a point estimator of $\eta(\tilde{\theta}_i)$, for each $1\le i\le\tilde{n}$, and then construct the empirical quantile estimators and CI from the estimators.

Figure 1 illustrates how the two proposed ratio estimators are constructed by pooling simulation outputs generated at $\{\theta_i\}_{1\le i\le n}$. For any $\tilde{\theta}\in\tilde{\Theta}$, we first propose the following kNN estimator

$$\widehat{\eta}_{kNN}(\tilde{\theta}) \triangleq \frac{\widehat{Y}_{kNN}(\tilde{\theta})}{\widehat{A}_{kNN}(\tilde{\theta})} = \frac{\frac{1}{k}\sum_{i=1}^{k}\overline{Y\left(\theta_{(i)}\right)}}{\frac{1}{k}\sum_{i=1}^{k}\overline{A\left(\theta_{(i)}\right)}} = \frac{\frac{1}{k}\sum_{i=1}^{k}\frac{1}{r}\sum_{j=1}^{r}Y_j\left(\theta_{(i)}\right)}{\frac{1}{k}\sum_{i=1}^{k}\frac{1}{r}\sum_{j=1}^{r}A_j\left(\theta_{(i)}\right)}, \tag{6}$$

where $\theta_{(i)}$ denotes $\tilde{\theta}$'s ith nearest neighbor among $\{\theta_j\}_{1\le j\le n}$. As illustrated in Figure 1, for each $\theta_{(i)}$, we compute $\overline{Y\left(\theta_{(i)}\right)}$ and $\overline{A\left(\theta_{(i)}\right)}$ then take their averages across the k nearest neighbors of $\tilde{\theta}$ to obtain $\widehat{Y}_{kNN}(\tilde{\theta})$ and $\widehat{A}_{kNN}(\tilde{\theta})$, respectively. Namely, we train two kNN regression models on (θ,Y) and (θ,A) separately, and $\widehat{\eta}_{kNN}(\tilde{\theta})$ is the ratio of the two models evaluated at $\tilde{\theta}$. Note that $\overline{Y\left(\theta_{(i)}\right)}$ and $\overline{A\left(\theta_{(i)}\right)}$ are correlated as they are computed from the same simulation sample path. This correlation is considered in our analysis.

Compared to the standard estimator (2), the pooled estimator (6) is expected to reduce the variances of the numerator and denominator at the expense of additional biases in them, i.e., $\frac{1}{k}\sum_{i=1}^{k}\overline{Y(\theta_{(i)})}$ is no longer an unbiased estimator of $\mathbb{E}[Y|\tilde{\theta}]$.

The second estimator is proposed to reduce the extra bias introduced by the kNN method by combining it with the likelihood ratio (LR) method. Recall that $\mathbf{Z}_j(\theta) = \{Z_{j,1}(\theta), Z_{j,2}(\theta), \ldots, Z_{j,S(\theta)}(\theta)\}$ is the set of simulation inputs generated from the input model with parameter θ within the jth regenerative cycle. Henceforth, when no confusion arises we adopt the short-hand notation $Z_{j,\ell} = Z_{j,\ell}(\theta)$ for convenience. Let $p(Z|\theta)$ be the pdf of the input model, i.e., $Z_{j,\ell} \overset{\text{i.i.d.}}{\sim} p(Z|\theta)$ so the joint likelihood of $\mathbf{Z}_j(\theta)$ is $\prod_{\ell=1}^{S(\theta)} p(Z_{j,\ell}|\theta)$. Then, the sample likelihood ratio between θ and $\tilde{\theta}$ for the jth regenerative cycle is defined as $W_j\left(\theta;\tilde{\theta}\right) \triangleq \frac{\prod_{\ell=1}^{S(\theta)} p(Z_{j,\ell}|\tilde{\theta}))}{\prod_{\ell=1}^{S(\theta)} p(Z_{j,\ell}|\theta)}$. Under mild conditions (e.g., absolute continuity of $p(\cdot|\tilde{\theta})$ with respect to $p(\cdot|\theta)$), we have

$$\mathbb{E}\left[Y_j(\theta)W_j\left(\theta;\tilde{\theta}\right)\middle|\,\theta\right] = \int y(\mathbf{Z}_j)\frac{\prod_{\ell=1}^{S(\theta)} p(Z_{j,\ell}|\tilde{\theta})}{\prod_{\ell=1}^{S(\theta)} p(Z_{j,\ell}|\theta)}\left(\prod_{\ell=1}^{S(\theta)} p(Z_{j,\ell}|\theta)\right)\mathrm{d}\mathbf{Z}_j = \mathbb{E}\left[Y|\tilde{\theta}\right]. \tag{7}$$

Combining (6) with the LR method, we propose the kNN LR ratio estimator

$$\widehat{\eta}_{kNNLR}(\tilde{\theta}) \triangleq \frac{\widehat{Y}_{kNNLR}(\tilde{\theta})}{\widehat{A}_{kNNLR}(\tilde{\theta})} = \frac{\frac{1}{k}\sum_{i=1}^{k}\overline{Y\left(\theta_{(i)}\right)W\left(\theta_{(i)}\right)}}{\frac{1}{k}\sum_{i=1}^{k}\overline{A\left(\theta_{(i)}\right)W_j\left(\theta_{(i)}\right)}} = \frac{\frac{1}{k}\sum_{i=1}^{k}\frac{1}{r}\sum_{j=1}^{r}Y_j\left(\theta_{(i)}\right)W_j\left(\theta_{(i)};\tilde{\theta}\right)}{\frac{1}{k}\sum_{i=1}^{k}\frac{1}{r}\sum_{j=1}^{r}A_j\left(\theta_{(i)}\right)W_j\left(\theta_{(i)};\tilde{\theta}\right)}. \tag{8}$$

While the LR method alleviates the bias introduced by the kNN pooling, it has a drawback: $\frac{1}{r}\sum_{j=1}^{r}Y_j\left(\theta_{(i)}\right)W_j\left(\theta_{(i)};\tilde{\theta}\right)$ may have a large or even infinite variance when $\theta_{(i)}$ and $\tilde{\theta}$ significantly differ. For (8), however, this is somewhat regulated by that we only pool the observations at the k-nearest neighbors of $\tilde{\theta}$. The exact effect of the choice of k to the variance of (8) requires further analyses.

4 ASYMPTOTIC ANALYSIS

In this section, we examine the asymptotic properties of the kNN and kNN LR estimators. Section 4.1 establishes the mean squared error (MSE) of the kNN estimator, $\widehat{\eta}_{kNN}(\tilde{\theta})$, and a Central Limit Theorem (CLT) for it. Furthermore, we examine the convergence of the empirical quantile estimator constructed from $\widehat{\eta}_{kNN}(\tilde{\theta})$. Section 4.2 shows the bias and MSE of the kNN LR estimator, $\widehat{\eta}_{kNNLR}(\tilde{\theta})$, under a special case where the input distribution belongs in the exponential family. All proofs of the theoretical results in this section are omitted due to the space limit and will be made available in an online archive version.

In the following, we state two assumptions on f and the distributions of Y and A required for later theoretical developments.

Assumption 1 Given $\widehat{\theta}$, the following statements hold: (i) $\{\tilde{\theta}_i\}_{1\le i\le \tilde{n}} \subset \Theta$. (ii) For any $\theta \in \Theta$, $f(\theta|\widehat{\theta})$ is continuous and bounded for all $\theta \in \Theta$. (iii) For any $\theta \in \Theta$, $\mathbb{P}(\|\theta\| > t|\widehat{\theta}) = \mathscr{O}(t^{-\gamma})$ for some $\gamma > 0$ as $t \to \infty$.

Assumption 1(i) guarantees that for any $\tilde{\theta}$, $\inf_{1\le j\le n}\|\tilde{\theta} - \theta_j\| \to 0$ almost surely as n tends to infinity, which forms the fundamental basis of the kNN regression technique. Assumption 1(ii) makes f locally Lipschitz continuous. Assumption 1(iii) accommodates the case when Θ is unbounded, but also holds when Θ is bounded.

Next, let $g_Y(\theta,y|\widehat{\theta})$ and $g_A(\theta,a|\widehat{\theta})$ be the joint pdfs of (θ,Y) and (θ,A) conditional on $\widehat{\theta}$, respectively. Assumption 2 stipulates some differentiability conditions of moments of Y and A with respect to θ:

Assumption 2 Given $\widehat{\theta}$, for any $\theta \in \Theta$: (i) $\int y g_Y(\theta,y|\widehat{\theta})\mathrm{d}y$ and $\int a g_A(\theta,a|\widehat{\theta})\mathrm{d}a$ are bounded and twice differentiable in θ. (ii) $\int y^2 g_Y(\theta,y|\widehat{\theta})\mathrm{d}y$ and $\int a^2 g_A(\theta,a|\widehat{\theta})\mathrm{d}a$ are bounded and twice differentiable in θ.

Note that $\int y g_Y(\theta,y|\widehat{\theta})\mathrm{d}y = \int y g_{Y|\theta}(y|\theta) f(\theta|\widehat{\theta})\mathrm{d}y = \mathbb{E}[Y|\theta]f(\theta|\widehat{\theta})$. Therefore, Assumption 1(ii) and Assumption 2 together imply that the first two conditional moments of Y are bounded in the neighborhood of θ. The same implication holds for A.

Throughout the paper, we adopt the following notations to describe the limiting behavior of sequences: for positive sequences $\{a_n\}$ and $\{b_n\} \subset \mathbb{R}$, write $a_n = \mathscr{O}(b_n)$ if there exists constant $\overline{c} > 0$ such that $a_n \leq \overline{c} b_n$ holds for all $n \geq 1$; $a_n = o(b_n)$ if $\frac{a_n}{b_n} \to 0$ as $n \to \infty$; and $a_n = \mathscr{O}_{\mathbb{P}}(b_n)$ if for any $\varepsilon > 0$, there exists M and N such that $\mathbb{P}(|a_n/b_n| < M) > 1 - \varepsilon$ for all $n > N$.

4.1 Analysis on the kNN ratio estimator

The following Taylor expansion on $\frac{b}{a}$ is repeatedly used in our analyses: for $a \neq 0$ and $a + \Delta a \neq 0$,

$$\frac{b+\Delta b}{a+\Delta a} - \frac{b}{a} = -\frac{b\Delta a}{a^2} + \frac{\Delta b}{a} - \frac{\Delta a \Delta b}{a^2} + \frac{b}{a^3}(\Delta a)^2 + o((\Delta a)^2). \tag{9}$$

Let us first fix $\tilde{\theta}$ to be an arbitrary point in $\tilde{\Theta}$. Define $d(Y) = \widehat{Y}_{kNN} - \mathbb{E}[Y|\tilde{\theta}]$ and $d(A) = \widehat{A}_{kNN} - \mathbb{E}[A|\tilde{\theta}]$. Then, we have $\widehat{\eta}_{kNN}(\tilde{\theta}) = \frac{\widehat{Y}_{kNN}}{\widehat{A}_{kNN}} = \frac{\mathbb{E}[Y|\tilde{\theta}]+d(Y)}{\mathbb{E}[A|\tilde{\theta}]+d(A)}$. From (9),

$$\widehat{\eta}_{kNN}(\tilde{\theta}) - \eta(\tilde{\theta}) = -\frac{\mathbb{E}[Y|\tilde{\theta}]d(A)}{\left(\mathbb{E}[A|\tilde{\theta}]\right)^2} + \frac{d(Y)}{\mathbb{E}[A|\tilde{\theta}]} - \frac{d(A)d(Y)}{\left(\mathbb{E}[A|\tilde{\theta}]\right)^2} + \frac{\mathbb{E}[Y|\tilde{\theta}]}{\left(\mathbb{E}[A|\tilde{\theta}]\right)^3}(d(A))^2 + o((d(A))^2). \tag{10}$$

To derive the mean squared error (MSE) of $\widehat{\eta}_{kNN}$, we analyze the moments of $d(A)$ and $d(Y)$. Let $R_{n,k+1}$ be the $(k+1)$th nearest neighbor's distance from $\tilde{\theta}$, i.e., $R_{n,k+1} = \|\tilde{\theta} - \theta_{(k+1)}\|$. Conditional on $R_{n,k+1}$, the k nearest neighbors $\theta_{(1)}, \ldots, \theta_{(k)}$ are i.i.d. (Mack and Rosenblatt 1979). Further, let V_d denote the volume of the unit ball in $\mathbb{R}^d$ and $\widehat{f}_{n,k}(\tilde{\theta}|\widehat{\theta}) \triangleq \frac{k}{nV_d R_{n,k+1}^d}$. Note that the latter is a kNN density estimator of $f(\theta|\widehat{\theta})$. The random variable, $R_{n,k+1}$, is closely related to the Beta distribution and from this relationship, one can show that $\mathbb{E}\left[\widehat{f}_{n,k}(\tilde{\theta}|\widehat{\theta})/f(\tilde{\theta}|\widehat{\theta}) \middle| \tilde{\theta}, \widehat{\theta}\right] \to 1$ when Assumption 3(i) in the following holds. Additionally, if Assumption 3(ii) holds, $\widehat{f}_{n,k}(\tilde{\theta}|\widehat{\theta})$ is strongly uniformly consistent to $f(\tilde{\theta}|\widehat{\theta})$ (Devroye and Wagner 1977).

Assumption 3 The values of k and n satisfy (i) $k \to \infty$ and $\frac{k}{n} \to 0$ as $n \to \infty$; and (ii) $\frac{\log(n)}{k} \to 0$ as $n \to \infty$.

Consider the following scaled kNN estimator of $\mathbb{E}[Y|\tilde{\theta}]$:

$$\widehat{Y}^s_{kNN}(\tilde{\theta}) := \frac{\widehat{f}_{n,k}(\tilde{\theta}|\widehat{\theta})}{f(\tilde{\theta}|\widehat{\theta})}\widehat{Y}_{kNN}(\tilde{\theta}) = \frac{1}{nV_d R_{n,k+1}^d f(\tilde{\theta}|\widehat{\theta})} \sum_{i=1}^{n} \mathbb{1}\left\{\frac{\|\tilde{\theta} - \theta_i\|}{R_{n,k+1}} < 1\right\} \overline{Y(\theta_i)}, \tag{11}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. From the definition of $R_{n,k+1}$, there are k nonzero indicators in the sum in (11). We also define $\widehat{A}^s_{kNN}(\tilde{\theta}) := \frac{\widehat{f}_{n,k}(\tilde{\theta}|\widehat{\theta})}{f(\tilde{\theta}|\widehat{\theta})}\widehat{A}_{kNN}(\tilde{\theta})$. While the ratio of the scaled estimators remains unchanged, i.e., $\frac{\widehat{Y}^s_{kNN}(\tilde{\theta})}{\widehat{A}^s_{kNN}(\tilde{\theta})} = \frac{\widehat{Y}_{kNN}(\tilde{\theta})}{\widehat{A}_{kNN}(\tilde{\theta})} = \widehat{\eta}_{kNN}(\tilde{\theta})$, the scaling turns out to be useful in the subsequent analysis.

Define $(yf)(\theta) \triangleq \mathbb{E}[Y|\theta]f(\theta|\widehat{\theta})$ and $(af)(\theta) \triangleq \mathbb{E}[A|\theta]f(\theta|\widehat{\theta})$. Also, for any twice differentiable function $\psi(x)$, define $\Delta\psi(x) \triangleq \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}\psi(x)$. Inspired by Propositions 3 and 4 in Mack (1981), we establish the following convergnece results for the biases and variances of the scaled estimators.

Lemma 1 Suppose Assumptions 1 (i) (ii), 2 and 3 (i) hold. Then, conditional on $\widehat{\theta}$, for any $\tilde{\theta} \in \tilde{\Theta}$,

$$\mathbb{E}\left[\widehat{Y}^s_{kNN}(\tilde{\theta}) \middle| \tilde{\theta}, \widehat{\theta}\right] - \mathbb{E}[Y|\tilde{\theta}] = \frac{\Delta(yf)(\tilde{\theta})}{2(d+2)V_d^{\frac{2}{d}}\left(f(\tilde{\theta}|\widehat{\theta})\right)^{1+\frac{2}{d}}}\left(\frac{k}{n}\right)^{\frac{2}{d}} + o\left(\left(\frac{k}{n}\right)^{\frac{2}{d}}\right), \text{ and} \tag{12}$$

$$\mathrm{Var}\left[\widehat{Y}^s_{kNN}(\tilde{\theta})\middle|\tilde{\theta},\widehat{\theta}\right] \leq \frac{\mathrm{Var}[Y|\tilde{\theta}]}{rk} + \frac{2\left(\mathbb{E}[Y|\tilde{\theta}]\right)^2}{k} + \frac{1}{2(d+2)^2}\frac{\left(\Delta(yf)(\tilde{\theta})\right)^2}{V_d^{\frac{d}{4}}(f(\tilde{\theta}|\widehat{\theta}))^{2+\frac{d}{4}}}\left(\frac{k}{n}\right)^{\frac{4}{d}} + o\left(\left(\frac{k}{n}\right)^{\frac{4}{d}}\right) + o\left(\frac{1}{k}\right) + o\left(\frac{1}{rk}\right). \tag{13}$$

Similar statements can be made for $\mathbb{E}\left[\widehat{A}^s_{kNN}(\tilde{\theta})\middle|\tilde{\theta},\widehat{\theta}\right] - \mathbb{E}[A|\tilde{\theta}]$ and $\mathrm{Var}\left[\widehat{A}^s_{kNN}(\tilde{\theta})\middle|\tilde{\theta},\widehat{\theta}\right]$.

Note that $f(\tilde{\theta}|\widehat{\theta})$ is in the denominator of the dominant term in (12). This means that, all else equal, when $f(\tilde{\theta}|\widehat{\theta})$ is small so that the probability of sampling θ close to $\tilde{\theta}$ is low, the resulting estimator has a larger bias. A similar observation can be made for the third term of (13). Recall that we allow the choice of f to be flexible and not necessarily equal to the bootstrap distribution, $\tilde{f}(\cdot|\widehat{\theta})$. Thus, for fixed $\tilde{\theta}$ we can choose f to reduce the bias.

Without the scaling factor $\frac{\widehat{f}_{n,k}(\tilde{\theta}|\widehat{\theta})}{f(\tilde{\theta}|\widehat{\theta})}$, the bias of $\widehat{Y}_{kNN}(\tilde{\theta})$ is bounded as $\mathbb{E}[\widehat{Y}_{kNN}(\tilde{\theta})|\tilde{\theta},\widehat{\theta}] - \mathbb{E}[Y|\tilde{\theta}] = \mathscr{O}(\left(\frac{k}{n}\right)^{\frac{1}{d}})$. The same bound holds for the bias of $\widehat{A}_{kNN}(\tilde{\theta})$. If we additionally adopt Assumption 1 (iii), the variance (13) can be shown to have convergence rate $\mathscr{O}(1/k) + \mathscr{O}(1/(rk))$ (Mack 1981).

Combining Lemma 1 and the Taylor expansion (10), Proposition 1 derives the expression for the bias and MSE of $\widehat{\eta}_{kNN}(\tilde{\theta})$.

Proposition 1 Suppose Assumptions 1 (i) (ii), 2 and 3 (i) hold. Then, conditional on $\widehat{\theta}$, for any $\tilde{\theta} \in \tilde{\Theta}$,

$$\mathbb{E}[\widehat{\eta}_{kNN}(\tilde{\theta})\big|\tilde{\theta},\widehat{\theta}] - \eta(\tilde{\theta}) = \frac{\Delta(yf)(\tilde{\theta}) - \eta(\tilde{\theta})\Delta(af)(\tilde{\theta})}{2(d+2)\mathbb{E}[A|\tilde{\theta}]V_d^{\frac{2}{d}}(f(\tilde{\theta}|\widehat{\theta}))^{1+\frac{2}{d}}}\left(\frac{k}{n}\right)^{\frac{2}{d}} + o\left(\left(\frac{k}{n}\right)^{\frac{2}{d}}\right), \tag{14}$$

$$\mathbb{E}\left[\left(\widehat{\eta}_{kNN}(\tilde{\theta}) - \eta(\tilde{\theta})\right)^2\middle|\tilde{\theta},\widehat{\theta}\right] = \mathscr{O}\left(\left(\frac{k}{n}\right)^{\frac{4}{d}}\right) + \mathscr{O}\left(\frac{1}{rk}\right) + \mathscr{O}\left(\frac{1}{k}\right). \tag{15}$$

Theorem 1 establishes a CLT for $\widehat{\eta}_{kNN}(\tilde{\theta})$; note that $\xrightarrow{\mathscr{D}}$ denotes convergence in distribution.

Theorem 1 Suppose Assumptions 1, 2 and 3 hold. Additionally, for any $\tilde{\theta} \in \tilde{\Theta}$, suppose that $\mathbb{E}\left[\left|\overline{Y(\tilde{\theta})} - \eta(\tilde{\theta})\overline{A(\tilde{\theta})}\right|^3\middle|\tilde{\theta},\widehat{\theta}\right] < \infty$ and $\mathrm{Var}\left[\overline{Y(\tilde{\theta})} - \eta(\tilde{\theta})\overline{A(\tilde{\theta})}\middle|\tilde{\theta},\widehat{\theta}\right] > 0$. Let $k = o\left(n^{\frac{2}{2+d}}\right)$ and r be a constant, then, conditional on both $\widehat{\theta}$ and $\tilde{\theta}$,

$$\sqrt{rk}\left(\widehat{\eta}_{kNN}(\tilde{\theta}) - \eta(\tilde{\theta})\right) \xrightarrow{\mathscr{D}} \mathscr{N}\left(0, V_d^{-1}\mathrm{Var}\left[Y(\tilde{\theta}) - \eta(\tilde{\theta})A(\tilde{\theta})\middle|\tilde{\theta}\right]\right).$$

Next, we proceed to show that the empirical α-quantile of $\widehat{\eta}_{kNN}(\tilde{\theta}_1),\ldots,\widehat{\eta}_{kNN}(\tilde{\theta}_{\tilde{n}})$, that is, $\widehat{q}_{\alpha,n}(\widehat{\eta}_{kNN}(\tilde{\theta}_i))$, converges to $q_\alpha(\eta(\tilde{\theta}))$ where $\tilde{\theta} \sim \tilde{f}(\tilde{\theta}|\widehat{\theta})$. Let $\Phi(x) = \mathbb{P}(\eta(\tilde{\theta}) \leq x)$ and $\phi(x)$ be the cdf and the pdf of $\eta(\tilde{\theta})$, respectively. Recall that in our design we generate $\tilde{\theta}_i \overset{i.i.d.}{\sim} \tilde{f}(\tilde{\theta}|\widehat{\theta})$, for $i = 1,\ldots,\tilde{n}$. Then, the ecdf of $\widehat{\eta}_{kNN}(\tilde{\theta})$ is defined as $\Phi_{\tilde{n},r}(x) = \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\mathbb{1}\{\widehat{\eta}_{kNN}(\tilde{\theta}_i) \leq x\}$. This is not a typical ecdf constructed from i.i.d. observations as $\widehat{\eta}_{kNN}(\tilde{\theta}_1),\ldots,\widehat{\eta}_{kNN}(\tilde{\theta}_{\tilde{n}})$ are correlated. Nevertheless, we show that $\Phi_{\tilde{n},r}(x)$ is a consistent estimator of $\Phi(x)$ below.

Let us define the scaled simulation errors $\varepsilon_i = \sqrt{rk}(\widehat{\eta}_{kNN}(\tilde{\theta}_i) - \eta(\tilde{\theta}_i))$ and denote the conditional joint distribution of the pair$(\eta(\tilde{\theta}_i),\varepsilon_i)$ by $h_i(\eta,\varepsilon|\widehat{\theta})$. Moreover, let $h_{i,j}(\eta_i,\eta_j,\varepsilon_i,\varepsilon_j|\widehat{\theta})$ represent the conditional joint distribution of $(\eta(\tilde{\theta}_i),\eta(\tilde{\theta}_j),\varepsilon_i,\varepsilon_j)$. We first make the following assumption to show the consistency result for $\Phi_{\tilde{n},r}(x)$.

Assumption 4 The following conditions hold:

(i) $\Phi(x)$ is absolutely continuous with continuous pdf $\phi(x)$ and $f(\cdot|\widehat{\theta})$ is bounded away from zero on Θ.

(ii) For any $\tilde{\theta} \in \tilde{\Theta}$ and any $1 \le i \le \tilde{n}$, $h_i(\eta, \varepsilon|\widehat{\theta})$ is differentiable with respect to η. There exists $p_{0,n,r}(\varepsilon) > 0$ and $p_{1,n,r}(\varepsilon) > 0$ such that $h_i(\eta, \varepsilon|\widehat{\theta}) \le p_{0,n,r}(\varepsilon)$ and $\left|\frac{\partial}{\partial \eta} h_i(\eta, \varepsilon|\widehat{\theta})\right| \le p_{1,n,r}(\varepsilon)$. Moreover, $\sup_n \sup_r \int_{-\infty}^{\infty} |\varepsilon|^q p_{l,n,r}(\varepsilon) \mathrm{d}\varepsilon < \infty$ for $l = 0, 1$ and $0 \le q \le 2$.

(iii) For any $\tilde{\theta} \in \tilde{\Theta}$ and any $1 \le i, j \le \tilde{n}$ with $i \ne j$, $h_{i,j}(\eta_i, \eta_j, \varepsilon_i, \varepsilon_j|\widehat{\theta})$ is differentiable with respect to both η_i and η_j. There exists $p_{0,n,r}(\varepsilon_i, \varepsilon_j) > 0$ and $p_{1,n,r}(\varepsilon_i, \varepsilon_j) > 0$ such that $h_{i,j}(\eta_i, \eta_j, \varepsilon_i, \varepsilon_j|\widehat{\theta}) \le p_{0,n,r}(\varepsilon_i, \varepsilon_j)$ and $\max\left\{\left|\frac{\partial}{\partial \eta_i} h_{i,j}(\eta_i, \eta_j, \varepsilon_i, \varepsilon_j)\right|, \left|\frac{\partial}{\partial \eta_j} h_{i,j}(\eta_i, \eta_j, \varepsilon_i, \varepsilon_j)\right|\right\} \le p_{1,n,r}(\varepsilon_i, \varepsilon_j)$. Moreover, $\sup_n \sup_r \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |\varepsilon_i|^{q_i} |\varepsilon_j|^{q_j} p_{l,n,r}(\varepsilon) \mathrm{d}\varepsilon < \infty$ for $l = 0, 1$ and $0 \le q_i, q_j \le 2$ with $q_i + q_j \le 3$.

The following lemma establishes the point-wise bias and variance of ecdf $\Phi_{\tilde{n},s}$.

Lemma 2 Suppose Assumptions 1 (i) (ii), 2, 3 and 4 hold. Then, $\mathbb{E}[\Phi_{\tilde{n},s}(x)|\widehat{\theta}] = \Phi(x) + \mathscr{O}((\frac{k}{n})^{\frac{2}{d}}) + \mathscr{O}\left(\frac{1}{rk}\right)$ and $\mathrm{Var}[\Phi_{\tilde{n},r}(x)|\widehat{\theta}] \le \mathscr{O}\left(\frac{1}{rk}\right) + \mathscr{O}\left(\frac{1}{k}\right) + \mathscr{O}\left(\frac{1}{\tilde{n}}\right) + \mathscr{O}((\frac{k}{n})^{\frac{4}{d}})$.

Lemma 3 extends Lemma 2 a Glivenko-Cantelli-type uniform weak consistency result for $\Phi_{\tilde{n},r}$.

Lemma 3 Suppose Assumptions 1 (i) (ii), 2, 3 and 4 hold. Then, conditional on $\widehat{\theta}$,

$$\sup_{x \in \mathbb{R}} |\Phi_{\tilde{n},r}(x) - \Phi(x)| = \mathscr{O}_{\mathbb{P}}(\tfrac{1}{\sqrt{rk}}) + \mathscr{O}_{\mathbb{P}}(\tfrac{1}{\sqrt{\tilde{n}}}) + \mathscr{O}_{\mathbb{P}}((\tfrac{k}{n})^{\frac{2}{d}}).$$

Finally, Proposition 2 states a weak consistency result for the proposed quantile estimator.

Proposition 2 Suppose Assumptions 1 (i) (ii), 2, 3 and 4 hold. Then conditional on $\widehat{\theta}$, $|\widehat{q}_{\alpha,n}(\widehat{\eta}_{kNN}(\tilde{\theta}_i)) - q_\alpha(\eta(\theta))| = \mathscr{O}_{\mathbb{P}}(\frac{1}{\sqrt{rk}}) + \mathscr{O}_{\mathbb{P}}(\frac{1}{\sqrt{\tilde{n}}}) + \mathscr{O}_{\mathbb{P}}((\frac{k}{n})^{\frac{2}{d}})$.

Recall that we may choose $n \ne \tilde{n}$. Since, the total simulation cost of estimating the quantile is determined by n, i.e., nr regenerative cycles, without inflating the simulation cost, $\tilde{n}$ can be chosen sufficiently large so that the estimation error of $\widehat{q}_{\alpha,n}(\widehat{\eta}_{kNN}(\tilde{\theta}_i))$ is not dominated by $\tilde{n}$.

Under conditions similar to Assumption 4, one can show that $|\widehat{q}_\alpha(\widehat{\eta}_{std}(\theta)) - q_\alpha(\eta(\theta))| = \mathscr{O}_{\mathbb{P}}(\frac{1}{\sqrt{n}}) + \mathscr{O}_{\mathbb{P}}(\frac{1}{\sqrt{r}})$. Therefore, Proposition 2 implies that when $d \ge 4$, $\widehat{q}_{\alpha,n}(\widehat{\eta}_{kNN}(\tilde{\theta}_i))$ is less efficient than $\widehat{q}_\alpha(\widehat{\eta}_{std}(\theta))$. Indeed, such a shortcoming is directly related to that the MSE convergence rate of $\widehat{\eta}_{kNN}(\tilde{\theta}_i)$ slows down for higher d as stipulated in Proposition 1.

4.2 Analysis on the *k*NN LR ratio estimator

In this subsection, we study the kNN LR estimator, as proposed in (8), which effectively reduces the bias introduced by the kNN approach (6). Recall that $\mathbf{Z}_j(\theta)$ is the set of simulation inputs generated with parameter θ within the jth cycle and the joint likelihood of $\mathbf{Z}_j(\theta)$ is $\prod_{\ell=1}^{S(\theta)} p(Z_{j,\ell}|\theta)$. We constrain our discussion to input models in the exponential family with canonical form, namely, $\prod_{\ell=1}^{S(\theta)} p(Z_{j,\ell}|\theta) = p_b(\mathbf{Z}_j) \exp(\theta^\top U(\mathbf{Z}_j) - L(\theta))$, where p_b is called the base measure, U is the sufficient statistics and L is the log-partition function. We denote $\mathrm{Int}(\Theta)$ as the interior of Θ and make the following assumptions.

Assumption 5 (i) Suppose that $\prod_{\ell=1}^{S(\theta)} p(Z_{j,\ell}|\theta)$ belongs to the exponential family and is in the canonical form. Further, suppose that for any $\tilde{\theta} \in \mathrm{Int}(\Theta)$, there exists a neighborhood $N(\tilde{\theta})$ such that for $\forall \theta \in N(\tilde{\theta})$, $\mathbb{E}[U(\mathbf{Z})|\theta] < \infty$. (ii) Suppose that for any $\tilde{\theta} \in \mathrm{Int}(\Theta)$, there exist a neighborhood $N(\tilde{\theta})$ and $\beta_Y, \beta_A > 0$ such that for $\forall \theta \in N(\tilde{\theta})$, $\mathbb{E}[Y^2|\tilde{\theta}] = \mathbb{E}[Y^2|\theta] + \mathscr{O}(\|\tilde{\theta} - \theta\|^{\beta_Y})$ and $\mathbb{E}[A^2|\tilde{\theta}] = \mathbb{E}[A^2|\theta] + \mathscr{O}(\|\tilde{\theta} - \theta\|^{\beta_A})$.

Assumption 5(i) implies that the joint pdf has bounded derivative, as it can be calculated that $\nabla_\theta \exp(L(\theta)) = \mathbb{E}[U(\mathbf{Z})|\theta]$ if we allow the exchange of integral and differential operators. Assump-

tion 5(ii) imposes a smoothness condition of $\mathbb{E}[Y^2|\tilde{\theta}]$ and $\mathbb{E}[A^2|\tilde{\theta}]$ with respect to $\tilde{\theta}$. Employed with Assumption 5, we can show the bias and MSE of $\widehat{\eta}_{kNNLR}(\tilde{\theta})$ decreases at the order of $\frac{1}{rk}$.

Lemma 4 Suppose Assumption 1 (i) (ii), 3 and 5 hold. Then for n and k sufficiently large,

$$\mathbb{E}[\widehat{\eta}_{kNNLR}(\tilde{\theta})\big|\,\tilde{\theta},\widehat{\theta}]-\eta(\tilde{\theta})=\frac{1}{rk}\frac{\eta(\tilde{\theta})\mathrm{Var}[A|\tilde{\theta}]-\mathrm{Cov}[Y,A|\tilde{\theta}]}{\mathbb{E}[A^2|\tilde{\theta}]}+o\left(\frac{1}{rk}\right),$$

$$\mathbb{E}\left[\left\|\widehat{\eta}_{kNNLR}(\tilde{\theta})-\eta(\tilde{\theta})\right\|^2\Big|\,\tilde{\theta}\right]=\frac{2}{rk}\frac{\mathrm{Var}[Y|\tilde{\theta}]+\eta^2(\tilde{\theta})\mathrm{Var}[A|\tilde{\theta}]}{\left(\mathbb{E}[A|\tilde{\theta}]\right)^2}+o\left(\frac{1}{rk}\right).$$

In contrast to Proposition 1, the incorporation of the LR method effectively eliminates the $\mathscr{O}(\left(\frac{k}{n}\right)^{\frac{4}{d}})$ term in the MSE analysis, which happens to be the square of the bias in $\widehat{\eta}_{kNN}$. However, a drawback of Lemma 4 is that it only allows for asymptotic claims when n and k are sufficiently large. This is because we cannot assume uniformity in Assumption 5. The convergence of the quantile estimator based on $\widehat{\eta}_{kNNLR}$ necessitates further investigation due to the same aforementioned reason.

5 EMPIRICAL ANALYSIS

To demonstrate the performances of the proposed estimators, we examine the empirical coverage probabilities of the percentile bootstrap CIs constructed from the standard, kNN and kNN LR estimators; and their respective widths using two $M/M/1/10$ queueing examples. The performance measure of interest is the steady-state expected number in system. We assume that the input distribution family is known, but the parameters need to be estimated from observed data. In each macro replication, we collect 100 observations generated from the two exponential distributions with true rates $\theta^c=(\lambda^c,\mu^c)$, then compute the maximum likelihood estimators, $\widehat{\theta}=(\widehat{\lambda},\widehat{\mu})$, of θ^c. Let $\tilde{f}(\widehat{\theta}|\theta^c)$ be the distribution for $\widehat{\theta}$.

For each $i=1,2,\ldots,\tilde{n}$, 100 interarrival and service times are generated from the exponential distributions with rate vector $\widehat{\theta}$, and $\tilde{\theta}_i$ is the MLE of the sample. We consider two ways to generate the experiment set: (i) We sample $\{\theta_i\}_{1\le i\le n}$ from $\tilde{f}(\tilde{\theta}|\widehat{\theta})$ independently from $\{\tilde{\theta}_i\}_{1\le i\le\tilde{n}}$. (ii) We first construct the smallest ellipsoid that encapsulates $\{\tilde{\theta}_i\}_{1\le i\le\tilde{n}}$ and then sample $\{\theta_i\}_{1\le i\le n}$ uniformly from the ellipsoid. In the second implementation, the sampling density $f(\theta|\widehat{\theta})\neq\tilde{f}(\tilde{\theta}|\widehat{\theta})$ and f is clearly dependent on $\{\tilde{\theta}_i\}_{1\le i\le\tilde{n}}$.

The queueing simulation model is initialized with an empty system and the regenerative state is chosen as when the server becomes idle. We pool simulation outputs $Y_j\left(\theta_{(i)}\right)$ and $A_j\left(\theta_{(i)}\right)$ to calculate the estimators $\{\widehat{\eta}_{kNN}(\tilde{\theta})\}_{1\le i\le\tilde{n}}$ and $\{\widehat{\eta}_{kNNLR}(\tilde{\theta})\}_{1\le i\le\tilde{n}}$ as defined in (6) and (8), respectively The resulting percentile bootstrap CIs are then compared to those constructed using the standard estimator (2).

In our experiments, we fix the simulation budget at $nr=10{,}000$ and change the value of n and r to examine the resilience of each method to the choice of experiment design parameters. The parameter k is selected to be $\sqrt{n}/\log(\log(\log(n)))$, which satisfies the conditions for k and n in Theorem 1 and Proposition 2. We also fix the size of the bootstrap set to be $\tilde{n}=1{,}000$ and repeat 1,000 independent macro runs. The column labeled with $\widehat{\eta}_{kNN}$ and $\widehat{\eta}_{kNN}$-ellipsoid summarize the results for kNN method that use parametric bootstrapping and the ellipsoid implementation to generate the simulation set, respectively. A similar labeling applies to $\widehat{\eta}_{kNNLR}$ and $\widehat{\eta}_{kNNLR}$-ellipsoid.

Table 1 presents the empirical coverage probabilities and average CI widths where the true unknown parameters are $\lambda^c=0.8$ and $\mu^c=1$. The target coverage is 99% (95%) for Tables 1(a) and 1(b) (Tables 1(c) and 1(d)). We observe that $\widehat{\eta}_{std}$ tends to show over-coverage when n is big and r is small. Among our choices of parameters, the standard estimator matches the coverage target the best when $n=r=100$.

We also observe under-coverage for $\widehat{\eta}_{kNN}$, especially when n is small. We have empirically observed that $\tilde{f}(\tilde{\theta}|\widehat{\theta})$ has relatively small values at $\tilde{\theta}$ that are mapped to an extreme (lower) quantile of $\eta(\tilde{\theta})$. Hence, when f is identical to $\tilde{f}$, there are very few points in the simulation set near those $\tilde{\theta}$s. This makes the kNN distance from such $\tilde{\theta}$ large and increases the bias in $\widehat{\eta}_{kNN}$, as discussed in Proposition 1, which in

Table 1: Comparison of empirical coverage probabilities and CI widths with 1000 macro-runs on an $M/M/1/10$ system with $\lambda^c = 0.8$ and $\mu^c = 1$. Numbers in parenthesis are the respective standard errors.

n	r	$\widehat{\eta}_{std}$	$\widehat{\eta}_{kNN}$	$\widehat{\eta}_{kNNLR}$	$\widehat{\eta}_{kNN}$-ellipsoid	$\widehat{\eta}_{kNNLR}$-ellipsoid
100	100	98.90% (0.33%)	75.70% (1.36%)	98.40% (0.40%)	96.40% (0.59%)	98.40% (0.40%)
200	50	99.50% (0.22%)	86.40% (1.08%)	98.60% (0.37%)	98.60% (0.37%)	98.80% (0.34%)
400	25	100.00% (0.00%)	91.60% (0.88%)	98.80% (0.34%)	98.90% (0.33%)	98.70% (0.36%)
1000	10	100.00% (0.00%)	95.10% (0.68%)	98.60% (0.37%)	98.70% (0.36%)	98.70% (0.36%)
5000	2	100.00% (0.00%)	97.70% (0.47%)	99.00% (0.31%)	99.30% (0.26%)	99.40% (0.24%)

(a) Empirical coverage probabilities with target coverage 99%.

n	r	$\widehat{\eta}_{std}$	$\widehat{\eta}_{kNN}$	$\widehat{\eta}_{kNNLR}$	$\widehat{\eta}_{kNN}$-ellipsoid	$\widehat{\eta}_{kNNLR}$-ellipsoid
100	100	5.15 (0.03)	2.68 (0.02)	4.63 (0.03)	5.28 (0.03)	4.78 (0.03)
200	50	5.42 (0.03)	3.19 (0.02)	4.70 (0.03)	5.28 (0.03)	4.87 (0.03)
400	25	5.85 (0.03)	3.56 (0.02)	4.76 (0.03)	5.19 (0.03)	4.92 (0.03)
1000	10	6.41 (0.03)	3.97 (0.03)	4.82 (0.03)	5.10 (0.03)	4.95 (0.03)
5000	2	6.99 (0.02)	4.53 (0.03)	4.96 (0.03)	5.13 (0.03)	5.08 (0.03)

(b) Average CI widths with target coverage 99%.

n	r	$\widehat{\eta}_{std}$	$\widehat{\eta}_{kNN}$	$\widehat{\eta}_{kNNLR}$	$\widehat{\eta}_{kNN}$-ellipsoid	$\widehat{\eta}_{kNNLR}$-ellipsoid
100	100	96.20% (0.60%)	72.30% (1.42%)	94.70% (0.71%)	91.00% (0.91%)	95.10% (0.68%)
200	50	98.00% (0.44%)	81.80% (1.22%)	94.40% (0.73%)	93.60% (0.77%)	95.30% (0.67%)
400	25	99.20% (0.28%)	86.00% (1.10%)	94.90% (0.70%)	95.00% (0.69%)	95.30% (0.67%)
1000	10	99.80% (0.14%)	91.20% (0.90%)	95.20% (0.68%)	95.20% (0.68%)	95.70% (0.64%)
5000	2	99.80% (0.14%)	94.80% (0.70%)	95.80% (0.63%)	96.10% (0.61%)	96.00% (0.62%)

(c) Empirical coverage probabilities with target coverage 95%.

n	r	$\widehat{\eta}_{std}$	$\widehat{\eta}_{kNN}$	$\widehat{\eta}_{kNNLR}$	$\widehat{\eta}_{kNN}$-ellipsoid	$\widehat{\eta}_{kNNLR}$-ellipsoid
100	100	4.21 (0.03)	2.48 (0.02)	3.76 (0.03)	4.50 (0.03)	3.82 (0.03)
200	50	4.48 (0.03)	2.89 (0.02)	3.80 (0.02)	4.33 (0.03)	3.86 (0.03)
400	25	4.90 (0.03)	3.16 (0.02)	3.84 (0.03)	4.18 (0.03)	3.89 (0.03)
1000	10	5.52 (0.03)	3.44 (0.02)	3.88 (0.03)	4.08 (0.03)	3.92 (0.03)
5000	2	5.90 (0.03)	3.81 (0.02)	4.01 (0.03)	4.10 (0.03)	4.05 (0.03)

(d) Average CI widths with target coverage 95%.

turn causes a poor coverage. As n increases, more data points are sampled around $\tilde{\theta}$, reducing its kNN distance and bias in $\widehat{\eta}_{kNN}$, thereby improving the coverage.

For the ellipsoid implementation, the likelihood of sampling scenarios near $\tilde{\theta}$ that are mapped to extreme quantiles increases. This alleviates the bias in $\widehat{\eta}_{kNN}$ and so the coverage of $\widehat{\eta}_{kNN}$-ellipsoid is significantly improved compared to that of $\widehat{\eta}_{kNN}$. The $\widehat{\eta}_{kNN}$-ellipsoid also exhibits robustness to the choice of n. Interestingly, observe that $\widehat{\eta}_{kNNLR}$ performs well with both sampling schemes and the CI width is slightly inflated for the ellipsoid implementation. The choice of f does not appear to have a significant impact on the performance of $\widehat{\eta}_{kNNLR}$. This is because the bias caused by the kNN method is eliminated by the LR method. Both $\widehat{\eta}_{kNNLR}$ and $\widehat{\eta}_{kNNLR}$-ellipsoid show robustness to the choice of n.

We also examine a lightly loaded system, where $\lambda^c = 0.5$ and $\mu^c = 1.5$ with coverage targets 99% and 95%. Table 2 summarizes the experiment results. Similar to the first case, $\widehat{\eta}_{std}$ shows over-coverage with larger CI widths, especially when n is large. The $\widehat{\eta}_{kNN}$ again exhibits under-coverage when n is small, but the under-coverage is improved for $\widehat{\eta}_{kNN}$-ellipsoid. Similar to the heavily loaded system, $\widehat{\eta}_{kNNLR}$ performs well with both sampling schemes, but with a slightly bigger standard error in all statistics.

Table 2: Comparison of empirical coverage probabilities and CI width with 1000 macro-runs on an $M/M/1/10$ system with $\lambda^c = 0.5$ and $\mu^c = 1.5$. Numbers in parenthesis are the respective standard errors.

n	r	$\widehat{\eta}_{std}$	$\widehat{\eta}_{kNN}$	$\widehat{\eta}_{kNNLR}$	$\widehat{\eta}_{kNN}$-ellipsoid	$\widehat{\eta}_{kNNLR}$-ellipsoid
100	100	99.60% (0.20%)	72.10% (1.42%)	98.00% (0.44%)	94.90% (0.70%)	98.80% (0.34%)
200	50	100% (0.00%)	84.60% (1.14%)	98.60% (0.37%)	98.30% (0.41%)	99.00% (0.31%)
400	25	100% (0.00%)	90.30% (0.94%)	98.40% (0.40%)	98.70% (0.36%)	98.90% (0.33%)
1000	10	100% (0.00%)	95.40% (0.66%)	99.00% (0.31%)	99.00% (0.31%)	99.20% (0.28%)
5000	2	100% (0.00%)	98.80% (0.34%)	99.30% (0.26%)	99.50% (0.22%)	99.50% (0.22%)

(a) Empirical coverage probabilities with target coverage 99%.

n	r	$\widehat{\eta}_{std}$	$\widehat{\eta}_{kNN}$	$\widehat{\eta}_{kNNLR}$	$\widehat{\eta}_{kNN}$-ellipsoid	$\widehat{\eta}_{kNNLR}$-ellipsoid
100	100	0.877 (0.011)	0.270 (0.003)	0.666 (0.008)	0.590 (0.008)	0.668 (0.008)
200	50	1.013 (0.010)	0.344 (0.004)	0.671 (0.008)	0.646 (0.008)	0.674 (0.008)
400	25	1.284 (0.012)	0.409 (0.005)	0.679 (0.008)	0.672 (0.008)	0.680 (0.008)
1000	10	1.742 (0.014)	0.498 (0.005)	0.700 (0.008)	0.695 (0.008)	0.695 (0.008)
5000	2	2.480 (0.013)	0.665 (0.007)	0.779 (0.009)	0.767 (0.008)	0.775 (0.008)

(b) Average CI widths with target coverage 99%.

n	r	$\widehat{\eta}_{std}$	$\widehat{\eta}_{kNN}$	$\widehat{\eta}_{kNNLR}$	$\widehat{\eta}_{kNN}$-ellipsoid	$\widehat{\eta}_{kNNLR}$-ellipsoid
100	100	98.10% (0.43%)	67.40% (1.48%)	94.90% (0.70%)	90.50% (0.93%)	95.40% (0.66%)
200	50	99.60% (0.20%)	79.20% (1.28%)	95.10% (0.68%)	94.30% (0.73%)	95.90% (0.63%)
400	25	100% (0.00%)	85.50% (1.11%)	95.50% (0.66%)	95.20% (0.68%)	96.10% (0.61%)
1000	10	100% (0.00%)	91.40% (0.89%)	96.10% (0.61%)	95.40% (0.66%)	95.80% (0.63%)
5000	2	100% (0.00%)	95.90% (0.63%)	97.40% (0.50%)	96.90% (0.55%)	97.20% (0.52%)

(c) Empirical coverage probabilities with target coverage 95%.

n	r	$\widehat{\eta}_{std}$	$\widehat{\eta}_{kNN}$	$\widehat{\eta}_{kNNLR}$	$\widehat{\eta}_{kNN}$-ellipsoid	$\widehat{\eta}_{kNNLR}$-ellipsoid
100	100	0.631 (0.007)	0.246 (0.003)	0.479 (0.005)	0.466 (0.006)	0.479 (0.005)
200	50	0.732 (0.007)	0.306 (0.003)	0.482 (0.005)	0.494 (0.006)	0.484 (0.005)
400	25	0.896 (0.008)	0.355 (0.004)	0.487 (0.005)	0.499 (0.006)	0.490 (0.005)
1000	10	1.170 (0.009)	0.415 (0.004)	0.503 (0.005)	0.507 (0.005)	0.502 (0.005)
5000	2	1.642 (0.008)	0.517 (0.005)	0.561 (0.006)	0.557 (0.006)	0.556 (0.006)

(d) Average CI width with target coverage 95%.

In summary, we recommend the kNN estimator with the ellipsoid sampling implementation and the kNNLR estimator due to their computational efficiencies demonstrated by the empirical analyses.

ACKNOWLEDGMENTS

This work is partially supported by the National Science Foundation (NSF) grants DMS-1854659 and CMMI-2045400. This work is partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants RGPIN-2018-03755.

REFERENCES

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. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Barton, R. R., H. Lam, and E. Song. 2022. "Input Uncertainty in Stochastic Simulation". In *The Palgrave Handbook of Operations Research*, edited by S. Salhi and J. Boylan, 573–620. Cham, Switzerland: Springer International Publishing.

Barton, R. R., and L. W. Schruben. 2001. "Resampling Methods for Input Modeling". In *Proceedings of the 2001 Winter Simulation Conference*, edited by B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, 372–378. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and Their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge, United Kingdom: Cambridge University Press.

Devroye, L. P., and T. J. Wagner. 1977. "The Strong Uniform Consistency of Nearest Neighbor Density Estimates". *The Annals of Statistics* 5(3):536–540.

Glynn, P. W. 2006. "Simulation Algorithms for Regenerative Processes". In *Handbooks in Operations Research and Management Science*, edited by S. G. Henderson and B. L. Nelson, Volume 13, 477–500. Elsevier.

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.

Henderson, S. G., and P. W. Glynn. 2001. "Regenerative Steady-State Simulation of Discrete-Event Systems". *ACM Transactions on Modeling and Computer Simulation* 11(4):313–345.

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

Mack, Y., and M. Rosenblatt. 1979. "Multivariate k-Nearest Neighbor Density Estimates". *Journal of Multivariate Analysis* 9(1):1–15.

Mack, Y.-P. 1981. "Local Properties of k-NN Regression Estimates". *SIAM Journal on Algebraic Discrete Methods* 2(3):311–323.

Ross, S. M. 1995. *Stochastic Processes*. Hoboken, New Jersey: John Wiley & Sons.

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

AUTHOR BIOGRAPHIES

LINYUN HE is a Ph.D. student in the School of Industrial and Systems Engineering at Georgia Institute of Technology. His research interests include simulation optimization, stochastic optimization, non-parametric methods and high-dimensional statistics. His email address is lhe85@gatech.edu. His website is https://dongfengguzhu.github.io.

BEN FENG is an Assistant Professor in actuarial science at the University of Waterloo. He earned his Ph.D. in the Department of Industrial Engineering and Management Sciences at Northwestern University. He is an Associate of the Society of Actuaries (ASA). His research interests include stochastic simulation design and analysis, optimization via simulation, nonlinear optimization, and financial and actuarial applications of simulation and optimization methodologies. His e-mail address is ben.feng@uwaterloo.ca. His website is http://www.math.uwaterloo.ca/∼mbfeng/.

EUNHYE SONG is a Coca-Cola Foundation Early Career Professor and Assistant Professor in the School of Industrial and Systems Engineering at Georgia Institute of Technology. She earned her PhD degree in Industrial Engineering and Management Sciences at Northwestern University. Her research interests include simulation design of experiments, uncertainty and risk quantification, and simulation optimization. Her email address is eunhye.song@isye.gatech.edu. Her website is http://eunhyesong.info.