Submitted to *Operations Research*
manuscript OPRE-2014-08-465-final

Quantile Estimation With Latin Hypercube Sampling

Hui Dong
Supply Chain Management and Marketing Sciences Dept., Rutgers Univ., Newark, NJ 07102, huidong@rutgers.edu
Marvin K. Nakayama
Computer Science Department, New Jersey Institute of Technology, Newark, NJ 07102, marvin@njit.edu

Quantiles are often used to measure risk of stochastic systems. We examine quantile estimators obtained using simulation with Latin hypercube sampling (LHS), a variance-reduction technique that efficiently extends stratified sampling to higher dimensions and produces negatively correlated outputs. We consider single-sample LHS (ssLHS), which minimizes the variance that can be obtained from LHS, and also replicated LHS (rLHS). We develop a consistent estimator of the asymptotic variance of the ssLHS quantile estimator's central limit theorem, enabling us to provide the first confidence interval (CI) for a quantile when applying ssLHS. For rLHS, we construct CIs using batching and sectioning. On average, our rLHS CIs are shorter than previous rLHS CIs and only slightly wider than the ssLHS CI. We establish the asymptotic validity of the CIs by first proving that the quantile estimators satisfy Bahadur representations, which show that the quantile estimators can be approximated by linear transformations of estimators of the cumulative distribution function (CDF). We present numerical results comparing the various CIs.

Key words: simulation: efficiency, statistical analysis; reliability: system safety

History:

1. Introduction

For a given constant $0 < p < 1$, the p-quantile of a continuous random variable Y is a constant such that Y has probability p of lying below the constant. We can also express the p-quantile in terms of the inverse of the CDF of Y. For example, the median corresponds to the 0.5-quantile.

Many application areas use quantiles to measure risk of stochastic systems. In financial portfolio management, a quantile is called a value-at-risk, and portfolio risk and capital adequacy are often assessed via a 0.99-quantile; e.g., see Chapter 9 of Glasserman (2004). A project manager may want to determine the 0.95-quantile of the project completion time, with which he/she can then determine how long the start of a project can be delayed and still have a high chance of it completing by a specified deadline. Safety and uncertainty analyses of nuclear power plants are usually based on 0.95-quantiles; e.g., see U.S. Nuclear Regulatory Commission (2010).

This paper studies simulation methods to estimate a quantile. The typical approach first estimates the CDF, which is inverted to obtain a quantile estimator. We also want a confidence interval (CI) for the quantile to provide a measure of the estimator's error. Indeed, in safety analyses of nuclear power plants, the U.S. Nuclear Regulatory Commission requires that plant licensees verify, with 95% confidence, that a 0.95-quantile lies below a mandated threshold. This is known as a *95/95 criterion*; e.g., see Section 24.9 of U.S. Nuclear Regulatory Commission (2011).

When *simple random sampling* (SRS; i.e., simulation without any variance reduction) is applied, there are several methods to construct a CI for a quantile. One approach exploits a binomial property of the independent and identically distributed (i.i.d.) outputs to obtain a nonparametric CI based on order statistics; e.g., see Section 2.6.1 of Serfling (1980). An alternative method starts by establishing a central limit theorem (CLT) for the SRS quantile estimator (Serfling 1980, Section 2.3.3), and then unfolds the CLT to obtain a CI with a consistent estimator of the CLT's asymptotic variance. Techniques for consistently estimating the asymptotic variance include a finite difference (FD; Bloch and Gastwirth (1968), Bofinger (1975)) and kernel methods (Falk (1986)). These approaches require the user to specify a parameter known as the *bandwidth*, and choosing a "good" bandwidth value can be difficult in practice. The bootstrap has also been applied to construct a CI for a quantile when applying SRS (Hall and Martin (1988)). Unfortunately, the bootstrap quantile variance estimator converges more slowly than the consistent estimators above, which adversely affects the coverage of the bootstrap CI for a quantile (Hall and Martin (1989)).

We can avoid having to consistently estimate the asymptotic variance by employing a method that cancels out the asymptotic variance in a relevant limit theorem. One such technique is *batching* (e.g., p. 491 of Glasserman (2004)), also known as subsampling. This method first divides the n outputs into $b \geq 2$ batches, each of size $m = n/b$, and computes a quantile estimator from each batch. Then the sample mean and sample variance of the quantile estimates across the batches are used to construct a CI for the quantile. A problem with batching arises from the fact that quantile estimators are generally biased (Avramidis and Wilson (1998)). While the bias vanishes as the sample size grows to infinity, the batching point estimator can have significant bias, which is determined by the batch size $m < n$. Thus, when the overall sample size n is not large, the batching CI may have poor coverage because it is centered at a highly contaminated point estimator.

To address this issue, we can instead apply *sectioning*, which was originally proposed in Section III.5a of Asmussen and Glynn (2007) for SRS and extended to certain variance-reduction techniques (VRTs) by Nakayama (2014). Similar to batching, sectioning replaces the batching point estimator with the overall point estimator throughout the batching CI. Since the overall quantile estimator has bias determined by the overall sample size, it is typically less biased than the batching point estimator, and this can lead to improved CI coverage when the sample size is small.

Because the SRS CI for the p-quantile can be unusably wide, particularly when $p \approx 0$ or $p \approx 1$, we may instead apply a VRT. This is especially important when a single simulation run takes enormous computational effort to complete, as in many application areas. The current paper focuses on Latin hypercube sampling (LHS), originally developed by McKay et al. (1979) as a way to efficiently extend stratified sampling to higher dimensions, producing negatively correlated outputs. Stein (1987) further analyzes the approach and shows that the LHS estimator of a mean has asymptotic variance that is no larger than its SRS counterpart. Owen (1992) proves a CLT for the LHS estimator of a mean of bounded outputs, and Loh (1996) extends this to outputs having a finite absolute third moment. Avramidis and Wilson (1998) apply LHS to estimate a quantile and verify that the LHS quantile estimator satisfies a CLT. The intuitive appeal of and ease of implementing

LHS make it perhaps the most common VRT in certain fields, such as nuclear engineering, although not currently for estimating quantiles. Citing over 300 references, Helton and Davis (2003) survey the extensive literature on LHS, with a particular focus on its applications in nuclear engineering.

Other VRTs used in quantile estimation include control variates (CV) (Hsu and Nelson (1990), Hesterberg and Nelson (1998)); importance sampling (IS) (Glynn (1996), Sun and Hong (2010)); combined IS and stratified sampling (IS+SS) (Glasserman et al. (2000)); and correlation-induction methods, such as antithetic variates (AV) and LHS (Avramidis and Wilson (1998), Jin et al. (2003)). While these papers develop VRT quantile estimators, most do not consider how to construct a CI. One difficulty is that it is nontrivial to construct a consistent estimator of the asymptotic variance in the CLT for the VRT quantile estimator.

Chu and Nakayama (2012a) develop a general asymptotically-valid approach for VRT CIs, and they show how it applies for IS, IS+SS, CV, and AV. Nakayama (2011) extends this to a type of replicated LHS (rLHS), in which r independent LHS samples are generated, each of a fixed size m, where the asymptotic validity holds as $r \to \infty$ with m fixed. The approach in these papers first establishes that the VRT quantile estimator satisfies a so-called Bahadur representation (Bahadur (1966), Ghosh (1971)), which approximates a quantile estimator as the true quantile plus a linear transformation of a CDF estimator, and then leverages this result to consistently estimate the asymptotic variance using a FD. Other methods for constructing asymptotically valid CIs for quantiles when applying VRTs include batching and sectioning (Nakayama (2014) for IS and CV) and the bootstrap (Liu and Yang (2012) for IS).

In this paper, we consider LHS quantile estimators. We examine both a single-sample LHS (ssLHS) estimator, where there is only one LHS sample of size n, and rLHS, with b independent LHS samples, each of size m. ***We prove that each quantile estimator obeys a Bahadur representation*** (as $n \to \infty$ for ssLHS, and as $m \to \infty$ with b fixed for rLHS). These lead to CLTs, but the ssLHS CLT's asymptotic variance is "neither known nor easily estimated" (Glasserman 2004, p. 242). Indeed, although Avramidis and Wilson (1998) derive the variance's form, they do

not give a variance estimator nor a quantile CI. To address this shortcoming, ***we now develop an ssLHS CI by providing a consistent estimator for the ssLHS quantile estimator's asymptotic variance***, which is a ratio of two terms. For the numerator, we modify an approach of Owen (1992), which consistently estimates the asymptotic variance in the CLT for the estimator of a mean of bounded, computable outputs. In our case, though, directly applying Owen's method would require knowing the true value of the p-quantile, which obviously is unknown; nevertheless, we surmount this issue. We estimate the denominator with a FD, as in Chu and Nakayama (2012a).

We also develop rLHS CIs using batching and sectioning. We fix the number $b \geq 2$ of batches, each with a single LHS sample of size m, for bm total outputs. Naming this method *single replicated LHS (srLHS)*, ***we prove the validity of the srLHS batching and sectioning CIs*** as $m \to \infty$ with b fixed. This contrasts an alternative approach of Nakayama (2012), which we call *multiple replicated LHS (mrLHS)*. In mrLHS, we increase the number r of independent replicated LHS samples, each of fixed size m (for rm total outputs); the r LHS replicates are divided into a fixed number $b \geq 2$ of batches, with each batch comprising r/b independent LHS samples. The rLHS methods differ in their asymptotic regimes: srLHS is valid as $m \to \infty$ with b fixed, whereas mrLHS requires $r \to \infty$ with b and m fixed. The distinction is important as the LHS sample size determines how much the variance is lessened. For a given overall sample size n, srLHS allows for the LHS size to be large, but mrLHS has fixed LHS size, which is small since r must be large. Thus, ***srLHS can reduce variance more than mrLHS***, and our numerical results confirm this. Moreover, we provide theoretical and numerical evidence showing that for the same n, ***the variance of srLHS is comparable to ssLHS, which minimizes the variance that can be obtained by LHS***.

The rest of the paper progresses as follows. Section 2 reviews SRS CIs for quantiles. We discuss ssLHS quantile estimation in Section 3. Section 4 presents CIs based on ssLHS and srLHS, and provides an asymptotic theoretical comparison of the methods. Section 5 contains numerical results, and we make concluding remarks in Section 6. All proofs are in appendices. Our srLHS batching and sectioning CIs also appear (without proofs) in Dong and Nakayama (2014).

2. Background

Consider a random variable Y that can be expressed as

$$Y = g(U_1, U_2, \ldots, U_d), \tag{1}$$

where $g : \Re^d \to \Re$ is a given (measurable) function that can be computed but may not be known in closed form, and $U_1, U_2, \ldots, U_d$ are d i.i.d. unif$[0,1]$ random variables, where d is fixed. Let F denote the CDF of Y, and for $0 < p < 1$, we define the p-quantile of F (or equivalently of Y) to be $\xi_p = F^{-1}(p) \equiv \inf\{x : F(x) \geq p\}$. Our goal is to estimate and construct a CI for ξ_p via simulation.

We assume that F is unknown but that we can still generate observations of Y through (1). Our framework includes as a special case a complicated computer code g_1 that takes as input a fixed number of random variables with specified joint distribution to produce a single output. For example, suppose that $X_1, X_2, \ldots, X_d$ are independent (but not necessarily identically distributed) random inputs, where H_j is the marginal CDF of X_j, which we can generate as $X_j = H_j^{-1}(U_j)$. Then the output is $Y = g_1(X_1, X_2, \ldots, X_d) = g_1(H_1^{-1}(U_1), H_2^{-1}(U_2), \ldots, H_d^{-1}(U_d)) \equiv g(U_1, U_2, \ldots, U_d)$.

2.1. Constructing a Confidence Interval When Applying SRS

To apply SRS to generate n i.i.d. outputs, we first generate $n \times d$ i.i.d. unif$[0,1]$ random variables

$$\begin{array}{cccc} U_{1,1} & U_{1,2} & \cdots & U_{1,d} \\ U_{2,1} & U_{2,2} & \cdots & U_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ U_{s,1} & U_{s,2} & \cdots & U_{s,d} \end{array} \tag{2}$$

for $s = n$. We apply function g from (1) to each row to generate a sample of n i.i.d. outputs $Y_i = g(U_{i,1}, U_{i,2}, \ldots, U_{i,d})$, $i = 1, 2, \ldots, n$. Let F_n be the SRS estimator of the CDF F, where $F_n(y) = (1/n)\sum_{1 \leq i \leq n} I(Y_i \leq y)$ and $I(\cdot)$ is the indicator function, which equals 1 (resp., 0) when its argument is true (resp., false). The SRS estimator of the p-quantile $\xi_p = F^{-1}(p)$ is $\xi_{p,n} = F_n^{-1}(p) = Y_{n:\lceil np \rceil}$, where $Y_{n:1} \leq Y_{n:2} \leq \cdots \leq Y_{n:n}$ are the sample's order statistics and $\lceil \cdot \rceil$ is the ceiling function.

Let F' denote the derivative (when it exists) of F. If $F'(\xi_p) > 0$, the SRS p-quantile estimator $\xi_{p,n}$ satisfies a CLT $(\sqrt{n}/\tau_p)(\xi_{p,n} - \xi_p) \Rightarrow N(0,1)$ as $n \to \infty$ (Serfling 1980, Section 2.3.3), where $\tau_p = \sqrt{p(1-p)}/F'(\xi_p)$, $N(a, b^2)$ is a normal random variable with mean a and variance b^2, and $\Rightarrow$ denotes convergence in distribution (e.g., see Section 25 of Billingsley (1995)). Unfolding the CLT leads to $C_n = [\xi_{p,n} \pm z_{\alpha/2}\tau_p/\sqrt{n}]$ as an asymptotic $100(1-\alpha)\%$ confidence interval for ξ_p, where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ and Φ is the $N(0,1)$ CDF. But the CI C_n is unusable since $F'(\xi_p)$ is unknown.

One way to consistently estimate τ_p when applying SRS employs a finite difference to estimate the *sparsity function* $\lambda_p \equiv 1/F'(\xi_p)$. Because $\frac{d}{dp}F^{-1}(p) = \lim_{h\to 0}[F^{-1}(p+h) - F^{-1}(p-h)]/(2h) = 1/F'(F^{-1}(p)) = \lambda_p$, we estimate λ_p by $\lambda_{p,n} = [F_n^{-1}(p+h_n) - F_n^{-1}(p-h_n)]/(2h_n)$, where $h_n > 0$ is a user-specified *bandwidth*. Then $\tau_{p,n} \equiv \sqrt{p(1-p)}\lambda_{p,n} \Rightarrow \tau_p$ as $n \to \infty$ when $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$; see Bloch and Gastwirth (1968) and Bofinger (1975). Hence, a two-sided $100(1-\alpha)\%$ confidence interval for ξ_p is $C_{n,\mathrm{SRS}} = [\xi_{p,n} \pm z_{\alpha/2}\tau_{p,n}/\sqrt{n}]$, which is *asymptotically valid*, i.e., the *coverage* $P(\xi_p \in C_{n,\mathrm{SRS}}) \to 1-\alpha$ as $n \to \infty$. The choice of bandwidth h_n in the FD has a significant impact on the quality of the estimator $\tau_{p,n}$, which can affect the coverage for small n, and it can be difficult to specify an appropriate value for h_n in practice; e.g., see Bofinger (1975), Hall and Sheather (1988) and Chu and Nakayama (2012b).

2.2. Batching and Sectioning CIs When Using SRS

Rather than trying to consistently estimate τ_p, we can construct a CI by instead applying a *cancellation method*, which cancels τ_p in the relevant limit theorem in a manner analogous to the Student t statistic. Two such techniques are batching (also known as subsampling) and sectioning. Batching divides the overall sample $Y_1, Y_2, \ldots, Y_n$, of size n into $b \geq 2$ batches, each containing a sample of size $m = n/b$. For each batch $j = 1, 2, \ldots, b$, let $Y_{j,i} = Y_{(j-1)m+i}$, $i = 1, 2, \ldots, m$, be the ith output in the jth batch, and define a CDF estimator for batch j as $F_{j,m}(y) = (1/m)\sum_{1\leq i\leq m} I(Y_{j,i} \leq y)$. The p-quantile estimator from the jth batch is $\xi_{p,j,m} = F_{j,m}^{-1}(p)$, and we define $\bar{\xi}_{p,b,m} = (1/b)\sum_{1\leq j\leq b}\xi_{p,j,m}$ as the SRS batching p-quantile estimator. The sample variance of $\xi_{p,j,m}$, $j = 1, 2, \ldots, b$, is $S^2_{b,m,\mathrm{batch}} = (1/(b-1))\sum_{1\leq j\leq b}(\xi_{p,j,m} - \bar{\xi}_{p,b,m})^2$. Let $t_{b-1,\alpha/2} = G^{-1}(1-\alpha/2)$, where G is the CDF of a Student t

random variable with $b-1$ degrees of freedom. The SRS batching two-sided $100(1-\alpha)\%$ confidence interval for ξ_p is then $C_{b,m,\text{batch}} = [\bar{\xi}_{p,b,m} \pm t_{b-1,\alpha/2} S_{b,m,\text{batch}}/\sqrt{b}]$, which is asymptotically valid.

A problem for the batching CI stems from quantile estimators generally being biased. Although the bias vanishes as the sample size grows large (Avramidis and Wilson (1998)), it can be significant for small sample sizes. The batch size $m = n/b < n$ determines the bias of the batching estimator $\bar{\xi}_{p,b,m}$, so batching centers its CI at an estimator that may be considerably biased. Hence, the batching CI may have poor coverage, especially when the overall sample size n is not very large.

Sectioning (Section III.5a of Asmussen and Glynn (2007) for SRS, Nakayama (2014) for IS and CV) addresses this issue by replacing the batching point estimator $\bar{\xi}_{p,b,m}$ in the batching CI with the overall quantile estimator $\xi_{p,n}$, based on the CDF estimator F_n from all n outputs. The SRS sectioning two-sided $100(1-\alpha)\%$ CI for ξ_p is $C_{b,m,\text{sec}} = [\xi_{p,n} \pm t_{b-1,\alpha/2} S_{b,m,\text{sec}}/\sqrt{b}]$, where $S^2_{b,m,\text{sec}} = (1/(b-1)) \sum_{1 \le j \le b} (\xi_{p,j,m} - \xi_{p,n})^2$.

While $C_{b,m,\text{batch}}$ and $C_{b,m,\text{sec}}$ are asymptotically valid CIs, the sectioning CI usually has better coverage, especially when the overall sample size $n = bm$ is small. For both batching and sectioning, Section III.5a of Asmussen and Glynn (2007) suggests choosing $b \le 30$. Nakayama (2014) presents numerical results with $b = 10$ and $b = 20$ when applying each of SRS, IS, and CV, and found that $b = 10$ resulted in significantly better coverage than $b = 20$ for the same overall sample size n when n is small and $p \approx 1$.

3. Latin Hypercube Sampling

We now describe how to generate a single LHS sample of size s. Start with $s \times d$ i.i.d. unif$[0,1]$ random variables $U_{i,j}$, $i = 1, 2, \ldots, s$, $j = 1, 2, \ldots, d$, as in (2). Also, generate $\pi_1, \pi_2, \ldots, \pi_d$ as d independent random permutations of $(1, 2, \ldots, s)$, where $\pi_j = (\pi_j(1), \pi_j(2), \ldots, \pi_j(s))$, $\pi_j(i)$ is the value to which i is mapped in the jth permutation, and each of the $s!$ permutations is equally likely. For $i = 1, 2, \ldots, s$, and $j = 1, 2, \ldots, d$, define

$$V_{i,j} = \frac{\pi_j(i) - 1 + U_{i,j}}{s}, \tag{3}$$

which we arrange in a grid

$$\begin{matrix} V_{1,1} & V_{1,2} & \cdots & V_{1,d} \\ V_{2,1} & V_{2,2} & \cdots & V_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ V_{s,1} & V_{s,2} & \cdots & V_{s,d} \end{matrix}. \tag{4}$$

Then, apply function $g:\Re^d \to \Re^1$ from (1) to each row of (4) to obtain the s outputs

$$\hat{Y}_i = g(V_{i,1}, V_{i,2}, \cdots, V_{i,d}), \quad i = 1,2,\ldots,s. \tag{5}$$

We call $(\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_s)$ an LHS sample of size s. Each row i in (4) has d i.i.d. unif[0,1] random variables, so $\hat{Y}_i \sim F$, $i = 1,2,\ldots,s$. But since the entries in each column j in (4) share the same permutation π_j, the rows in (4) are dependent, making $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_s$ dependent as well.

LHS yields smaller asymptotic variance than SRS when estimating a mean (Stein (1987)) or a quantile (Avramidis and Wilson (1998)). Under the conditions below, Avramidis and Wilson (1998) prove a CLT for a quantile estimator using single-sample LHS. In this paper, we further show (Theorem 1 below) that the ssLHS quantile estimator obeys a weak Bahadur (1966) representation (Ghosh (1971)) under two continuity conditions from Avramidis and Wilson (1998), where $\mathbf{U} = (U_1, U_2, \ldots, U_d)$ denotes a vector of i.i.d. unif$[0,1]$ random variables:

CC1 *The function $g(\cdot)$ in (1) has a finite set of discontinuities $\mathcal{D}$.*

CC2 *There exists a neighborhood $\mathcal{N}(\xi_p)$ of ξ_p such that for each $x \in \mathcal{N}(\xi_p)$ and for each $j = 1,\ldots,d$, there exists a finite set $\mathcal{Q}_j(x)$ with $P(g(\mathbf{U}) = x \,|\, U_j = u_j) = 0$ for every $u_j \in [0,1] - \mathcal{Q}_j(x)$.*

Define the ssLHS CDF estimator based on a single LHS sample of size $s = n$ as

$$\hat{F}_n(y) = \frac{1}{n} \sum_{1 \le i \le n} I(\hat{Y}_i \le y), \tag{6}$$

and let $\hat{\xi}_{p,n} = \hat{F}_n^{-1}(p)$ be the ssLHS p-quantile estimator. We then have the following result:

THEOREM 1. *Suppose continuity conditions CC1 and CC2 hold, CDF F has a bounded second derivative in a neighborhood of ξ_p, and $F'(\xi_p) > 0$. Let $p_n = p + cn^{-1/2}$ for any constant $c \in \Re$, and*

let $\hat{\xi}_{p_n,n}$ be the ssLHS estimator of the p_n-quantile $\xi_{p_n} = F^{-1}(p_n)$, with ssLHS sample size n. Then $\hat{\xi}_{p_n,n}$ satisfies the weak Bahadur representation

$$\hat{\xi}_{p_n,n} = \xi^{\#}_{p_n} - [\hat{F}_n(\xi_p) - p]/F'(\xi_p) + \hat{R}_n, \quad \textit{with } n^{1/2}\hat{R}_n \Rightarrow 0 \quad \textit{as} \quad n \to \infty, \tag{7}$$

where $\xi^{\#}_{p_n} = \xi_p + (p_n - p)/F'(\xi_p)$ and $\hat{F}_n(\cdot)$ is defined in (6).

The dependence among $\hat{Y}_i$, $i = 1, 2, \ldots, n$, complicates Theorem 1's proof (in Appendix A). To handle this, we modify and extend arguments that Avramidis and Wilson (1996, 1998) develop to demonstrate the ssLHS quantile estimator $\hat{\xi}_{p,n}$ obeys a CLT.

Setting $c = 0$ in Theorem 1 corresponds to what we call a *fixed* weak Bahadur representation:

$$\hat{\xi}_{p,n} = \xi_p - [\hat{F}_n(\xi_p) - p]/F'(\xi_p) + \hat{R}_n \quad \text{with} \quad n^{1/2}\hat{R}_n \Rightarrow 0 \quad \text{as} \quad n \to \infty. \tag{8}$$

The fixed version shows that the p-quantile estimator has $n^{1/2}$-asymptotics determined by the CDF estimator at ξ_p. Because $\hat{F}_n(\xi_p)$ is the average of $I(\hat{Y}_i \le \xi_p)$, which are bounded with mean p,

$$n^{1/2}[\hat{F}_n(\xi_p) - p] \Rightarrow N(0, \psi_p^2) \tag{9}$$

as $n \to \infty$ for some $\psi_p > 0$ (defined later) by the CLT of Owen (1992), which holds for an LHS estimator of a mean of bounded outputs. Thus, rearranging (8) and multiplying by $n^{1/2}$ leads to

$$n^{1/2}(\hat{\xi}_{p,n} - \xi_p) = -n^{1/2}[\hat{F}_n(\xi_p) - p]/F'(\xi_p) + n^{1/2}\hat{R}_n \Rightarrow N(0, \eta_p^2) \tag{10}$$

as $n \to \infty$ by Slutsky's theorem (e.g., p. 19 of Serfling (1980)), where

$$\eta_p = \psi_p/F'(\xi_p). \tag{11}$$

While the limiting result given by the outer terms in (10) was previously proven by Avramidis and Wilson (1998), the fixed weak Bahadur representation provides further insight into why the ssLHS quantile estimator, which is *not* a sample average, satisfies the CLT. Moreover, the fixed weak form in (8) and the analogue for single replicated LHS suffice to establish the asymptotic validity of our srLHS sectioning CI in Section 4.2 below.

But we also develop an ssLHS CI with a finite difference to estimate λ_p. We only consider using a FD bandwidth $h_n = cn^{-1/2}$ for any constant $c > 0$. The consistency of the FD will then follow from the *perturbed* weak Bahadur representation (7) for perturbed $p_n = p \pm cn^{-1/2}$ in Theorem 1. Working with only p_n of this form facilitates our proof of Theorem 1 as it allows us to leverage results from Avramidis and Wilson (1996). Also, Chu and Nakayama (2012a) present numerical results when applying other VRTs showing that bandwidths of the form $h_n = cn^{-1/2}$ led to FD CIs with better small-sample coverage than setting $h_n = cn^{-a}$ for $a = 1/3$ and $1/5$, especially for $p \approx 1$.

4. Confidence Intervals When Applying LHS

We now develop three methods for constructing a CI for a quantile when applying LHS. The first approach uses only a single LHS sample and is based on consistently estimating the asymptotic variance constant η_p^2 in the CLT in (10) for the ssLHS quantile estimator $\hat{\xi}_{p,n}$. The other two CIs employ batching and sectioning, and these require independent replicated LHS samples.

4.1. CI Using ssLHS

Under the conditions in Theorem 1, Avramidis and Wilson (1996, 1998) prove the ssLHS quantile estimator $\hat{\xi}_{p,n}$ obeys the CLT in (10). However, as noted by (Glasserman 2004, p. 242), the asymptotic variance η_p^2 in (10) is "neither known nor easily estimated," making it "difficult" to obtain a CI from the ssLHS CLT. Indeed, Avramidis and Wilson (1996, 1998) do not discuss how to estimate η_p nor how to construct a CI for ξ_p when applying ssLHS. We now tackle these issues.

The difficulties in consistently estimating $\eta_p = \psi_p / F'(\xi_p)$ from (11) are twofold. First, the numerator ψ_p has a complicated form (described shortly) and depends on the unknown ξ_p, and we extend a method of Owen (1992) to consistently estimate ψ_p. The sparsity function $\lambda_p = 1/F'(\xi_p)$ also poses problems as it requires estimating the derivative F' of the unknown CDF F at the unknown ξ_p; we handle this using a finite difference with a technique developed in Chu and Nakayama (2012a), which relies on the perturbed Bahadur representation established in Theorem 1.

To give an expression for the numerator ψ_p in (11), we review concepts from Stein (1987) and Avramidis and Wilson (1998). Let $\phi(\cdot)$ be a real-valued, square-intergrable function defined over

the d-dimensional unit cube $[0,1]^d$. Let $\mu_\phi = \mathrm{E}[\phi(\mathbf{U})] = \int_{[0,1]^d} \phi(\mathbf{u})\, d\mathbf{u}$, where $\mathbf{U} = (U_1, U_2, \ldots, U_d)$ is a vector of i.i.d. unif$[0,1]$ random variables. Then define the jth *main effect* of $\phi(\cdot)$ as

$$\phi_j(u_j) = \mathrm{E}[\phi(\mathbf{U}) \,|\, U_j = u_j] = \int_{[0,1]^{d-1}} \phi(u_1, \ldots, u_j, \ldots, u_d) \prod_{\substack{1 \le j' \le d \\ j' \ne j}} du_{j'} \tag{12}$$

for $u_j \in [0,1]$ and $j = 1, 2, \ldots, d$; the *additive part* of $\phi(\cdot)$ is

$$\phi_{\mathrm{add}}(\mathbf{u}) = \sum_{1 \le j \le d} \phi_j(u_j) - (d-1)\mu_\phi, \tag{13}$$

where $\mathbf{u} = (u_1, u_2, \ldots, u_d) \in [0,1]^d$; and the *residual from additivity* of $\phi(\cdot)$ is

$$\phi_{\mathrm{res}}(\mathbf{u}) = \phi(\mathbf{u}) - \phi_{\mathrm{add}}(\mathbf{u}). \tag{14}$$

Recall g in (1), and define $\chi(\mathbf{u}) = I(g(\mathbf{u}) \le \xi_p)$, which is square-integrable as it is bounded. Let $\chi_j(\cdot)$, $\chi_{\mathrm{add}}(\cdot)$, and $\chi_{\mathrm{res}}(\cdot)$ be its jth main effect, additive part, and residual from additivity, respectively. Avramidis and Wilson (1998) show that

$$\psi_p^2 = \mathrm{E}[\chi_{\mathrm{res}}^2(\mathbf{U})] = \mathrm{Var}[\chi_{\mathrm{res}}(\mathbf{U})] = \mathrm{Var}[\chi(\mathbf{U})] - \sum_{j=1}^{d} \mathrm{Var}[\chi_j(U_j)], \tag{15}$$

where the second equality holds since $\mathrm{E}[\chi_{\mathrm{res}}(\mathbf{U})] = 0$ and the third because $\mathrm{Cov}[\chi(\mathbf{U}), \chi_j(U_j)] = \mathrm{Var}[\chi_j(U_j)]$, $j = 1, 2, \ldots, d$. Thus, LHS removes the variability of the additive part from the original response $\chi(\mathbf{U})$. The difficulties in estimating ψ_p^2 arise from the facts that $\chi_{\mathrm{res}}(\cdot)$ is not observable and depends implicitly on the unknown ξ_p.

To estimate ψ_p^2, we modify an estimator from Owen (1992), who gives a consistent estimator of $\mathrm{E}[\phi_{\mathrm{res}}^2(\mathbf{U})]$ for a bounded computable function ϕ. In our case, even though g in (1) can be computed, χ cannot because ξ_p is unknown, which makes estimating $\mathrm{E}[\chi_{\mathrm{res}}^2(\mathbf{U})]$ more difficult. Instead, we replace ξ_p in χ with $\hat{\xi}_{p,n}$, which consistently estimates ξ_p, but this significantly complicates the analysis of the resulting estimator. We first describe how to adapt Owen's estimator to estimate $\psi_p^2 = \mathrm{E}[\chi_{\mathrm{res}}^2(\mathbf{U})]$ under the assumption that ξ_p is known. Let $W_i = I(\hat{Y}_i \le \xi_p)$, $i = 1, 2, \ldots, n$. For $j = 1, 2, \ldots, d$, define the operator N_j such that

$$N_j W_i = \begin{cases} W_m & \text{if } \pi_j(i) < n, \text{ where } \pi_j(m) = \pi_j(i) + 1 \\ \bar{W}_n & \text{if } \pi_j(i) = n \end{cases}, \tag{16}$$

where $\pi_j(i)$ is used in the definition of $V_{i,j}$ in (3) and $\bar{W}_n = n^{-1}\sum_{1\le i\le n} W_i$. Thus, in the first case of (16), N_jW_i is the output W_m corresponding to the next larger input in the jth coordinate. Then under a Lipschitz condition (see (21) below), we can consistently estimate $\mathrm{E}[\chi^2_{\mathrm{res}}(\mathbf{U})]$ by

$$\frac{1}{2n}\sum_{j=1}^{d}\sum_{i=1}^{n}[W_i - N_jW_i]^2 - (d-1)p(1-p). \tag{17}$$

(Owen (1992) also considers another estimator, but for simplicity, we only work with (17).)

Unfortunately, the estimator in (17) is not implementable because each W_i depends on the unknown ξ_p, so we now modify (17) to account for this. For each $x \in \Re$, define $W_i(x) = I(\hat{Y}_i \le x)$, and let $\bar{W}_n(x) = n^{-1}\sum_{1\le i\le n} W_i(x) = \hat{F}_n(x)$. Then our estimator of the numerator ψ_p^2 in (11) and (15) is

$$\hat{\psi}^2_{p,n} = \frac{1}{2n}\sum_{j=1}^{d}\sum_{i=1}^{n}\left[W_i(\hat{\xi}_{p,n}) - N_jW_i(\hat{\xi}_{p,n})\right]^2 - (d-1)p(1-p) \tag{18}$$

with $N_jW_i(x) = W_m(x)$ if $\pi_j(i) < n$, where $\pi_j(m) = \pi_j(i)+1$, and $N_jW_i(x) = \bar{W}_n(x)$ if $\pi_j(i) = n$.

To handle the denominator in (11), we estimate $\lambda_p = \frac{d}{dp}F^{-1}(p)$ using the finite difference

$$\hat{\lambda}_{p,n} = [\hat{F}_n^{-1}(p+h_n) - \hat{F}_n^{-1}(p-h_n)]/(2h_n) \tag{19}$$

with bandwidth $h_n > 0$. The terms in the numerator of (19) are p_n-quantile estimators for $p_n = p \pm h_n$, which can be analyzed through perturbed Bahadur representations from (7). Thus, we can estimate η_p in (11) with $\hat{\eta}_{p,n} = \hat{\psi}_{p,n}\hat{\lambda}_{p,n}$, and we obtain an approximate $100(1-\alpha)\%$ CI for ξ_p as

$$C_{n,\mathrm{ssLHS}} = [\hat{\xi}_{p,n} \pm z_{\alpha/2}\hat{\eta}_{p,n}/\sqrt{n}]. \tag{20}$$

Define $\zeta(\mathbf{u}, x) = I(g(\mathbf{u}) \le x)$, and $\zeta_j(u_j, x)$ as the jth main effect of $\zeta(\mathbf{u}, x)$. The following result, proven in Appendix B, shows that $\hat{\psi}_{p,n}$ and $\hat{\lambda}_{p,n}$ consistently estimate ψ_p and λ_p, respectively, and the CI $C_{n,\mathrm{ssLHS}}$ is asymptotically valid.

THEOREM 2. *Under the assumptions of Theorem 1,* $\hat{\lambda}_{p,n} \Rightarrow \lambda_p$ *as* $n \to \infty$ *when the bandwidth in* (19) *is* $h_n = cn^{-1/2}$ *for any constant* $c > 0$. *Suppose that in addition, for each* $j = 1, 2, \ldots, d$, *and each* x *in a neighborhood* N *of* ξ_p, *the* j*th main effect* $\zeta_j(u_j, x)$ *satisfies a Lipschitz condition, i.e.,*

there exists a constant $c_0 > 0$ *such that* $|\zeta_j(u,x) - \zeta_j(v,x)| < c_0|u-v|$ *for every* $u, v \in [0,1]$. (21)

Then $\hat{\psi}^2_{p,n} \Rightarrow \psi_p^2$ *and* $P(\xi_p \in C_{n,ssLHS}) \to 1-\alpha$ *as* $n \to \infty$.

4.2. CIs Using srLHS

Rather than trying to consistently estimate η_p in (11), we now describe CIs that circumvent this issue. We build these intervals using replicated LHS with batching or sectioning. Let $b \geq 2$ be the number of batches. For each batch, a single LHS sample of size m is generated by letting $s = m$ in (5), and LHS samples across batches are generated independently. The overall sample size across all batches is $n = bm$. Because a single LHS sample is used in each batch, where we increase the batch size m, Section 1 referred to this method as single replicated LHS (srLHS). This differs from the rLHS approach proposed by Nakayama (2012), which increases the total number r of independent LHS samples, each with a *fixed* size m, and divides the r LHS samples into b batches. Section 1 called the latter technique multiple replicated LHS (mrLHS), and batching and sectioning with mrLHS are asymptotically valid when $r \to \infty$ with b and m fixed (Nakayama (2012)).

For srLHS, let $\tilde{Y}_{j,i}$, $i = 1, 2, \ldots, m$, be the m outputs in the jth batch, $j = 1, 2, \ldots, b$. Because the batches are independent, $\tilde{Y}_{j,i}$ and $\tilde{Y}_{j',i'}$ are independent for $j \neq j'$, but $\tilde{Y}_{j,i}$ and $\tilde{Y}_{j,i'}$ are dependent as they are in the same LHS sample j. Then we define the srLHS overall p-quantile estimator

$$\check{\xi}_{p,b,m} = \check{F}^{-1}_{b,m}(p), \quad \text{where} \quad \check{F}_{b,m}(y) = \frac{1}{b} \sum_{1 \leq j \leq b} \frac{1}{m} \sum_{1 \leq i \leq m} I(\tilde{Y}_{j,i} \leq y), \tag{22}$$

which is based on all $n = bm$ outputs. Let $\tilde{F}_{j,m}$ be the CDF estimator from batch j, i.e.,

$$\tilde{F}_{j,m}(y) = \frac{1}{m} \sum_{1 \leq i \leq m} I(\tilde{Y}_{j,i} \leq y). \tag{23}$$

The corresponding p-quantile estimator from batch j is $\tilde{\xi}_{p,j,m} = \tilde{F}^{-1}_{j,m}(p)$, and

$$\acute{\xi}_{p,b,m} = \frac{1}{b} \sum_{1 \leq j \leq b} \tilde{\xi}_{p,j,m} \tag{24}$$

is the srLHS batching p-quantile estimator. The sample variance of $\tilde{\xi}_{p,j,m}$, $j = 1, 2, \ldots, b$, is

$$\check{S}^2_{b,m,\text{batch}} = \frac{1}{b-1} \sum_{1 \leq j \leq b} (\tilde{\xi}_{p,j,m} - \acute{\xi}_{p,b,m})^2. \tag{25}$$

The two-sided $100(1-\alpha)\%$ CI for ξ_p using srLHS batching is

$$\check{C}_{b,m,\text{batch}} = [\acute{\xi}_{p,b,m} \pm t_{b-1,\alpha/2} \check{S}_{b,m,\text{batch}} / \sqrt{b}]. \tag{26}$$

As with SRS, the srLHS sectioning CI will usually have better coverage for small n than the batching CI (26) because of the bias of batching point estimator $\acute{\xi}_{p,b,m}$. Sectioning replaces $\acute{\xi}_{p,b,m}$ with the overall quantile estimator $\check{\xi}_{p,b,m}$ from (22), and instead of (25), we compute $\check{S}^2_{b,m,\text{sec}} = (1/(b-1))\sum_{1\le j\le b}(\tilde{\xi}_{p,j,m} - \check{\xi}_{p,b,m})^2$. Then the two-sided $100(1-\alpha)\%$ CI for ξ_p using srLHS sectioning is

$$\check{C}_{b,m,\text{sec}} = [\check{\xi}_{p,b,m} \pm t_{b-1,\alpha/2}\check{S}_{b,m,\text{sec}}/\sqrt{b}], \tag{27}$$

which is centered at a less-biased point estimator $\check{\xi}_{p,b,m}$ than the batching CI (26).

To establish the asymptotic validity of the srLHS sectioning method, we first prove that the srLHS overall quantile estimator has a weak Bahadur representation, which is given next. Its proof, in Appendix C, exploits Theorem 1 and the independence of the LHS samples across batches.

THEOREM 3. *Assume the same conditions as in Theorem 1, and for $p_n = p + cn^{-1/2}$ for any constant $c \in \Re$, consider the srLHS overall p_n-quantile estimator $\check{\xi}_{p_n,b,m}$ based on srLHS with overall sample size $n = bm$ comprising $b \ge 2$ independent LHS samples, each of size m. Then*

$$\check{\xi}_{p_n,b,m} = \xi^{\#}_{p_n} - [\check{F}_{b,m}(\xi_p) - p]/F'(\xi_p) + \check{R}_{b,m} \quad \text{with} \quad n^{1/2}\check{R}_{b,m} = b^{1/2}m^{1/2}\check{R}_{b,m} \Rightarrow 0$$

as $m \to \infty$ with $b \ge 2$ fixed, where $\xi^{\#}_{p_n}$ is defined in Theorem 1.

The next result, whose proof in Appendix D utilizes Theorem 3 for fixed $p_n = p$, verifies the asymptotic validity of the CI estimators by the srLHS approach.

THEOREM 4. *Under the same conditions as in Theorem 1, $P(\xi_p \in C) \to 1-\alpha$ as $m \to \infty$ with $b \ge 2$ fixed, where $C \in \{\check{C}_{b,m,batch}, \check{C}_{b,m,sec}\}$.*

4.3. Asymptotic Comparison of ssLHS, srLHS and mrLHS

We first use our various Bahadur representations to compare the ssLHS and srLHS quantile estimators. By (8), (9), and (11) (and assuming appropriate uniform integrability, e.g., p. 338 of Billingsley (1995)), we can approximate the variance of the ssLHS quantile estimator as

$$\text{Var}[\hat{\xi}_{p,n}] \approx \text{Var}[\hat{F}_n(\xi_p)]/[F'(\xi_p)]^2 \approx \psi_p^2/(n[F'(\xi_p)]^2) = \eta_p^2/n \tag{28}$$

for large ssLHS size n. For the srLHS overall quantile estimator, it follows from Theorem 3 with fixed $p_n = p$ that $\mathrm{Var}[\check{\xi}_{p,b,m}] \approx \mathrm{Var}[\check{F}_{b,m}(\xi_p)]/[F'(\xi_p)]^2$ for large LHS size m with $b \geq 2$ fixed. By (22) and (23), we see that $\check{F}_{b,m}(\xi_p)$ is the average of $\tilde{F}_{j,m}(\xi_p)$, $j = 1, 2, \ldots, b$, which are i.i.d., so $\mathrm{Var}[\check{F}_{b,m}(\xi_p)] = \mathrm{Var}[\tilde{F}_{j,m}(\xi_p)]/b$. When the LHS size m of each batch j is large, (9) implies $\mathrm{Var}[\tilde{F}_{j,m}(\xi_p)] \approx \psi_p^2/m$. We then see that the srLHS overall quantile estimator has variance $\mathrm{Var}[\check{\xi}_{p,b,m}] \approx \psi_p^2/(bm[F'(\xi_p)]^2) = \eta_p^2/(bm)$. Comparing this to (28), we see that when srLHS is implemented with each LHS size m large and a fixed number b of batches, ***the srLHS overall quantile estimator has roughly the same variance as the ssLHS quantile estimator*** with overall sample size $n = bm$.

A similar asymptotic analysis shows that the srLHS batching quantile estimator $\acute{\xi}_{p,b,m}$ in (24) also possesses about the same variance. But in general, the srLHS overall quantile estimator will typically have less bias than the srLHS batching quantile estimator, so the srLHS sectioning CI should have better coverage than the batching CI, even though they have roughly the same widths.

We further note that the ssLHS CI in (20) is based on a normal approximation, whereas the srLHS batching and sectioning CIs in (26) and (27), respectively, follow from a Student t limit. Hence, the batching and sectioning CIs asymptotically will be somewhat wider than the ssLHS CI, with the difference vanishing as $b \to \infty$. However, for a fixed overall sample size $n = bm$, choosing b large will necessitate small m, but the asymptotics for srLHS require m to be large.

The main difference of our srLHS methods and the mrLHS approaches in Nakayama (2012) lies in their asymptotic regimes. The mrLHS batching and sectioning CIs are asymptotically valid when the number r of independent LHS samples satisfies $r \to \infty$, with a fixed size m for each LHS sample and a fixed number $b \geq 2$ of batches, with each batch containing r/b independent LHS samples. Thus, mrLHS methods use rm total observations. In contrast, the srLHS CIs require a fixed number $b \geq 2$ of independent LHS samples, each of size $m \to \infty$. The total number of observations for srLHS is then bm. If the total sample size is fixed at n for both methods, srLHS allows larger LHS size than mrLHS. Hence, srLHS can reduce variance more than mrLHS because the variance of LHS estimators decreases as the LHS sample size grows. Our numerical results in Section 5 confirm this.

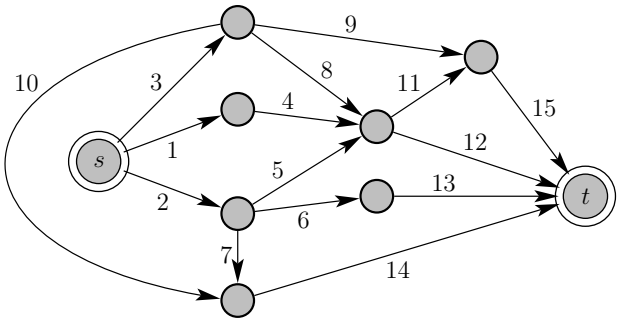

Figure 1 A stochastic activity network.

5. Numerical Results

We now present numerical results from constructing various CIs for quantiles of two stochastic models. The first is a stochastic activity network (SAN), previously studied in Juneja et al. (2007), Chu and Nakayama (2012b), and Nakayama (2014). The second model considers the sojourn time (waiting time plus service time) of customer $v = 5$ in an M/M/1 queue.

5.1. SAN Model

A SAN models the time to complete a project having activities with random durations and precedence constraints. Figure 1 displays a SAN with $d = 15$ activities, which correspond to the edges in the network. The length X_j of edge j is the time to complete activity j. The activity durations are independent exponential random variables, with activities $j \leq 8$, having mean 2, and the rest with mean 1. The network has $q = 10$ paths from nodes s to t, which are $B_1 = \{1, 4, 11, 15\}$, $B_2 = \{1, 4, 12\}$, $B_3 = \{2, 5, 11, 15\}$, $B_4 = \{2, 5, 12\}$, $B_5 = \{2, 6, 13\}$, $B_6 = \{2, 7, 14\}$, $B_7 = \{3, 8, 11, 15\}$, $B_8 = \{3, 8, 12\}$, $B_9 = \{3, 9, 15\}$, and $B_{10} = \{3, 10, 14\}$. Let $Y = \max_{k=1,2,\ldots,q} \sum_{j \in B_k} X_j$, which corresponds to the time to complete the project. Let F denote the CDF of Y, and we are interested in constructing confidence intervals for the p-quantile of F for different values of p using SRS and various versions of LHS.

To simulate the model using SRS, we generate $d = 15$ i.i.d. unif$[0, 1]$ random numbers, which we transform via inversion to get X_j, $j = 1, 2, \ldots, 15$, and then compute the output Y. We repeat this n independent times to obtain n i.i.d. outputs $Y_1, Y_2, \ldots, Y_n$. For LHS, we generate the (dependent)

outputs as in Section 3. From one SRS simulation with $n = 10^7$, we estimated the "true value" of the p-quantile as $\xi_p = 11.7659$ for $p = 0.8$ and $\xi_p = 15.3478$ for $p = 0.95$, which we used in our coverage experiments. For each nominal 90% CI C_n based on an overall sample size n, we ran 10^4 independent experiments to estimate the coverage $P(\xi_p \in C_n)$ and average half-width (AHW).

Table 1 presents results from constructing CIs when applying SRS, ssLHS, srLHS and mrLHS. (Dong and Nakayama (2014) present similar results for a smaller SAN with $d = 5$ activities using batching and sectioning only.) For SRS, we constructed the finite-difference CI from Section 2.1, and the batching and sectioning CIs from Section 2.2. These correspond to the columns labeled "FD", "Batch", and "Section". All batching and sectioning CIs use $b = 10$ batches. (Experiments in Nakayama (2014) with $b = 10$ and $b = 20$ using SRS, importance sampling and control variates show that $b = 20$ often results in poorer coverage than $b = 10$ for the same overall sample size n, especially when n is small or $p \approx 1$.) The ssLHS CI is from (20), which combines the modified Owen (1992) estimator (18) and the finite difference (19); this column is also labeled "MOFD" for "modified Owen-FD". For srLHS, the batching and sectioning CIs in (26) and (27), respectively, again have $b = 10$ independent batches, with each batch consisting of a single LHS sample of size $m = n/b$. We construct mrLHS CIs with a fixed LHS size $m = 10$, and the number of independent LHS samples as $r = n/m$, which are divided into $b = 10$ batches; thus, the number r of independent LHS samples grows with n, but each LHS sample size remains fixed at $m = 10$. When $n = 100$, the performance of srLHS and mrLHS are comparable as the two methods are then identical, but as n grows, the advantage of srLHS over mrLHS appears and increases with n.

The FD estimators used bandwidth $h_n = n^{-1/2}$. For small n and $p \approx 1$, we can have $p + h_n \geq 1$, which causes problems for the FD because the inverse of the estimated CDF is evaluated outside its domain $(0, 1)$; e.g., see (19). In this case, we replace $p + h_n$ and $p - h_n$ in the FD with $1 - (1-p)/10$ and $2p - 1 + (1-p)/10$, respectively, the latter chosen so that the two are symmetric about p. Some sort of adjustment like this needs to be made for the finite difference to be well-defined.

We first analyze the coverage in Table 1 of the CIs. For large n, the coverages of all the CIs are close to nominal, demonstrating their asymptotic validity. The coverages for $p = 0.95$ converge

Table 1 Coverages (and average half-widths) of nominal 90% confidence intervals for the p-quantile ($p=0.5$ and $p=0.8$) of the SAN model using FD, modified Owen-FD (MOFD), batching and sectioning methods with SRS, ssLHS, srLHS ($b=10$, $m=n/b$), and mrLHS ($b=10$, $m=10$).

n	$p=0.8$							
	SRS			ssLHS	srLHS		mrLHS	
	FD	Batch	Section	MOFD	Batch	Section	Batch	Section
100	0.902 (1.038)	0.666 (0.890)	0.892 (0.970)	0.898 (0.844)	0.629 (0.734)	0.892 (0.812)	0.617 (0.728)	0.886 (0.806)
400	0.883 (0.463)	0.837 (0.484)	0.902 (0.498)	0.888 (0.327)	0.803 (0.344)	0.904 (0.361)	0.825 (0.405)	0.902 (0.419)
1600	0.882 (0.226)	0.889 (0.249)	0.905 (0.252)	0.883 (0.153)	0.867 (0.167)	0.904 (0.170)	0.877 (0.207)	0.897 (0.209)
6400	0.893 (0.115)	0.895 (0.125)	0.901 (0.126)	0.895 (0.077)	0.892 (0.084)	0.902 (0.084)	0.899 (0.104)	0.902 (0.105)
n	$p=0.95$							
	SRS			ssLHS	srLHS		mrLHS	
	FD	Batch	Section	MOFD	Batch	Section	Batch	Section
100	0.943 (2.259)	0.854 (1.674)	0.860 (1.724)	0.954 (2.576)	0.879 (1.587)	0.864 (1.635)	0.874 (1.580)	0.861 (1.626)
400	0.896 (0.926)	0.672 (0.841)	0.892 (0.913)	0.928 (0.797)	0.665 (0.719)	0.888 (0.785)	0.677 (0.788)	0.888 (0.856)
1600	0.890 (0.443)	0.833 (0.456)	0.897 (0.470)	0.909 (0.342)	0.820 (0.355)	0.897 (0.368)	0.830 (0.432)	0.896 (0.445)
6400	0.895 (0.218)	0.886 (0.235)	0.905 (0.237)	0.897 (0.163)	0.873 (0.174)	0.903 (0.177)	0.884 (0.223)	0.900 (0.225)

more slowly than for $p = 0.8$. In general, sectioning has better coverage than batching, especially for small n and the larger p, because the sectioning CI is centered at a less-biased estimator. Also, ssLHS and sectioning with srLHS and mrLHS have comparable coverages for $p = 0.8$.

In terms of AHW for large n, all of the various LHS methods outperform SRS. The AHW ratio (AHWR) of SRS with FD to ssLHS for $p = 0.8$ is $0.115/0.077 \approx 1.49$, which means that the sample size of SRS with FD would have to be increased by a factor of $(0.115/0.077)^2 \approx 2.2$ for its CI's AHW to approximately equal that of ssLHS. While the AHWR of SRS with FD to ssLHS is smaller ($0.218/0.163 \approx 1.34$) for $p = 0.95$, this still corresponds to SRS with FD requiring about an 80% larger sample size than ssLHS to get the same AHW. The srLHS AHW is slightly greater (about 9%) than for ssLHS, and the difference roughly matches the increase in the CIs' 0.95-critical points from a Student t distribution with $b - 1 = 9$ degrees of freedom to a normal (1.833 to 1.645); see the asymptotic analysis in Section 4.3. Comparing mrLHS to srLHS, the AHWR with batching and sectioning is around 1.25 for both $p = 0.8$ and $p = 0.95$, demonstrating the additional variance reduction that srLHS obtains over mrLHS from having a larger LHS sample size. The AHWR of SRS to mrLHS is approximately 1.2 for $p = 0.8$ and 1.05 for $p = 0.95$, so mrLHS may produce only a modest variance reduction compared to SRS for extreme quantiles. This illustrates a deficiency of mrLHS and further motivates the use of srLHS and ssLHS.

5.2. Sojourn Time in M/M/1 Queue

We also ran numerical experiments to estimate the p-quantile ξ_p of the sojourn time of customer $v = 5$ in an M/M/1 queue, where the first customer arrives at time 0 to an empty system. For $i \geq 1$, let A_i denote the time that elapses between the arrivals of customers i and $i + 1$, and $A_1, A_2, \ldots$ are i.i.d. exponential random variables with rate 1. Let S_i be the service time of customer i, and $S_1, S_2, \ldots$ are i.i.d. exponentials (independent of the interarrival times) with rate 10/9. In our experiments we set $p = 0.5$, 0.8, and 0.95. We implemented algorithms in Kaczynski et al. (2012) to numerically compute ξ_p, obtaining the true values $\xi_{0.5} = 1.670$, $\xi_{0.8} = 3.370$, and $\xi_{0.95} = 5.515$.

As with the SAN experiments in Section 5.1, our M/M/1 experiments apply SRS, ssLHS, srLHS, and mrLHS, using FD, MOFD, batching and sectioning. For SRS, we generate an i.i.d. uniform grid as in (2). When applying some form of LHS, we generate an LHS sample (of dependent uniforms) as in (4). In both (2) and (4), each row has $d = 2v - 1$ entries, which are used to generate by inversion the interarrival times $A_1, A_2, \ldots, A_{v-1}$, and the service times $S_1, S_2, \ldots, S_v$. We then compute the waiting time W'_j of customer $j \geq 2$ via Lindley's equation: $W'_j = \max(W'_{j-1} + S_{j-1} - A_{j-1}, 0)$, with $W'_1 = 0$. The sojourn time of customer v is then $W'_v + S_v$.

Table 2 (resp., 3) gives the M/M/1 results for $p = 0.5$ and 0.8 (resp., 0.95) from 10^4 independent experiments to estimate coverage and AHW for different overall sample sizes n. As with the SAN results, we see that for the M/M/1 model, coverages for all p appear to converge to the nominal level 0.9 as n gets large. For $p = 0.5$ and $n = 6400$, the AHWR for SRS with FD over ssLHS is $0.049/0.027 \approx 1.59$, which means that the sample size of SRS with FD would have to be increased by about a factor of $(0.049/0.027)^2 \approx 2.5$ for its CI's AHW to equal that of ssLHS. The AHWR of SRS with FD to ssLHS decreases to $0.069/0.045 \approx 1.53$ and $0.131/0.102 \approx 1.28$ for $p = 0.8$ and $p = 0.95$, respectively, but the latter still corresponds to SRS with FD requiring about a 65% larger sample size than ssLHS to get roughly the same AHW.

The results also show the superiority in AHW of ssLHS and srLHS over mrLHS. For $p = 0.95$ and $n = 6400$, we see that the AHWR of mrLHS with sectioning to ssLHS is $0.125/0.102 \approx 1.23$, which corresponds to mrLHS needing roughly a 50% larger sample size n than ssLHS to produce about the same AHW. Also, for $n = 6400$, the AHWRs of mrLHS to srLHS (both for sectioning) for $p = 0.5$, 0.8, and 0.95 are 1.07, 1.10, and 1.15, respectively. Thus, the advantage of srLHS over mrLHS can increase as p becomes more extreme, where for $p = 0.95$, mrLHS would require about a 32% larger sample size than that for srLHS to obtain approximately the same AHW.

6. Concluding Remarks

LHS is perhaps the most widely applied VRT in certain fields, such as nuclear engineering, because of its intuitive appeal and ease of implementation. However, LHS has not been adopted for performing safety and uncertainty analyses of nuclear facilities, which require CIs for quantiles; currently,

Table 2 Coverages (and average half-widths) of nominal 90% confidence intervals for the p-quantile ($p=0.5$ and $p=0.8$) of the sojourn time of customer $v=5$ in an M/M/1 queue using FD, modified Owen-FD (MOFD), batching and sectioning methods with SRS, ssLHS, srLHS ($b=10$, $m=n/b$), and mrLHS ($b=10$, $m=10$).

n	$p=0.5$							
	SRS			ssLHS	srLHS		mrLHS	
	FD	Batch	Section	MOFD	Batch	Section	Batch	Section
100	0.876 (0.342)	0.894 (0.347)	0.890 (0.354)	0.762 (0.195)	0.901 (0.215)	0.881 (0.220)	0.899 (0.215)	0.880 (0.220)
400	0.901 (0.176)	0.900 (0.182)	0.903 (0.185)	0.872 (0.108)	0.895 (0.114)	0.897 (0.115)	0.905 (0.118)	0.900 (0.120)
1600	0.897 (0.085)	0.898 (0.093)	0.897 (0.093)	0.886 (0.054)	0.896 (0.058)	0.897 (0.059)	0.902 (0.061)	0.902 (0.062)
6400	0.902 (0.043)	0.902 (0.046)	0.901 (0.046)	0.895 (0.027)	0.901 (0.029)	0.903 (0.029)	0.898 (0.031)	0.898 (0.031)
n	$p=0.8$							
	CMC			ssLHS	srLHS		mrLHS	
	FD	Batch	Section	MOFD	Batch	Section	Batch	Section
100	0.894 (0.581)	0.818 (0.527)	0.878 (0.548)	0.813 (0.372)	0.792 (0.363)	0.859 (0.382)	0.779 (0.363)	0.857 (0.382)
400	0.908 (0.289)	0.877 (0.291)	0.895 (0.296)	0.890 (0.189)	0.869 (0.191)	0.900 (0.196)	0.868 (0.210)	0.891 (0.215)
1600	0.894 (0.139)	0.899 (0.150)	0.902 (0.151)	0.893 (0.091)	0.888 (0.097)	0.900 (0.099)	0.895 (0.109)	0.901 (0.110)
6400	0.894 (0.069)	0.901 (0.076)	0.903 (0.076)	0.896 (0.045)	0.901 (0.049)	0.904 (0.050)	0.897 (0.055)	0.899 (0.055)

Table 3 Coverages (and average half-widths) of nominal 90% confidence intervals for the p-quantile ($p = 0.95$) of the sojourn time of customer $v = 5$ in an M/M/1 queue using FD, modified Owen-FD (MOFD), batching and sectioning methods with SRS, ssLHS, srLHS ($b = 10$, $m = n/b$), and mrLHS ($b = 10$, $m = 10$).

n	$p = 0.95$							
	SRS			ssLHS	srLHS		mrLHS	
	FD	Batch	Section	MOFD	Batch	Section	Batch	Section
100	0.685	0.409	0.823	0.908	0.427	0.806	0.420	0.798
	(0.688)	(0.757)	(0.895)	(1.278)	(0.625)	(0.741)	(0.624)	(0.742)
400	0.953	0.708	0.886	0.972	0.694	0.884	0.710	0.882
	(0.673)	(0.497)	(0.534)	(0.608)	(0.401)	(0.434)	(0.436)	(0.468)
1600	0.920	0.845	0.900	0.918	0.837	0.894	0.845	0.897
	(0.279)	(0.271)	(0.279)	(0.219)	(0.212)	(0.219)	(0.241)	(0.247)
6400	0.901	0.882	0.895	0.900	0.884	0.905	0.888	0.901
	(0.131)	(0.138)	(0.140)	(0.102)	(0.107)	(0.109)	(0.124)	(0.125)

only SRS is used. An initial investigation on applying mrLHS for safety and uncertainty analyses of nuclear facilities appears in Grabaskas et al. (2012), using an rLHS CI based on a finite difference developed by Nakayama (2011). But this rLHS method and mrLHS of Nakayama (2012) do not obtain the full variance reduction possible by LHS since they both require a fixed LHS size.

Our current paper developed LHS CIs for quantiles that maximize this VRT's variance reduction by allowing the LHS size to grow large. We considered ssLHS and srLHS, proving Bahadur representations for quantile estimators for both approaches. We leveraged these results to obtain asymptotically valid CIs for ssLHS and srLHS. For the ssLHS CI, we developed a consistent estimator of the asymptotic variance from the ssLHS quantile estimator's CLT; the previous work of Avramidis and Wilson (1998) on ssLHS quantile estimation did not address the issues of estimating the asymptotic variance nor of constructing a CI. We also use batching and sectioning to build CIs with srLHS, and these CIs have roughly the same average width as the ssLHS CI and are

shorter than mrLHS CIs (Nakayama (2012)). The improvement of srLHS over mrLHS results from srLHS allowing large LHS sizes, whereas mrLHS permits only fixed LHS sizes. Sectioning produces CIs with better coverage than batching, especially for small overall sample sizes, because the former centers its CI at a less-biased point estimator. Our methods have the potential for significant impact in applications such as safety and uncertainty analyses of nuclear power plants. Moreover, other areas, such as finance and service industries, can also greatly benefit from our results.

Appendix A: Proof of Theorem 1

For two functions f_1 and f_2, we write $f_1(n) = O(f_2(n))$ to mean that there exist constants c_1 and n_1 such that $|f_1(n)| \le |c_1 f_2(n)|$ for all $n \ge n_1$. Chu and Nakayama (2012a) establish the following three conditions, which, if satisfied, are sufficient to ensure that a weak Bahadur representation (7) holds for a quantile estimator $F_n^{-1}(p_n)$ with F_n a CDF estimator and $p_n = p + O(n^{-1/2})$ as $n \to \infty$.

Condition A1 *$P(T_n) \to 1$ as $n \to \infty$, where T_n is the event that the CDF estimator $F_n(y)$ is monotonically increasing in y.*

Condition A2 *For each $a_n = O(n^{-1/2})$, $D_n \equiv n^{1/2}[(F_n(\xi_p + a_n) - F_n(\xi_p)) - (F(\xi_p + a_n) - F(\xi_p))] \Rightarrow 0$ as $n \to \infty$.*

Condition A3 *$n^{1/2}[F_n(\xi_p) - F(\xi_p)] \Rightarrow N(0, \psi_p^2)$ as $n \to \infty$ for some $0 < \psi_p < \infty$.*

We now show that the ssLHS CDF estimator $\hat{F}_n$ in (6) satisfies Conditions A1, A2, and A3. In (6), each $I(\hat{Y}_i \le y)$ is monotonically increasing in y, so $\hat{F}_n(y)$ is also monotonically increasing in y; i.e., Condition A1 holds. Also, (9) implies Condition A3. For Condition A2, our theorem only considers perturbations $p_n = p + c/\sqrt{n}$ for a constant c, and a careful examination of the proof of Theorem 3.1(i) of Chu and Nakayama (2012a) reveals that in this case, it suffices to instead show

Condition A4 *For each $a_n = n^{-1/2}t$ with $t \in \Re$, $D_n \Rightarrow 0$ as $n \to \infty$.*

We prove this through Lemmas 1 and 2 below for a fixed $t \ge 0$. (The case $t < 0$ is similar.)

Recall $\chi(\mathbf{u}) = I(g(\mathbf{u}) \le \xi_p)$ for $\mathbf{u} \in [0,1]^d$, and its jth main effect $\chi_j(\cdot)$, additive part $\chi_{\rm add}(\cdot)$, and residual from additivity $\chi_{\rm res}(\cdot)$, as in (12)–(14). Avramidis and Wilson (1996) define $\chi^{(n)}(\mathbf{u}) = I(g(\mathbf{u}) \le \xi_p + n^{-1/2}t)$, its jth main effect $\chi_j^{(n)}(\cdot)$, additive part $\chi_{\rm add}^{(n)}(\cdot)$, and residual from additivity $\chi_{\rm res}^{(n)}(\cdot)$. Define $\omega^{(n)}(\mathbf{u}) = \chi^{(n)}(\mathbf{u}) - \chi(\mathbf{u})$, so $\omega^{(n)} \in \{0, 1\}$ when $t \ge 0$. It is easy to verify that its jth main effect $\omega_j^{(n)}(u_j) = \chi_j^{(n)}(u_j) - \chi_j(u_j)$, additive part $\omega_{\rm add}^{(n)}(\mathbf{u}) = \chi_{\rm add}^{(n)}(\mathbf{u}) - \chi_{\rm add}(\mathbf{u})$, and residual from additivity $\omega_{\rm res}^{(n)}(\mathbf{u}) = \chi_{\rm res}^{(n)}(\mathbf{u}) - \chi_{\rm res}(\mathbf{u})$. For a d-vector $\mathbf{U}$ of i.i.d. unif$[0,1]$ random variables,

$$\mathrm{E}[\omega^{(n)}(\mathbf{U})] = P(\xi_p < Y \le \xi_p + t/\sqrt{n}) \to 0 \tag{29}$$

as $n \to \infty$ because we assumed F' exists at ξ_p. Let $\mathcal{B}(i,n)$ be the subinterval $[(i-1)/n,\, i/n]$ of $[0,1]$ for $i = 1, 2, \ldots, n$.

LEMMA 1. *Under the conditions of Theorem 1,* $\lim_{n\to\infty} n \sum_{1\le i\le n} \left[\int_{\mathcal{B}(i,n)} \omega_j^{(n)}(u)\, du\right]^2 = 0$ *for* $j = 1, \ldots, d$.

Proof of Lemma 1 We will establish Lemma 1 by extending the proof of Lemma 4 in Avramidis and Wilson (1996). Fix $t \ge 0$; we can handle $t < 0$ similarly. Choose n_0 sufficiently large so that $\xi_p + n_0^{-1/2} t \in \mathcal{N}(\xi_p)$, and fix $\nu \ge n_0$. Note that $\omega_j^{(\nu)}(\cdot) = \chi_j^{(\nu)}(\cdot) - \chi_j(\cdot)$ has a finite number of discontinuities over $[0,1]$ because Avramidis and Wilson (1996) have proven in their Lemma 4 that under conditions CC1 and CC2, both $\chi_j^{(\nu)}(\cdot)$ and $\chi_j(\cdot)$ also have that property. Using the same arguments that Avramidis and Wilson (1996) applied to verify their equation (90), we can show that for each fixed $\nu \ge n_0$,

$$\lim_{n\to\infty} n \sum_{1\le i\le n} \left[\int_{\mathcal{B}(i,n)} (\chi_j^{(\nu)}(u) - \chi_j(u))\, du\right]^2 = \int_{[0,1]} (\chi_j^{(\nu)}(u) - \chi_j(u))^2\, du. \tag{30}$$

For all $n \ge \nu$ and $u \in [0,1]$,

$$\chi_j^{(\nu)}(u) - \chi_j(u) \ge \chi_j^{(n)}(u) - \chi_j(u) \ge \chi_j(u) - \chi_j(u) \tag{31}$$

because $t \ge 0$. By the fact that $\lim_{\nu\to\infty} \chi_j^{(\nu)}(u) = \chi_j(u)$ for all $u \in [0,1]$ (see equation (87) in Avramidis and Wilson (1996)) and the bounded convergence theorem (BCT; Theorem 16.5 of Billingsley (1995)), we then have that (30) and (31) imply

$$\begin{aligned}
0 &= \int_{[0,1]} (\chi_j(u) - \chi_j(u))^2\, du = \lim_{\nu\to\infty} \int_{[0,1]} \left(\chi_j^{(\nu)}(u) - \chi_j(u)\right)^2 du \\
&= \lim_{\nu\to\infty} \lim_{n\to\infty} n \sum_{1\le i\le n} \left[\int_{\mathcal{B}(i,n)} \left(\chi_j^{(\nu)}(u) - \chi_j(u)\right) du\right]^2 \\
&\ge \limsup_{n\to\infty} n \sum_{1\le i\le n} \left[\int_{\mathcal{B}(i,n)} \left(\chi_j^{(n)}(u) - \chi_j(u)\right) du\right]^2 \ge \liminf_{n\to\infty} n \sum_{1\le i\le n} \left[\int_{\mathcal{B}(i,n)} \left(\chi_j^{(n)}(u) - \chi_j(u)\right) du\right]^2 \\
&\ge \lim_{n\to\infty} n \sum_{1\le i\le n} \left[\int_{\mathcal{B}(i,n)} \left(\chi_j(u) - \chi_j(u)\right) du\right]^2 = \int_{[0,1]} (\chi_j(u) - \chi_j(u))^2\, du = 0.
\end{aligned}$$

Therefore, Lemma 1 holds because

$$\lim_{n\to\infty} n \sum_{1\le i\le n} \left[\int_{\mathcal{B}(i,n)} \omega_j^{(n)}(u)\, du\right]^2 = \lim_{n\to\infty} n \sum_{1\le i\le n} \left\{\left[\int_{\mathcal{B}(i,n)} \left(\chi_j^{(n)}(u) - \chi_j(u)\right) du\right]^2\right\} = 0. \qquad \square$$

LEMMA 2. *Under the conditions of Theorem 1, the ssLHS CDF estimator* $\hat{F}_n$ *satisfies*

$$\hat{D}_n \equiv n^{1/2} \left[\left(\hat{F}_n(\xi_p + n^{-1/2}t) - \hat{F}_n(\xi_p)\right) - \left(F(\xi_p + n^{-1/2}t) - F(\xi_p)\right)\right] \Rightarrow 0 \tag{32}$$

as $n \to \infty$ *for each* $t \in \Re$, *so Condition A4 holds for* $a_n = n^{-1/2}t$, *as required for Theorem 1.*

Proof of Lemma 2 Fix $t \geq 0$. (The argument for $t < 0$ is analogous.) By definition, $\hat{F}_n(\xi_p + n^{-1/2}t) - \hat{F}_n(\xi_p) = n^{-1}\sum_{1\leq i\leq n}\omega^{(n)}(\mathbf{U}^{(i)})$, where $\mathbf{U}^{(i)}$, $i = 1, 2, \ldots, n$, are the input variables of a Latin hypercube sample; i.e., $\mathbf{U}^{(i)} = (V_{i,1}, V_{i,2}, \ldots, V_{i,d})$ is the ith row from (4). By decomposing $\omega^{(n)}(\cdot)$ into its additive and residual parts and using (14), we can write $\hat{D}_n$ from (32) as

$$\hat{D}_n = A_n + B_n, \tag{33}$$

where $A_n = n^{-1/2}\sum_{i=1}^{n}\left[\omega^{(n)}_{\text{add}}(\mathbf{U}^{(i)}) - \left(F(\xi_p + n^{-1/2}t) - F(\xi_p)\right)\right]$ and $B_n = n^{-1/2}\sum_{i=1}^{n}\omega^{(n)}_{\text{res}}(\mathbf{U}^{(i)})$. We will prove Lemma 2 by showing that both A_n and B_n weakly vanish as $n \to \infty$ via some arguments similar to those in the proof of Lemma 6 of Avramidis and Wilson (1996).

Let $\mathrm{E}_{\text{LH}}[\cdot]$, $\mathrm{Var}_{\text{LH}}[\cdot]$ and $\mathrm{Cov}_{\text{LH}}[\cdot]$ stand for the expectation, variance and covariance, respectively, under LHS. We also continue to use $\mathrm{E}[\cdot]$ and $\mathrm{Var}[\cdot]$ when considering quantities that do not depend on LHS. To verify that $A_n \Rightarrow 0$, it suffices to show that $v_n \equiv \mathrm{Var}_{\text{LH}}[A_n] \to 0$ as $n \to \infty$ by Chebyshev's inequality because $\mathrm{E}_{\text{LH}}[A_n] = 0$. We start by noting that

$$v_n = \frac{1}{n}\mathrm{Var}_{\text{LH}}\left[\sum_{i=1}^{n}\omega^{(n)}_{\text{add}}(\mathbf{U}^{(i)})\right] = \mathrm{Var}[\omega^{(n)}_{\text{add}}(\mathbf{U})] + (n-1)\mathrm{Cov}_{\text{LH}}[\omega^{(n)}_{\text{add}}(\mathbf{U}^{(1)}), \omega^{(n)}_{\text{add}}(\mathbf{U}^{(2)})] \tag{34}$$

because $\omega^{(n)}_{\text{add}}(\mathbf{U}^{(i)})$, $i = 1, 2, \ldots, n$, are exchangeable. We first prove $\mathrm{Var}[\omega^{(n)}_{\text{add}}(\mathbf{U})] = o(1)$, where for two functions f_1 and f_2, we write $f_1(n) = o(f_2(n))$ to mean that $f_1(n)/f_2(n) \to 0$ as $n \to \infty$. Equation (100) of Avramidis and Wilson (1996) establishes that $\mathrm{E}[\chi^{(n)}_{\text{add}}(\mathbf{U})] = \mathrm{E}[\chi_{\text{add}}(\mathbf{U})] + o(1)$ and $\mathrm{Var}[\chi^{(n)}_{\text{add}}(\mathbf{U})] = \mathrm{Var}[\chi_{\text{add}}(\mathbf{U})] + o(1)$ as $n \to \infty$. Thus, because $\omega^{(n)}_{\text{add}}(\mathbf{u}) = \chi^{(n)}_{\text{add}}(\mathbf{u}) - \chi_{\text{add}}(\mathbf{u})$, we have

$$\mathrm{E}[\omega^{(n)}_{\text{add}}(\mathbf{U})] = o(1) \tag{35}$$

as $n \to \infty$. Also, (13) and the independent components of $\mathbf{U} = (U_1, U_2, \ldots, U_d)$ imply

$$\begin{aligned} 0 \leq \mathrm{Var}[\omega^{(n)}_{\text{add}}(\mathbf{U})] &= \sum_{1\leq j\leq d}\mathrm{Var}[\omega^{(n)}_j(U_j)] = \sum_{1\leq j\leq d}\left(\mathrm{E}[(\omega^{(n)}_j(U_j))^2] - \mathrm{E}^2[\omega^{(n)}_j(U_j)]\right) \\ &\leq \sum_{1\leq j\leq d}\left(\mathrm{E}[\omega^{(n)}_j(U_j)] - \mathrm{E}^2[\omega^{(n)}_j(U_j)]\right) \end{aligned}$$

because $0 \leq \omega^{(n)}_j(\cdot) \leq 1$. Thus, $\mathrm{E}[\omega^{(n)}_j(U_j)] = \mathrm{E}[\omega^{(n)}(\mathbf{U})] \to 0$ as $n \to \infty$ by (29) leads to

$$\mathrm{Var}[\omega^{(n)}_{\text{add}}(\mathbf{U})] = o(1). \tag{36}$$

Recalling (34), we next show that $c_n \equiv \mathrm{Cov}_{\text{LH}}[\omega^{(n)}_{\text{add}}(\mathbf{U}^{(1)}), \omega^{(n)}_{\text{add}}(\mathbf{U}^{(2)})] = o(n^{-1})$. By an analogue of equation (A.3) of Stein (1987) and the uniform boundedness of $\omega^{(n)}_{\text{add}}(\mathbf{u})$ in n and $\mathbf{u}$, we have

$$c_n = dn^{-1}\mathrm{E}^2\left[\omega^{(n)}_{\text{add}}(\mathbf{U})\right] - n^{-1}\sum_{1\leq j\leq d}\left\{n\sum_{1\leq i\leq n}\left[\int_{\mathcal{B}(i,n)}\omega^{(n)}_j(u)\,du\right]^2\right\} + o(n^{-1}). \tag{37}$$

By (35), we have that $dn^{-1}\mathrm{E}^2\left[\omega_{\mathrm{add}}^{(n)}(\mathbf{U})\right] = o(n^{-1})$. In addition, Lemma 1 implies the quantity in the braces on the right side of (37) is $o(1)$ for each $j = 1, 2, \ldots, d$, so the entire second term in (37) is $o(n^{-1})$ because d is fixed. Thus, all three terms in (37) are $o(n^{-1})$, so $c_n = o(n^{-1})$ as $n \to \infty$. Putting this and (36) into (34) yields $\lim_{n\to\infty} v_n = 0$, so $A_n \Rightarrow 0$ as $n \to \infty$.

Now we prove that B_n from (33) vanishes as $n \to \infty$. Since $\omega_{\mathrm{res}}^{(n)}(\mathbf{u})$ are bounded uniformly in n and $\mathbf{u}$, equation (103) in Avramidis and Wilson (1996) still applies to our B_n, i.e., for each $q \geq 1$,

$$\mathrm{E}_{\mathrm{LH}}[B_n^q] = \mathrm{E}_{\mathrm{IR}}[B_n^q] + o(1) \tag{38}$$

as $n \to \infty$, where $\mathrm{E}_{\mathrm{IR}}[\cdot]$ denotes expectation when using independent replications (IR); i.e., when $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(n)}$ are independent vectors of independent uniforms. Note that $\mathrm{E}_{\mathrm{LH}}[B_n] = 0$ since

$$\mathrm{E}[\omega_{\mathrm{res}}^{(n)}(\mathbf{U})] = \mathrm{E}[\omega^{(n)}(\mathbf{U})] - \mathrm{E}[\omega_{\mathrm{add}}^{(n)}(\mathbf{U})] = 0. \tag{39}$$

It then suffices to prove $\lim_{n\to\infty} \mathrm{Var}_{\mathrm{LH}}[B_n] = 0$. By (38) with $q = 2$, we have that

$$\begin{aligned}\mathrm{Var}_{\mathrm{LH}}[B_n] &= \mathrm{E}_{\mathrm{LH}}[B_n^2] = \mathrm{E}_{\mathrm{IR}}[B_n^2] + o(1)\\ &= \frac{1}{n}\mathrm{E}_{\mathrm{IR}}\left[\sum_{1\leq i\leq n}(\omega_{\mathrm{res}}^{(n)}(\mathbf{U}^{(i)}))^2 + \sum_{1\leq i\leq n}\sum_{\substack{1\leq i'\leq n:\\ i'\neq i}}\omega_{\mathrm{res}}^{(n)}(\mathbf{U}^{(i)})\,\omega_{\mathrm{res}}^{(n)}(\mathbf{U}^{(i')})\right] + o(1).\end{aligned} \tag{40}$$

Under IR, we have that $\mathrm{E}_{\mathrm{IR}}[\omega_{\mathrm{res}}^{(n)}(\mathbf{U}^{(i)})\,\omega_{\mathrm{res}}^{(n)}(\mathbf{U}^{(i')})] = \mathrm{E}[\omega_{\mathrm{res}}^{(n)}(\mathbf{U}^{(i)})]\,\mathrm{E}[\omega_{\mathrm{res}}^{(n)}(\mathbf{U}^{(i')})] = 0$ by (39). Let $\mu_{\omega,n} = \mathrm{E}[\omega^{(n)}(\mathbf{U})] = \mathrm{E}[\omega_{\mathrm{add}}^{(n)}(\mathbf{U})] = \mathrm{E}[\omega_j^{(n)}(U_j))]$. Thus, (40) and (39) yield

$$\begin{aligned}\mathrm{Var}_{\mathrm{LH}}[B_n] &= \mathrm{E}\left[(\omega_{\mathrm{res}}^{(n)}(\mathbf{U}))^2\right] + o(1) = \mathrm{Var}\left[\omega_{\mathrm{res}}^{(n)}(\mathbf{U})\right] + \mathrm{E}^2\left[\omega_{\mathrm{res}}^{(n)}(\mathbf{U})\right] + o(1)\\ &= \mathrm{Var}\Big[\omega^{(n)}(\mathbf{U}) - \omega_{\mathrm{add}}^{(n)}(\mathbf{U})\Big] + o(1) = \mathrm{Var}\Big[\omega^{(n)}(\mathbf{U}) - \sum_{1\leq j\leq d}\omega_j^{(n)}(U_j) + (d-1)\mu_{\omega,n}\Big] + o(1)\\ &= \mathrm{Var}[\omega^{(n)}(\mathbf{U})] + \sum_{1\leq j\leq d}\mathrm{Var}[\omega_j^{(n)}(U_j)] - \sum_{1\leq j\leq d}2\mathrm{Cov}[\omega^{(n)}(\mathbf{U}), \omega_j^{(n)}(U_j)] + o(1)\\ &= \mathrm{Var}[\omega^{(n)}(\mathbf{U})] - \sum_{1\leq j\leq d}\mathrm{Var}[\omega_j^{(n)}(U_j)] + o(1),\end{aligned} \tag{41}$$

where the last line follows from the fact that

$$\begin{aligned}&\mathrm{Cov}[\omega^{(n)}(\mathbf{U}), \omega_j^{(n)}(U_j)] = \mathrm{E}[\omega^{(n)}(\mathbf{U})\,\omega_j^{(n)}(U_j)] - \mu_{\omega,n}^2 = \mathrm{E}[\mathrm{E}[\omega^{(n)}(\mathbf{U})\,\omega_j^{(n)}(U_j)\,|\,U_j]] - \mu_{\omega,n}^2\\ &= \mathrm{E}[\omega_j^{(n)}(U_j)\,\mathrm{E}[\omega^{(n)}(\mathbf{U})\,|\,U_j]] - \mu_{\omega,n}^2 = \mathrm{E}[(\omega_j^{(n)}(U_j))^2] - \mu_{\omega,n}^2 = \mathrm{Var}[\omega_j^{(n)}(U_j)].\end{aligned}$$

In addition, because $[\omega^{(n)}(\mathbf{U})]^2 = \omega^{(n)}(\mathbf{U})$ as it is binary, (41) implies

$$\begin{aligned}\mathrm{Var}_{\mathrm{LH}}[B_n] &= \Big(\mathrm{E}[\omega^{(n)}(\mathbf{U})] - \mathrm{E}^2[\omega^{(n)}(\mathbf{U})]\Big) - \sum_{1\leq j\leq d}\Big(\mathrm{E}[(\omega_j^{(n)}(U_j))^2] - \mathrm{E}^2[\omega_j^{(n)}(U_j)]\Big) + o(1)\\ &= (\mu_{\omega,n} + (d-1)\mu_{\omega,n}^2) - \sum_{1\leq j\leq d}\mathrm{E}[(\omega_j^{(n)}(U_j))^2] + o(1).\end{aligned} \tag{42}$$

Because $0 \le \omega_j^{(n)}(U_j) \le 1$, we see that $0 \le \mathrm{E}[(\omega_j^{(n)}(U_j))^2] \le \mathrm{E}[\omega_j^{(n)}(U_j)] = \mathrm{E}[\omega^{(n)}(\mathbf{U})] = \mu_{\omega,n} \to 0$ as $n \to \infty$ by (29). Therefore, the sum in (42) satisfies $\lim_{n\to\infty} \sum_{1\le j\le d} \mathrm{E}[(\omega_j^{(n)}(U_j))^2] = 0$, so we have $\lim_{n\to\infty} \mathrm{Var}_{\mathrm{LH}}[B_n] = 0$, which, combined with the previously established $\mathrm{E}_{\mathrm{LH}}[B_n] = 0$, implies $B_n \Rightarrow 0$ as $n \to \infty$. We have thus proven that both A_n and B_n in (33) weakly converge to 0 as $n \to \infty$. Hence, $\hat{D}_n \Rightarrow 0$ as $n \to \infty$ by Slutsky's theorem, completing the proofs of Lemma 2 and Theorem 1. □

Appendix B: Proof of Theorem 2

Theorem 3.3(i) of Chu and Nakayama (2012a) establishes that if a perturbed p_n-quantile estimator satisfies the perturbed Bahadur representation in (7) with $p_n = p + cn^{-1/2}$ for any constant $c \neq 0$, as we proved in Theorem 1, then $\hat{\lambda}_{p,n} \Rightarrow \lambda_p$ as $n \to \infty$. Hence, if we show that

$$\hat{\psi}_{p,n}^2 \Rightarrow \psi_p^2 \quad \text{as} \quad n \to \infty, \tag{43}$$

we then have that $\hat{\eta}_{p,n} = \hat{\psi}_{p,n}\hat{\lambda}_{p,n} \Rightarrow \psi_p\lambda_p = \eta_p$ as $n \to \infty$ by Slutsky's theorem, and the asymptotic validity of CI $C_{n,\mathrm{ssLHS}}$ in (20) follows from the CLT (10). Thus, what remains is to establish (43).

We do this by applying the following lemma, which we verify by modifying arguments used in the proof of Theorem 4.1(ii) of Chu and Nakayama (2012a).

Lemma 3. *Suppose $l(x)$ is a real-valued function that is continuous at $x = q$. Also suppose that $L_n(x) \Rightarrow l(x)$ as $n \to \infty$ for each x in a neighborhood N of q and that for each n, $L_n(x)$ is monotonic in x for $x \in N$. Then $Q_n \Rightarrow q$ as $n \to \infty$ implies $L_n(Q_n) \Rightarrow l(q)$ as $n \to \infty$.*

Proof of Lemma 3 We will demonstrate that $b_n \equiv P(|L_n(Q_n) - l(q)| > \epsilon) \to 0$ as $n \to \infty$ for each $\epsilon > 0$. Because $l(x)$ is continuous at $x = q$, there exists a $\delta > 0$ such that

$$\max\left(|l(q+\delta) - l(q)|, |l(q-\delta) - l(q)|\right) < \epsilon/2. \tag{44}$$

We can choose $\delta > 0$ such that $q + \delta$ and $q - \delta$ lie in the neighborhood N. We then express

$$b_n = P(|L_n(Q_n) - l(q)| > \epsilon, |Q_n - q| \le \delta) + P(|L_n(Q_n) - l(q)| > \epsilon, |Q_n - q| > \delta) \equiv r_n + s_n. \tag{45}$$

The assumed monotonicity of $L_n(x)$ in x for $x \in N$ implies that $|L_n(Q_n) - l(q)| \le \max(|L_n(q+\delta) - l(q)|, |L_n(q-\delta) - l(q)|)$ when $|Q_n - q| \le \delta$. Hence,

$$\begin{aligned} r_n &\le P\left(\max(|L_n(q+\delta) - l(q)|, |L_n(q-\delta) - l(q)|) > \epsilon, |Q_n - q| \le \delta\right) \\ &\le P\left(|L_n(q+\delta) - l(q)| > \epsilon, |Q_n - q| \le \delta\right) + P\left(|L_n(q-\delta) - l(q)| > \epsilon, |Q_n - q| \le \delta\right) \equiv r_{n,1} + r_{n,2}. \end{aligned}$$

Also $r_{n,1} \le P\left(|L_n(q+\delta) - l(q)| > \epsilon\right) \le P\left(|L_n(q+\delta) - l(q+\delta)| + |l(q+\delta) - l(q)| > \epsilon\right) \le P\left(|L_n(q+\delta) - l(q+\delta)| > \epsilon/2\right) \to 0$ as $n \to \infty$, where the last inequality follows from (44), and the convergence holds because $q + \delta \in N$ and we assumed $L_n(x) \Rightarrow l(x)$ for $x \in N$. Similarly, $r_{n,2} \to 0$ as $n \to \infty$, so $r_n \to 0$. Also, $s_n \le P(|Q_n - q| > \delta) \to 0$ as $n \to \infty$ because $Q_n \Rightarrow q$ as $n \to \infty$. Combining these results with (45) establishes that $b_n \to 0$, completing the proof of Lemma 3. □

We now return to verifying that (43) holds. By (18),

$$\hat{\psi}^2_{p,n} = \sum_{j=1}^{d} \hat{\sigma}^2_j(\hat{\xi}_{p,n}) - (d-1)p(1-p), \text{ where } \hat{\sigma}^2_{j,n}(x) = \frac{1}{2n} \sum_{1\le i\le n} \left[W_i(x) - N_j W_i(x)\right]^2, \ x \in \Re. \quad (46)$$

We will apply Lemma 3 to find the weak limit of each $\hat{\sigma}^2_{j,n}(\hat{\xi}_{p,n})$ as $n \to \infty$ using the following approach. First, for each $1 \le j \le d$ and x in the neighborhood N of ξ_p for which the Lipschitz condition (21) holds, we show that $\hat{\sigma}^2_{j,n}(x)$ has a deterministic weak limit $\sigma^2_j(x)$, which is continuous at $x = \xi_p$; see Lemma 4 below. Invoking Lemma 3 also requires that $\hat{\sigma}^2_{j,n}(x)$ is monotone in x in the neighborhood N, but it is not clear this holds. Instead, we next derive (in the proof of Lemma 5 below) an alternative representation of $\sigma^2_j(x)$ as the sum of three terms, each continuous at $x = \xi_p$ and whose corresponding estimators are monotone in x everywhere, and so also in N. We then apply Lemma 3 to each of the three estimators to verify that $\hat{\sigma}^2_{j,n}(\hat{\xi}_{p,n}) \Rightarrow \sigma^2_j(\xi_p)$ as $n \to \infty$. We now provide the details.

LEMMA 4. *Under the conditions of Theorem 2,*

$$\hat{\sigma}^2_{j,n}(x) \Rightarrow \sigma^2_j(x) \equiv \mathit{Var}[\zeta(\mathbf{U},x)] - \mathit{Var}[\zeta_j(U_j,x)] = E[\mathit{Var}[\zeta(\mathbf{U},x) \,|\, U_j]] \quad (47)$$

as $n \to \infty$ for each x in the neighborhood N for which (21) *holds, and $\sigma_j(x)$ is continuous at $x = \xi_p$.*

Proof of Lemma 4 We first show that for each $x \in N$ and $j = 1, 2, \ldots, d$, the weak convergence in (47) holds as $n \to \infty$ under the Lipschitz condition (21). We will establish this by applying ideas similar to those employed in the proof of Theorem 2 of Owen (1992).

For $i = 1, 2, \ldots, n$, recall that we previously defined $\mathbf{U}^{(i)} = (V_{i,1}, V_{i,2}, \ldots, V_{i,d})$ as the ith row of LHS inputs from (4) with $s = n$. To emphasize the dependence of the output $\hat{Y}_i$ on the input $\mathbf{U}^{(i)}$, we recall the function $\zeta(\mathbf{u}, x) = I(g(\mathbf{u}) \le x)$ for $\mathbf{u} \in [0,1]^d$ and $x \in \Re$, where g is defined in (1). Thus, $W_i(x) = \zeta(\mathbf{U}^{(i)}, x)$, and $\chi(\mathbf{u}) = I(g(\mathbf{u}) \le \xi_p) = \zeta(\mathbf{u}, \xi_p)$. For $\mathbf{U} = (U_1, U_2, \ldots, U_d)$ a vector of i.i.d. unif$[0,1]$ random variables and $u_j \in [0,1]$, let $\zeta_j(u_j, x) = \mathrm{E}[\zeta(\mathbf{U}, x) \,|\, U_j = u_j]$ be the jth main effect of $\zeta(\cdot, x)$, and define $\zeta_{\mathrm{res}}(\mathbf{u}, x)$ as its residual from additivity. Note that $\mathrm{E}[\zeta(\mathbf{U}, x)] = \mathrm{E}[\zeta_j(U_j, x)] = F(x)$. Hence, the last step in (47) holds by a variance decomposition.

Continuing now with establishing the weak convergence in (47), we next write

$$\begin{aligned}
\hat{\sigma}^2_{j,n}(x) &= \frac{1}{2n} \sum_{1\le i\le n} \left[W_i(x) - \zeta_j(V_{i,j},x) + \zeta_j(V_{i,j},x) - N_j W_i(x)\right]^2 \\
&= \frac{1}{2n} \sum_{1\le i\le n} \left[W_i(x) - \zeta_j(V_{i,j},x)\right]^2 + \frac{1}{2n} \sum_{1\le i\le n} \left[N_j W_i(x) - \zeta_j(V_{i,j},x)\right]^2 \\
&\quad - \frac{1}{n} \sum_{1\le i\le n} \left[W_i(x) - \zeta_j(V_{i,j},x)\right] \left[N_j W_i(x) - \zeta_j(V_{i,j},x)\right] \equiv H_{1,n} + H_{2,n} - H_{3,n}. \quad (48)
\end{aligned}$$

We separately analyze each term in (48), and we will show that as $n \to \infty$,

$$H_{1,n} \Rightarrow \frac{1}{2}\sigma_j^2(x), \tag{49}$$

$$H_{2,n} \Rightarrow \frac{1}{2}\sigma_j^2(x), \tag{50}$$

$$H_{3,n} \Rightarrow 0, \tag{51}$$

which will prove the weak convergence in (47) by Slutsky's theorem.

For the first term in (48), Owen (1992) states without proof that something similar to (49) holds in the line immediately following his equation (13), and we now provide the details for completeness. To do this, we will prove that

$$\mathrm{E}_{\mathrm{LH}}[H_{1,n}] = \frac{1}{2}\sigma_j^2(x) \tag{52}$$

for each n, and that

$$\mathrm{Var}_{\mathrm{LH}}[H_{1,n}] \to 0 \text{ as } n \to \infty, \tag{53}$$

which together ensure mean-square convergence of $H_{1,n}$ to $\sigma_j^2(x)/2$, implying (49) (e.g., see Theorem 1.3.2 of Serfling (1980)).

To prove (52), we have that because $W_i(x) = \zeta(\mathbf{U}^{(i)}, x)$,

$$\begin{aligned}\mathrm{E}_{\mathrm{LH}}[H_{1,n}] &= \frac{1}{2n}\sum_{1\le i\le n} \mathrm{E}_{\mathrm{LH}}\left[\left(\zeta(\mathbf{U}^{(i)},x) - \zeta_j(V_{i,j},x)\right)^2\right] \\ &= \frac{1}{2}\mathrm{E}\left[\zeta^2(\mathbf{U},x)\right] - \mathrm{E}\left[\zeta(\mathbf{U},x)\zeta_j(U_j,x)\right] + \frac{1}{2}\mathrm{E}\left[\zeta_j^2(U_j,x)\right] \end{aligned} \tag{54}$$

by expanding the square and because the LHS rows $\mathbf{U}^{(i)}$, $i = 1, 2, \ldots, n$, are identically distributed as $\mathbf{U}$. The middle term in (54), by iterated expectations, satisfies

$$\mathrm{E}\left[\zeta(\mathbf{U},x)\zeta_j(U_j,x)\right] = \mathrm{E}\left[\mathrm{E}\left[\zeta(\mathbf{U},x)\zeta_j(U_j,x)\,|\,U_j\right]\right] = \mathrm{E}\left[\zeta_j(U_j,x)\mathrm{E}\left[\zeta(\mathbf{U},x)\,|\,U_j\right]\right] = \mathrm{E}\left[\zeta_j^2(U_j,x)\right].$$

Thus, (54) becomes

$$\begin{aligned}\mathrm{E}_{\mathrm{LH}}[H_{1,n}] &= \frac{1}{2}\left(\mathrm{E}\left[\zeta^2(\mathbf{U},x)\right] - \mathrm{E}\left[\zeta_j^2(U_j,x)\right]\right) \\ &= \frac{1}{2}\left(\mathrm{Var}[\zeta(\mathbf{U},x)] - \mathrm{Var}[\zeta_j(U_j,x)]\right) = \frac{1}{2}\sigma_j^2(x)\end{aligned}$$

by the definition of $\sigma_j^2(x)$ in (47), where the second equality holds because $(\mathrm{E}[\zeta(\mathbf{U},x)])^2 = (\mathrm{E}[\zeta_j(U_j,x)])^2$ by iterated expectations. Hence, we have established (52).

To verify (53), let $v_j(\mathbf{u}) = \left[\zeta(\mathbf{u},x) - \zeta_j(u_j,x)\right]^2/2$ for $\mathbf{u} = (u_1, u_2, \ldots, u_d)$, so $H_{1,n} = (1/n)\sum_{1\le i\le n} v_j(\mathbf{U}^{(i)})$. Then Proposition 3 of Owen (1997) implies that

$$\mathrm{Var}_{\mathrm{LH}}\left[H_{1,n}\right] \le \frac{\mathrm{Var}[v_j(\mathbf{U})]}{n-1} \to 0 \text{ as } n \to \infty$$

by the boundedness of $v_j(\mathbf{U})$. Thus, (53) holds, which combined with (52) establishes (49).

Next we show that the second term, $H_{2,n}$, in (48) satisfies (50). To do this, we extend the operator N_j in (16) to further apply to $V_{i,j}$, with $N_jV_{i,j} = V_{m,j}$ if $\pi_j(i) < n$, where $\pi_j(m) = \pi_j(i) + 1$; and $N_jV_{i,j} = 1/2$ if $\pi_j(i) = n$. Then write

$$\begin{aligned} H_{2,n} &= \frac{1}{2n} \sum_{1\le i\le n} \left[N_jW_i(x) - \zeta_j(N_jV_{i,j},x) + \zeta_j(N_jV_{i,j},x) - \zeta_j(V_{i,j},x)\right]^2 \\ &= \frac{1}{2n} \sum_{1\le i\le n} \left[N_jW_i(x) - \zeta_j(N_jV_{i,j},x)\right]^2 + \frac{1}{2n} \sum_{1\le i\le n} \left[\zeta_j(N_jV_{i,j},x) - \zeta_j(V_{i,j},x)\right]^2 \\ &+ \frac{1}{n} \sum_{1\le i\le n} \left[N_jW_i(x) - \zeta_j(N_jV_{i,j},x)\right]\left[\zeta_j(N_jV_{i,j},x) - \zeta_j(V_{i,j},x)\right] \equiv J_{1,n} + J_{2,n} + J_{3,n}. \end{aligned} \tag{55}$$

The collections of summands for $J_{1,n}$ and $H_{1,n}$ differ in only the two extreme cases when $\pi_j(\cdot)$ maps to 1 and to n. Specifically, define integer s so that $\pi_j(s) = 1$, and N_j maps nothing to the element with index s. Also, when $\pi_j(i) = n$, we have that $N_jW_i(x) = \bar{W}(n)$ and $N_jV_{i,j} = 1/2$. Therefore,

$$J_{1,n} = H_{1,n} - \frac{1}{2n}[W_s(x) - \zeta_j(V_{s,j},x)]^2 + \frac{1}{2n}[\bar{W}(n) - \zeta_j(1/2,x)]^2 \Rightarrow \frac{1}{2}\sigma_j^2(x) \tag{56}$$

as $n \to \infty$ by (49) because $W_s(x)$, $\bar{W}_n$, and $\zeta_j(\cdot,x)$ are bounded between 0 and 1. We now show that $J_{2,n} \Rightarrow 0$. For each i such that $\pi_j(i) < n$, we have that $N_jV_{i,j}$ and $V_{i,j}$ lie in adjacent strata for coordinate j. Thus, $|N_jV_{i,j} - V_{i,j}| \le 2/n$, so $x \in N$ and the Lipschitz condition (21) ensure that

$$|\zeta_j(N_jV_{i,j},x) - \zeta_j(V_{i,j},x)| \le 2c_0/n \quad \text{when} \quad \pi_j(i) < n. \tag{57}$$

Consequently, defining integer s' so that $\pi_j(s') = n$, we get

$$\begin{aligned} |J_{2,n}| &= \frac{1}{2n} \sum_{i:\pi_j(i)<n} \left[\zeta_j(N_jV_{i,j},x) - \zeta_j(V_{i,j},x)\right]^2 + \frac{1}{2n}\left[\zeta_j(N_jV_{s',j},x) - \zeta_j(V_{s',j},x)\right]^2 \\ &\le \frac{n-1}{2n}\left(\frac{2c_0}{n}\right)^2 + \frac{1}{2n} = O(n^{-1}) \end{aligned} \tag{58}$$

by (57) and because $0 \le \zeta_j(\cdot,x) \le 1$. Similarly, we can again use (57) to show that $J_{3,n} = O(n^{-1})$, which combined with (55), (56), and (58) yields (50).

Finally, for the third term, $H_{3,n}$, in (48), we will verify (51) by arguing that

$$\mathrm{E}_{\mathrm{LH}}[H_{3,n}^2] \to 0 \text{ as } n \to \infty \tag{59}$$

and applying Theorem 1.3.2 of Serfling (1980). Note that

$$\begin{aligned} H_{3,n}^2 &= \frac{1}{n^2} \sum_{1\le i\le n} \sum_{1\le k\le n} \left[W_i(x) - \zeta_j(V_{i,j},x)\right]\left[W_k(x) - \zeta_j(V_{k,j},x)\right] \\ &\quad \times \left[N_jW_i(x) - \zeta_j(V_{i,j},x)\right]\left[N_jW_k(x) - \zeta_j(V_{k,j},x)\right] \equiv K_{1,n} + K_{2,n}, \end{aligned} \tag{60}$$

where $K_{1,n}$ is the sum of the terms for which $i=k$, and $K_{2,n}$ sums the terms with $i\neq k$. The summands of $H_{3,n}^2$ are bounded, so we can establish that the expectation of certain sums of terms from $H_{3,n}^2$ asymptotically vanish by simply showing that their number of summands is $o(n^2)$. For example, $K_{1,n}$ has n summands, so multiplying their sum by $1/n^2$ leads to

$$\mathrm{E}_{\mathrm{LH}}[K_{1,n}]=O(n^{-1}). \tag{61}$$

To analyze $K_{2,n}$, which we recall includes the terms from $H_{3,n}^2$ with $i\neq k$, let $G_{l,j}=\zeta_j(N_jV_{l,j},x)-\zeta_j(V_{l,j},x)$ for $l=1,2,\ldots,n$. Also, let $N_{j,l}=m$ if $\pi_j(l)<n$, where $\pi_j(m)=\pi_j(l)+1$; and $N_{j,l}=0$ if $\pi_j(l)=n$. Moreover, define $W_0(x)=\bar{W}_n(x)$ and $V_{0,j}=1/2$. We then have that

$$\begin{aligned}
K_{2,n}&=\frac{1}{n^2}\sum_{1\le i\le n}\sum_{\substack{1\le k\le n:\\ k\neq i}}\left[W_i(x)-\zeta_j(V_{i,j},x)\right]\left[W_k(x)-\zeta_j(V_{k,j},x)\right]\\
&\quad\times\left[N_jW_i(x)-\zeta_j(N_jV_{i,j},x)+G_{i,j}\right]\left[N_jW_k(x)-\zeta_j(N_jV_{k,j},x)+G_{k,j}\right]\\
&=\frac{1}{n^2}\sum_{1\le i\le n}\sum_{\substack{1\le k\le n:\\ k\neq i}}\left[W_i(x)-\zeta_j(V_{i,j},x)\right]\left[W_k(x)-\zeta_j(V_{k,j},x)\right]\\
&\quad\times\left[W_{N_{j,i}}(x)-\zeta_j(V_{N_{j,i},j},x)\right]\left[W_{N_{j,k}}(x)-\zeta_j(V_{N_{j,k},j},x)\right]\\
&\quad+\frac{1}{n^2}\sum_{1\le i\le n}\sum_{\substack{1\le k\le n:\\ k\neq i}}\left[W_i(x)-\zeta_j(V_{i,j},x)\right]\left[W_k(x)-\zeta_j(V_{k,j},x)\right]\\
&\quad\times\left[\left(W_{N_{j,i}}(x)-\zeta_j(V_{N_{j,i},j},x)\right)G_{k,j}+\left(W_{N_{j,k}}(x)-\zeta_j(V_{N_{j,k},j},x)\right)G_{i,j}+G_{i,j}G_{k,j}\right]\\
&\equiv T_{1,n}+T_{2,n},
\end{aligned}$$

where $T_{1,n}$ is the first double sum, and $T_{2,n}$ is the second double sum. We next handle $T_{1,n}$ and $T_{2,n}$ by breaking down each into various sums over disjoint sets of index pairs (i,k), where one of those sums for $T_{1,n}$ has each of its summands being the product of four distinct terms, with $N_{j,i}\neq 0$ and $N_{j,k}\neq 0$. We will analyze that sum later after first handling all of the other cases.

For a given permutation π_j, the definition of $N_{j,l}$ precludes the possibility that both $N_{j,i}=k$ and $N_{j,k}=i$. (To see this, note that $N_{j,i}=k$ means that $\pi_j(k)=\pi_j(i)+1$, which cannot simultaneously occur with $\pi_j(i)=\pi_j(k)+1$, which corresponds to $N_{j,k}=i$.) Thus, we only need to consider summands in $T_{1,n}$ and $T_{2,n}$ with index pairs (i,k) for which $N_{j,i}\neq k$ or $N_{j,k}\neq i$. There are $O(n)$ summands with index pairs (i,k) for which $N_{j,i}=k$ and $N_{j,k}\neq i$, or for which $N_{j,i}\neq k$ and $N_{j,k}=i$. After multiplying by the leading $1/n^2$, the sum of those $O(n)$ terms contribute $O(n^{-1})$ to $\mathrm{E}_{\mathrm{LH}}[K_{2,n}]$. In addition, there are $O(n)$ summands with index pairs (i,k) for which $N_{j,i}=0$ or $N_{j,k}=0$, so again, the sum of those multiplied by $1/n^2$ add $O(n^{-1})$ to $\mathrm{E}_{\mathrm{LH}}[K_{2,n}]$. Let $\mathcal{A}_n$ be the set of remaining possible index pairs (i,k), i.e., for which $1\le i\le n$, $1\le k\le n$, $i\neq k$, $N_{j,i}\neq k$, $N_{j,k}\neq i$, $N_{j,i}\neq 0$, and $N_{j,k}\neq 0$ all hold. While $\mathcal{A}_n$ depends (solely) on π_j, its cardinality $|\mathcal{A}_n|=O(n^2)$ for all π_j.

Let $T'_{1,n}$ and $T'_{2,n}$ be the sum of the terms with indices $(i,k) \in \mathcal{A}_n$ in $T_{1,n}$ and $T_{2,n}$, respectively. Also, define $e(\mathbf{U}^{(l)}) = W_l(x) - \zeta_j(V_{l,j}, x) \;\; = \zeta(\mathbf{U}^{(l)}, x) - \zeta_j(V_{l,j}, x)$ for $1 \le l \le n$. Thus, we have that $T'_{1,n} = (1/n^2) \sum_{(i,k)\in\mathcal{A}_n} e(\mathbf{U}^{(i)}) e(\mathbf{U}^{(k)}) e(\mathbf{U}^{(N_{j,i})}) e(\mathbf{U}^{(N_{j,k})})$, and $(i,k) \in \mathcal{A}_n$ implies that each summand is the product of four different $e(\cdot)$ terms, with $N_{j,i} \neq 0$ and $N_{j,k} \neq 0$. Consequently, using iterated expectations yields

$$
\begin{aligned}
&\mathrm{E}_{\mathrm{LH}}[T'_{1,n}] = \mathrm{E}_{\mathrm{LH}}\left[\mathrm{E}_{\mathrm{LH}}[T'_{1,n} \,|\, \pi_j]\right] \\
&= \frac{1}{n^2} \mathrm{E}_{\mathrm{LH}}\left[\sum_{(i,k)\in\mathcal{A}_n} \mathrm{E}_{\mathrm{LH}}\left[\mathrm{E}_{\mathrm{LH}}[e(\mathbf{U}^{(i)}) e(\mathbf{U}^{(k)}) e(\mathbf{U}^{(N_{j,i})}) e(\mathbf{U}^{(N_{j,k})}) \,|\, \mathbf{U}^{(k)}, \mathbf{U}^{(N_{j,i})}, \mathbf{U}^{(N_{j,k})}, \pi_j] \Big| \pi_j \right] \right] \\
&= \frac{1}{n^2} \mathrm{E}_{\mathrm{LH}}\left[\sum_{(i,k)\in\mathcal{A}_n} \mathrm{E}_{\mathrm{LH}}\left[\mathrm{E}_{\mathrm{LH}}[e(\mathbf{U}^{(i)}) \,|\, \mathbf{U}^{(k)}, \mathbf{U}^{(N_{j,i})}, \mathbf{U}^{(N_{j,k})}, \pi_j] \, e(\mathbf{U}^{(k)}) e(\mathbf{U}^{(N_{j,i})}) e(\mathbf{U}^{(N_{j,k})}) \Big| \pi_j \right] \right] \\
&= \frac{1}{n^2} \mathrm{E}_{\mathrm{LH}}\left[\sum_{(i,k)\in\mathcal{A}_n} \mathrm{E}_{\mathrm{LH}}\left[\big(\mathrm{E}[e(\mathbf{U})] + O(n^{-1})\big) \, e(\mathbf{U}^{(k)}) e(\mathbf{U}^{(N_{j,i})}) e(\mathbf{U}^{(N_{j,k})}) \Big| \pi_j \right] \right] \qquad (62) \\
&= \frac{1}{n^2} \mathrm{E}_{\mathrm{LH}}\left[\sum_{(i,k)\in\mathcal{A}_n} \mathrm{E}_{\mathrm{LH}}\left[O(n^{-1}) e(\mathbf{U}^{(k)}) e(\mathbf{U}^{(N_{j,i})}) e(\mathbf{U}^{(N_{j,k})}) \Big| \pi_j \right] \right], \qquad (63)
\end{aligned}
$$

where (62) follows from Lemma 1 of Owen (1992), and the last step holds because $\mathrm{E}[e(\mathbf{U})] = \mathrm{E}[\zeta(\mathbf{U}, x)] - \mathrm{E}[\zeta_j(U_j, x)] = 0$. Then by the boundedness of $e(\cdot)$ and the fact that $|\mathcal{A}_n| = O(n^2)$, we have that (63) implies $\mathrm{E}_{\mathrm{LH}}[T'_{1,n}] = O(n^{-1})$. Thus, combining this with the other index pairs $(i,k) \notin \mathcal{A}_n$ considered before in the previous paragraph yields $\mathrm{E}_{\mathrm{LH}}[T_{1,n}] \to 0$ as $n \to \infty$. A similar argument shows that $\mathrm{E}_{\mathrm{LH}}[T_{2,n}] \to 0$ by using the fact that $|G_{l,j}| \le 2c_0/n$ for each l such that $\pi_j(l) < n$ by (57). Hence, combining these results implies $\mathrm{E}_{\mathrm{LH}}[K_{2,n}] \to 0$ as $n \to \infty$, which with (60) and (61) ensures that (59) holds. This verifies (51), which combined with (48), (49), and (50) establishes the weak convergence in (47).

We now show that $\sigma_j^2(x)$ is continuous at $x = \xi_p$, which we do by establishing the continuity at $x = \xi_p$ of its two terms in (47): $\mathrm{Var}[\zeta(\mathbf{U}, x)]$ and $\mathrm{Var}[\zeta_j(U_j, x)]$. Because $F'(\xi_p) > 0$, $F(x)$ is continuous at $x = \xi_p$, which implies the following. First, $\mathrm{Var}[\zeta(\mathbf{U}, x)] = F(x)[1 - F(x)]$ is also continuous at $x = \xi_p$, which handles the first term of $\sigma_j^2(x)$. Second, $P(Y = \xi_p) = P(g(\mathbf{U}) = \xi_p) = 0$, so $\lim_{y\to\xi_p} \zeta(\mathbf{U}, y) = \zeta(\mathbf{U}, \xi_p)$ almost surely (a.s.). Since $0 \le \zeta(\mathbf{u}, y) \le 1$ for all $\mathbf{u} \in [0,1]^d$ and all y, the conditional dominated convergence theorem (e.g., p. 88 of Williams (1991)) yields $\lim_{y\to\xi_p} \zeta_j(U_j, y) = \zeta_j(U_j, \xi_p)$ a.s. and $\lim_{y\to\xi_p} \zeta_j^2(U_j, y) = \zeta_j^2(U_j, \xi_p)$ a.s. Because $0 \le \zeta_j(u, y) \le 1$ for all $u \in [0,1]$ and all y, which ensures the same for $\zeta_j^2(u, y)$, the BCT guarantees that $\lim_{y\to\xi_p} \mathrm{Var}[\zeta_j(U_j, y)] = \mathrm{Var}[\zeta_j(U_j, \xi_p)]$, so $\mathrm{Var}[\zeta_j(U_j, x)]$ is continuous at $x = \xi_p$. Thus, $\sigma_j^2(x)$ is also, and Lemma 4 is proven. □

LEMMA 5. *Under the conditions of Theorem 2,* $\hat{\sigma}_{j,n}^2(\hat{\xi}_{p,n}) \Rightarrow \sigma_j^2(\xi_p)$ *as* $n \to \infty$.

Proof of Lemma 5 It is not clear that $\hat{\sigma}^2_{j,n}(x)$ is monotonic in a neighborhood of ξ_p, as required to apply Lemma 3. Instead, we now derive another form for $\sigma^2_j(x)$ as the sum of three terms, each continuous at $x=\xi_p$ and whose estimator has the desired monotonicity. By (46),

$$\hat{\sigma}^2_{j,n}(x) = L_{j,1,n}(x) + L_{j,2,n}(x) - L_{j,3,n}(x), \tag{64}$$

where $L_{j,1,n}(x) = (1/(2n))\sum_{1\le i\le n} W_i^2(x)$, $L_{j,2,n}(x) = (1/(2n))\sum_{1\le i\le n}[N_j W_i(x)]^2$, and $L_{j,3,n}(x) = (1/n)\sum_{1\le i\le n} W_i(x)[N_j W_i(x)]$. For each x, because $W_i(x)\in\{0,1\}$, we have that

$$L_{j,1,n}(x) = \frac{1}{2n}\sum_{1\le i\le n} W_i(x) = \hat{F}_n(x)/2 \Rightarrow F(x)/2 \equiv l_{j,1}(x) \tag{65}$$

as $n\to\infty$ by the CLT for LHS estimators of a mean of bounded outputs, from Owen (1992). Recalling (16), define s so that $\pi_j(s)=1$. Since $W_i(x)\in\{0,1\}$ and N_j maps nothing to index s,

$$L_{j,2,n}(x) = \left(\frac{1}{2n}\sum_{1\le i\le n} W_i(x)\right) - \frac{1}{2n}W_s(x) + \frac{1}{2n}(\bar{W}_n(x))^2 \Rightarrow F(x)/2 \equiv l_{j,2}(x) \tag{66}$$

because $W_s(x)$ and $\bar{W}_n(x)$ are bounded between 0 and 1. The weak limits in (47), (65), and (66) are all deterministic, so for each x in the neighborhood N for which (21) holds, the corresponding left sides jointly converge as $n\to\infty$ by Theorem 3.9 of Billingsley (1999). Thus, (64) and the continuous-mapping theorem (CMT; Theorem 29.2 of Billingsley (1995)) ensure that for each $x\in N$, $L_{j,3,n}(x) = L_{j,1,n}(x) + L_{j,2,n}(x) - \hat{\sigma}^2_{j,n}(x) \Rightarrow F(x) - \sigma^2_j(x) \equiv l_{j,3}(x)$ as $n\to\infty$. Because $F(x)$ and $\sigma^2_j(x)$ are both continuous at $x=\xi_p$, the latter by Lemma 4, we then have that $l_{j,3}(x)$ maintains the same property. Summarizing, we have shown that for each $k=1,2,3$, the deterministic limit $l_{j,k}(x)$ is continuous at $x=\xi_p$. Also, for all x in the neighborhood N for which (21) holds, the estimators $L_{j,k,n}(x) \Rightarrow l_{j,k}(x)$ as $n\to\infty$, and $\sigma^2_j(x) = l_{j,1}(x) + l_{j,2}(x) - l_{j,3}(x)$.

To apply Lemma 3, we also need that the estimators $L_{j,k,n}(x)$, $k=1,2,3$, are monotonic in x in the neighborhood N. This holds because $W_i(x)$ and $N_j W_i(x)$ are monotonically increasing in x (actually for all x, not just in N) and nonnegative. Since $\hat{\xi}_{p,n} \Rightarrow \xi_p$ as $n\to\infty$ by (8) and (9), Lemma 3 ensures that $L_{j,k,n}(\hat{\xi}_{p,n}) \Rightarrow l_{j,k}(\xi_p)$ as $n\to\infty$ for $k=1,2,3$, so $\hat{\sigma}^2_j(\hat{\xi}_{p,n}) \Rightarrow \sigma^2_j(\xi_p)$ as $n\to\infty$, proving Lemma 5. □

Finally, because $\mathrm{Var}[\zeta(\mathbf{U},\xi_p)] = p(1-p)$, (46) and Lemmas 4 and 5 imply

$$\begin{aligned}\hat{\psi}^2_{p,n} &\Rightarrow \sum_{j=1}^d \sigma^2_j(\xi_p) - (d-1)p(1-p) = \sum_{j=1}^d (\mathrm{Var}[\zeta(\mathbf{U},\xi_p)] - \mathrm{Var}[\zeta_j(U_j,\xi_p)]) - (d-1)(\mathrm{Var}[\zeta(\mathbf{U},\xi_p)] \\ &= \mathrm{Var}[\zeta(\mathbf{U},\xi_p)] - \sum_{j=1}^d \mathrm{Var}[\zeta_j(U_j,\xi_p)] = \mathrm{Var}[\zeta_{\mathrm{res}}(\mathbf{U},\xi_p)] = \mathrm{Var}[\chi_{\mathrm{res}}(\mathbf{U})] = \psi^2_p\end{aligned}$$

by (15) because $\zeta(\mathbf{U},\xi_p) = \chi(\mathbf{U})$, verifying that (43) holds, which completes the proof.

Appendix C: Proof of Theorem 3

We will show that Conditions A1, A3, and A4 from Appendix A hold for the srLHS CDF estimator $\check{F}_{b,m}$ in (22). In our definition of $\check{F}_{b,m}$, each $I(\tilde{Y}_{j,i} \le y)$ is monotonically increasing in y, so $\check{F}_{b,m}(y)$ is also monotonically increasing. Hence, Condition A1 is satisfied.

To handle Condition A4, first fix $t \ge 0$ (the case when $t < 0$ can be treated similarly). Define

$$\begin{aligned}\check{D}_n &= n^{1/2}\left[\left(\check{F}_{b,m}(\xi_p + n^{-1/2}t) - \check{F}_{b,m}(\xi_p)\right) - \left(F(\xi_p + n^{-1/2}t) - F(\xi_p)\right)\right] \\ &= b^{-1/2}\sum_{1\le j\le b} m^{1/2}\left[\left(\tilde{F}_{j,m}(\xi_p + m^{-1/2}b^{-1/2}t) - \tilde{F}_{j,m}(\xi_p)\right) - \left(F(\xi_p + m^{-1/2}b^{-1/2}t) - F(\xi_p)\right)\right]\end{aligned}$$

by (22) and (23) since $n = bm$, and we want to prove $\check{D}_n \Rightarrow 0$ as $n \to \infty$. Let $t' = b^{-1/2}t$. Because each batch j contains a single LHS sample of size m, Lemma 2 implies $m^{1/2}\left[\left(\tilde{F}_{j,m}(\xi_p + m^{-1/2}t') - \tilde{F}_{j,m}(\xi_p)\right) - \left(F(\xi_p + m^{-1/2}t') - F(\xi_p)\right)\right] \Rightarrow 0$ as $m \to \infty$, i.e., $n = bm \to \infty$. Therefore, $\check{D}_n \Rightarrow 0$ as $n \to \infty$ by the independence of the batches, Example 3.2 and Theorem 3.9 of Billingsley (1999), and the CMT. Thus, Condition A4 holds for $\check{F}_{b,m}$.

Now we prove $\check{F}_{b,m}$ satisfies Condition A3. The weak convergence in (9), which is true for ssLHS, implies that for each batch j in srLHS, we have $m^{1/2}[\tilde{F}_{j,m}(\xi_p) - F(\xi_p)] \Rightarrow M_j \sim N(0, \psi_p^2)$ as $m \to \infty$, where $\psi_p^2 = \mathrm{Var}[\chi_{\mathrm{res}}(\mathbf{U})]$ by (15). The LHS samples across batches are independent, so

$$\left(m^{1/2}[\tilde{F}_{j,m}(\xi_p) - F(\xi_p)] : j = 1, 2, \ldots, b\right) \Rightarrow \left(M_j : j = 1, 2, \ldots, b\right) \tag{67}$$

as $m \to \infty$ by Example 3.2 of Billingsley (1999), where $M_1, M_2, \ldots, M_b$ are mutually independent. Because $\check{F}_{b,m}(y) = (1/b)\sum_{1\le j\le b}\tilde{F}_{j,m}(y)$ and $n = bm$, the CMT guarantees that

$$n^{1/2}[\check{F}_{b,m}(\xi_p) - F(\xi_p)] = b^{1/2}\left[\frac{1}{b}\sum_{1\le j\le b} m^{1/2}\left(\tilde{F}_{j,m}(\xi_p) - F(\xi_p)\right)\right] \Rightarrow \frac{1}{b^{1/2}}\sum_{1\le j\le b} M_j \sim N(0, \psi_p^2)$$

as $n = bm \to \infty$ with b fixed. Thus, Condition A3 holds for srLHS, completing the proof.

Appendix D: Proof of Theorem 4

We first prove the srLHS batching CI $\check{C}_{b,m,\mathrm{batch}}$ from (26) is asymptotically valid. By (10), for each batch j in the srLHS batching method, the quantile estimator $\tilde{\xi}_{p,j,m}$ obeys a CLT $m^{1/2}[\tilde{\xi}_{p,j,m} - \xi_p] \Rightarrow M_j' \sim N(0, \eta_p^2)$ as $m \to \infty$. The batches are independent, so Example 3.2 of Billingsley (1999) implies

$$\left(m^{1/2}[\tilde{\xi}_{p,j,m} - \xi_p] : j = 1, 2, \ldots, b\right) \Rightarrow (M_j' : j = 1, 2, \ldots, b) \tag{68}$$

as $m \to \infty$, with $M_1', M_2', \ldots, M_b'$ independent. Let $\bar{M}_b' = (1/b)\sum_{1\le j\le b} M_j'$, and the CMT ensures

$$\begin{aligned}\frac{b^{1/2}[\acute{\xi}_{p,b,m} - \xi_p]}{\check{S}_{b,m,\mathrm{batch}}} &= \frac{b^{1/2}\left[(1/b)\sum_{1\le j\le b} m^{1/2}(\tilde{\xi}_{p,j,m} - \xi_p)\right]}{\left[(1/(b-1))\sum_{1\le j\le b}\left(m^{1/2}(\tilde{\xi}_{p,j,m} - \xi_p) - \frac{1}{b}\sum_{1\le k\le b} m^{1/2}(\tilde{\xi}_{p,k,m} - \xi_p)\right)^2\right]^{1/2}} \\ &\Rightarrow \frac{b^{1/2}[(1/b)\sum_{1\le j\le b} M_j']}{[(1/(b-1))\sum_{1\le j\le b}(M_j' - \bar{M}_b')^2]^{1/2}}\end{aligned} \tag{69}$$

as $m \to \infty$ with $b \geq 2$ fixed. Because the limiting distribution (Student t) in (69) is continuous, we have that $P(\xi_p \in \check{C}_{b,m,\text{batch}}) = P\left(-t_{b-1,\alpha/2} \leq b^{1/2}[\acute{\xi}_{p,b,m} - \xi_p]/\check{S}_{b,m,\text{batch}} \leq t_{b-1,\alpha/2}\right) \to 1-\alpha$ as $n \to \infty$ by the portmanteau theorem (Theorem 2.1 of Billingsley (1999)).

Now we will prove the asymptotic validity of the srLHS sectioning CI $\check{C}_{b,m,\text{sec}}$ in (27). By (10), for each $j = 1, 2, \ldots, b$, the batch-j quantile estimator satisfies $\tilde{\xi}_{p,j,m} = \xi_p - [\tilde{F}_{j,m}(\xi_p) - p]/F'(\xi_p) + \tilde{R}_{j,m}$ with $m^{1/2}\tilde{R}_{j,m} \Rightarrow 0$ as $m \to \infty$. Letting $\acute{R}_{b,m} = (1/b)\sum_{1 \leq j \leq b} \tilde{R}_{j,m}$, we then get $\acute{\xi}_{p,b,m} = (1/b)\sum_{1 \leq j \leq b} \tilde{\xi}_{p,j,m} = \xi_p - \left[(1/b)\sum_{1 \leq j \leq b} \tilde{F}_{j,m}(\xi_p) - p\right]/F'(\xi_p) + \acute{R}_{b,m}$. In addition, $(1/b)\sum_{1 \leq j \leq b} \tilde{F}_{j,m}(\xi_p) = (1/(bm))\sum_{1 \leq j \leq b}\sum_{1 \leq i \leq m} I(\tilde{Y}_{j,i} \leq \xi_p) = \check{F}_{b,m}(\xi_p)$, so $\acute{\xi}_{p,b,m} = \xi_p - [\check{F}_{b,m}(\xi_p) - p]/F'(\xi_p) + \acute{R}_{b,m}$ where $m^{1/2}\acute{R}_{b,m} = (1/b)\sum_{1 \leq j \leq b} m^{1/2}\tilde{R}_{j,m} \Rightarrow 0$ as $m \to \infty$ with b fixed. Theorem 3 with $p_n = p$ then yields $m^{1/2}[\acute{\xi}_{p,b,m} - \check{\xi}_{p,b,m}] = m^{1/2}[\acute{R}_{b,m} - \check{R}_{b,m}] \Rightarrow 0$ as $m \to \infty$ with b fixed. Thus,

$$\frac{b^{1/2}[\check{\xi}_{p,b,m} - \xi_p]}{\check{S}_{b,m,\text{sec}}} = \frac{b^{1/2}\left[m^{1/2}(\check{\xi}_{p,b,m} - \acute{\xi}_{p,b,m}) + m^{1/2}(\acute{\xi}_{p,b,m} - \xi_p)\right]}{\left[(1/(b-1))\sum_{1 \leq j \leq b}\left(m^{1/2}(\tilde{\xi}_{p,j,m} - \acute{\xi}_{p,b,m}) + m^{1/2}(\acute{\xi}_{p,b,m} - \check{\xi}_{p,b,m})\right)^2\right]}$$
$$\Rightarrow \frac{b^{1/2}[(1/b)\sum_{1 \leq j \leq b} M'_j]}{[(1/(b-1))\sum_{1 \leq j \leq b}(M'_j - \bar{M}'_b)^2]^{1/2}}$$

as $m \to \infty$ with $b \geq 2$ fixed by (68), (69), Theorem 3.9 of Billingsley (1999), and the CMT. Since the above limit has a Student t distribution with $b-1$ degrees of freedom, we have $P(\xi_p \in \check{C}_{b,m,\text{sec}}) \to 1-\alpha$ as $m \to \infty$ with b fixed by the portmanteau theorem, completing the proof.

Acknowledgments

This work has been supported in part by the National Science Foundation under Grant No. CMMI-1200065. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Asmussen, S., P. Glynn. 2007. *Stochastic Simulation: Algorithms and Analysis*. Springer, New York.

Avramidis, A. N., J. R. Wilson. 1996. Correlation-induction techniques for estimating quantiles in simulation experiments. Tech. Rep. NCSU-IE 95-5, North Carolina State University.

Avramidis, A. N., J. R. Wilson. 1998. Correlation-induction techniques for estimating quantiles in simulation. *Operations Research* **46** 574–591.

Bahadur, R. R. 1966. A note on quantiles in large samples. *Annals of Mathematical Statistics* **37** 577–580.

Billingsley, P. 1995. *Probability and Measure*. 3rd ed. John Wiley and Sons, New York.

Billingsley, P. 1999. *Convergence of Probability Measures*. 2nd ed. John Wiley and Sons, New York.

Bloch, D. A., J. L. Gastwirth. 1968. On a simple estimate of the reciprocal of the density function. *Annals of Mathematical Statistics* **39** 1083–1085.

Bofinger, E. 1975. Estimation of a density function using order statistics. *Australian Journal of Statistics* **17** 1–7.

Chu, F., M. K. Nakayama. 2012a. Confidence intervals for quantiles when applying variance-reduction techniques. *ACM Transactions On Modeling and Computer Simulation* **36** Article 7.

Chu, F., M. K. Nakayama. 2012b. Confidence intervals for quantiles when applying variance-reduction techniques. *ACM Transactions On Modeling and Computer Simulation* **36** Article 7 (25 pages plus 12–page online–only appendix).

Dong, H., M. K. Nakayama. 2014. Constructing confidence intervals for a quantile using batching and sectioning when applying Latin hypercube sampling. *Proceedings of the 2014 Winter Simulation Conference*. IEEE, 640–651.

Falk, M. 1986. On the estimation of the quantile density function. *Statistics and Probability Letters* **4** 69–73.

Ghosh, J. K. 1971. A new proof of the Bahadur representation of quantiles and an application. *Annals of Mathematical Statistics* **42** 1957–1961.

Glasserman, P. 2004. *Monte Carlo Methods in Financial Engineering*. Springer, New York.

Glasserman, P., P. Heidelberger, P. Shahabuddin. 2000. Variance reduction techniques for estimating value-at-risk. *Management Science* **46** 1349–1364.

Glynn, P. W. 1996. Importance sampling for Monte Carlo estimation of quantiles. *Mathematical Methods in Stochastic Simulation and Experimental Design: Proceedings of the 2nd St. Petersburg Workshop on Simulation*. Publishing House of St. Petersburg Univ., St. Petersburg, Russia, 180–185.

Grabaskas, D., R. Denning, T. Aldemir, M. K. Nakayama. 2012. The use of Latin hypercube sampling for the efficient estimation of confidence intervals. *Proceedings of the 2012 International Congress on Advances in Nuclear Power Plants (ICAPP '12)*. 1443–1452.

Hall, P., M. A. Martin. 1988. Exact convergence rate of bootstrap quantile variance estimator. *Probability Theory and Related Fields* **80** 261–268.

Hall, P., M. A. Martin. 1989. A note on the accuracy of bootstrap percentile method confidence intervals for a quantile. *Statistics and Probability Letters* **8** 197–200.

Hall, P., S. J. Sheather. 1988. On the distribution of a Studentized quantile. *Journal of the Royal Statistical Society B* **50** 381–391.

Helton, J. C., F. J. Davis. 2003. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering and System Safety* **81** 23–69.

Hesterberg, T. C., B. L. Nelson. 1998. Control variates for probability and quantile estimation. *Management Science* **44** 1295–1312.

Hsu, J. C., B. L. Nelson. 1990. Control variates for quantile estimation. *Management Science* **36** 835–851.

Jin, X., M. C. Fu, X. Xiong. 2003. Probabilistic error bounds for simulation quantile estimation. *Management Science* **49** 230–246.

Juneja, S., R. Karandikar, P. Shahabuddin. 2007. Asymptotics and fast simulation for tail probabilities of maximum of sums of few random variables. *ACM Transactions on Modeling and Computer Simulation* **17** article 2, 35 pages.

Kaczynski, W. H., L. M. Leemis, J. H. Drew. 2012. Transient queueing analysis. *INFORMS Journal on Computing* **24** 10–28.

Liu, J., X. Yang. 2012. The convergence rate and asymptotic distribution of bootstrap quantile variance estimator for importance sampling. *Advances in Applied Probability* **44** 815–841.

Loh, W.-L. 1996. On Latin hypercube sampling. *Annals of Statistics* **24** 2058–2080.

McKay, M. D., W. J. Conover, R. J. Beckman. 1979. A comparison of three methods for selecting input variables in the analysis of output from a computer code. *Technometrics* **21** 239–245.

Nakayama, M. K. 2011. Asymptotically valid confidence intervals for quantiles and values-at-risk when applying Latin hypercube sampling. *International Journal on Advances in Systems and Measurements* **4** 86–94.

Nakayama, M. K. 2012. Confidence intervals for quantiles when applying replicated Latin hypercube sampling and sectioning. J. Nutaro, R. J. Gay, eds., *Proceedings of the Energy, Climate and Environment*

Modeling & Simulation 2012 Conference (ECEMS 2012), Autumn Simulation Multiconference, *Simulation*, vol. 44. The Society for Modeling and Simulation International, San Diego, CA, 12–19.

Nakayama, M. K. 2014. Confidence intervals using sectioning for quantiles when applying variance-reduction techniques. *ACM Transactions on Modeling and Computer Simulation* **24** Article 19.

Owen, A. B. 1992. A central limit theorem for Latin hypercube sampling. *Journal of the Royal Statistical Society B* **54** 541–551.

Owen, A. B. 1997. Monte Carlo variance of scrambled net quadrature. *SIAM Journal of Numerical Analysis* **34** 1884–1910.

Serfling, R. J. 1980. *Approximation Theorems of Mathematical Statistics*. John Wiley and Sons, New York.

Stein, M. 1987. Large sample properties of simulations using Latin hypercube sampling. *Technometrics* **29** 143–151. Correction 32:367.

Sun, L., L. J. Hong. 2010. Asymptotic representations for importance-sampling estimators of value-at-risk and conditional value-at-risk. *Operations Research Letters* **38** 246–251.

U.S. Nuclear Regulatory Commission. 2010. Acceptance criteria for emergency core cooling systems for light-water nuclear power reactors. Title 10, Code of Federal Regulations §50.46, NRC, Washington, DC.

U.S. Nuclear Regulatory Commission. 2011. Applying statistics. U.S. Nuclear Regulatory Commission Report NUREG-1475, Rev 1, U.S. Nuclear Regulatory Commission, Washington, DC.

Williams, D. 1991. *Probability with Martingales*. Cambridge University Press.

Hui Dong is currently a research scientist at Amazon.com Corporate LLC, Seattle. She received her PhD in Management at Rutgers University. Her research interests include variance reduction in simulation, supply chain management, e-commerce logistics. She recently extended her interests into deep learning and artificial intelligence.

Marvin K. Nakayama is a professor of computer science at the New Jersey Institute of Technology. His research interests include simulation, modeling, applied probability, statistics, risk analysis, and energy.