Proceedings of the 2017 Winter Simulation Conference
W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds.

QUANTILE ESTIMATION USING CONDITIONAL MONTE CARLO AND LATIN HYPERCUBE SAMPLING

Hui Dong

Amazon.com Corporate LLC
Seattle, WA 98109, USA

Marvin K. Nakayama

Department of Computer Science
New Jersey Institute of Technology
Newark, NJ 07102, USA

ABSTRACT

Quantiles are often employed to measure risk. We combine two variance-reduction techniques, conditional Monte Carlo and Latin hypercube sampling, to estimate a quantile. Compared to either method by itself, the combination can produce a quantile estimator with substantially smaller variance. In addition to devising a point estimator for the quantile when applying the combined approaches, we also describe how to construct confidence intervals for the quantile. Numerical results demonstrate the effectiveness of the methods.

1 INTRODUCTION

For a random variable Y and fixed $0 < p < 1$, the p-quantile of Y is a constant ξ such that p of the mass of the distribution of Y lies below ξ. The median, which is the 0.5-quantile, provides a well-known example. Many different fields measure risk through a p-quantile with p close to 0 or 1. In finance, a quantile is called a value-at-risk, and it is employed to assess portfolio risk. For example, the Basel II Accord (Basel Committee on Banking Supervision 2004) specifies capital requirements in terms of 0.99-quantiles. The U.S. Nuclear Regulatory Commission (NRC) permits nuclear-plant licensees to use a 0.95-quantile to demonstrate compliance with federal regulations (U.S. Nuclear Regulatory Commission 2010).

Our paper studies Monte Carlo simulation methods for quantile estimation. In addition to giving a point estimate for a quantile, we also want to specify a confidence interval (CI) for ξ to account for sampling error. Indeed, the NRC further requires providing a 95% CI for a 0.95-quantile (Section 3.2 of U.S. Nuclear Regulatory Commission 2005; Section 24.9 of U.S. Nuclear Regulatory Commission 2011).

Quantile estimation via simple random sampling (SRS) is extensively studied (Chapter 2 of Serfling 1980), but SRS may result in an unusably wide quantile CI. Instead, we apply variance-reduction techniques (VRTs) to estimate and construct a CI for ξ; see Chapter V of Asmussen and Glynn (2007) for an overview of VRTs to estimate a mean. Hong, Hu, and Liu (2014) survey simulation methods for quantile estimation. We focus on two well-known VRTs: Latin hypercube sampling (LHS; McKay, Beckman, and Conover 1979) and conditional Monte Carlo (CMC; Section V.4 of Asmussen and Glynn 2007).

Simultaneously stratifying each input coordinate of the function used to produce the random output, LHS is known to yield estimates of a mean with substantially reduced variance when the simulated response is nearly an additive function of the input variables. For quantile estimation, the relevant response corresponds to the indicator function that the output variable lies below the quantile. But an additive approximation to the indicator function typically has a poor fit, so LHS by itself does not reduce variance by much. CMC replaces the indicator with its conditional expectation, which is a smoother response. As the additive fit is now often much better, the combination of CMC and LHS can greatly reduce variance, as our numerical results show. Avramidis and Wilson (1996), Nakayama (2015), and Alban et al. (2017) observe similar synergies between CMC and LHS when estimating a probability.

The rest of the paper progresses as follows. Section 2 provides an overview of quantile estimation. We review how to use simple random sampling and LHS to estimate a quantile in Sections 3 and 4, respectively.

We explain in Section 5 quantile estimation via CMC, and we combine CMC with LHS in Section 6. Section 7 contains numerical results, and we close with some concluding remarks in Section 8.

2 OVERVIEW OF QUANTILE ESTIMATION

Let $U[0,1]$ denote a uniform distribution on $[0,1]$, and let $U_1, U_2, \ldots, U_d$ be d independent and identically distributed (IID) $U[0,1]$ random numbers. Then for a deterministic function $c_Y : [0,1]^d \to \Re$, define

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

We can think of c_Y as being a computer code that, when fed d IID uniform random numbers as input, produces an output random variable Y. The computer code c_Y may first convert the uniforms into a random vector $W = (W_1, W_2, \ldots, W_k)$ having some joint distribution, with possible dependence among the entries of W, and then performs further computations using W to obtain the output Y. For example, nuclear engineers employ computer codes in this manner (Hess et al. 2009) to perform probabilistic safety assessments of nuclear power plants. Each run of the computer code may require substantial computation time; e.g., it may numerically solve systems of differential equations. This motivates applying VRTs.

Let F be the cumulative distribution function (CDF) of Y, denoted by $Y \sim F$. We assume that the function c_Y is sufficiently complicated that we cannot compute F directly, but we can sample $Y \sim F$ by (1). For a fixed constant $0 < p < 1$, we define the p-quantile of F (or equivalently of Y) as $\xi = F^{-1}(p) \equiv \inf\{y : F(y) \geq p\}$. The goal is to use Monte Carlo simulation to estimate and construct a confidence interval (CI) for ξ.

The typical approach to estimate ξ via simulation first applies some Monte Carlo method with a sample size of n to construct an estimator $\hat{F}_n$ of F. Because the true p-quantile satisfies $\xi = F^{-1}(p)$, we then invert $\hat{F}_n$ to get a quantile estimator $\hat{\xi}_n = \hat{F}_n^{-1}(p)$. This idea works for many Monte Carlo methods, including simple random sampling, conditional Monte Carlo, and Latin hypercube sampling, among others.

Although the quantile estimator $\hat{\xi}_n$ is *not* a sample average, it still typically satisfies a central limit theorem (CLT). Proving a CLT for $\hat{\xi}_n$ requires that $f(\xi) > 0$, where f is the derivative (when it exists) of F. One way to establish a CLT for $\hat{\xi}_n$ employs a Bahadur (1966) representation, which roughly states that $\hat{\xi}_n \approx \xi + [p - \hat{F}_n(\xi)]/f(\xi)$, so $\hat{\xi}_n$ can be approximated as a linear transformation of $\hat{F}_n(\xi)$. As $\hat{F}_n(\xi)$ is usually a sample average, it obeys a CLT, which implies $\hat{\xi}_n$ also does.

To prove his representation for SRS, Bahadur (1966) first shows that $\hat{F}_n(y) \approx \hat{F}_n(\xi) + F(y) - F(\xi)$ uniformly for y in some neighborhood B_n of ξ, and that $\hat{\xi}_n \in B_n$ for all sufficiently large n. Thus, we get

$$p \approx \hat{F}_n(\hat{\xi}_n) \approx \hat{F}_n(\xi) + F(\hat{\xi}_n) - F(\xi) \approx \hat{F}_n(\xi) + f(\xi)[\hat{\xi}_n - \xi],$$

where the last step holds by a Taylor approximation. Rearranging terms, Bahadur (1966) formalizes the argument when $f(\xi) > 0$ and F has bounded second derivative in a neighborhood of ξ, establishing that

$$\hat{\xi}_n = \xi + [p - \hat{F}_n(\xi)]/f(\xi) + R_n, \tag{2}$$

where the remainder $R_n = O(n^{-3/4} \log n)$ almost surely (a.s.) as $n \to \infty$. Only assuming that $f(\xi) > 0$, Ghosh (1971) shows the remainder satisfies $\sqrt{n}R_n \Rightarrow 0$ as $n \to \infty$, where $\Rightarrow$ denotes convergence in distribution (Chapter 5 of Billingsley 1995). Sun and Hong (2010) prove an a.s. rate at which the remainder R_n vanishes in (2) for an importance-sampling quantile estimator. Chu and Nakayama (2012) establish that $\sqrt{n}R_n \Rightarrow 0$ under a general set of conditions on the Monte Carlo method applied, which is satisfied by many VRTs, including importance sampling, stratified sampling, antithetic variates, and control variates. Dong and Nakayama (2017b) verify that the conditions of Chu and Nakayama (2012) hold for LHS, so the LHS quantile estimator obeys the Bahadur representation in (2) with $\sqrt{n}R_n \Rightarrow 0$.

Returning now to establishing a CLT for the quantile estimator $\hat{\xi}_n$, we rearrange (2) and multiply through by $\sqrt{n}$ to obtain

$$\sqrt{n}[\hat{\xi}_n - \xi] = \sqrt{n}[p - \hat{F}_n(\xi)]/f(\xi) + \sqrt{n}R_n. \tag{3}$$

Now assume that the CDF estimator $\hat{F}_n$ satisfies a CLT

$$\sqrt{n}[p - \hat{F}_n(\xi)] \Rightarrow N(0, \psi^2) \tag{4}$$

as $n \to \infty$, where $N(a, s^2)$ denotes a normal random variable with mean a and variance s^2, and ψ^2 is the *asymptotic variance* in the CLT (4). For many Monte Carlo methods, (4) holds (under appropriate regularity conditions) because $\hat{F}_n$ is a sample average. Hence, putting (4) into (3) and using that $\sqrt{n}R_n \Rightarrow 0$ as $n \to \infty$ from a Bahadur representation, Slutsky's theorem (p. 19 of Serfling 1980) ensures

$$\sqrt{n}[\hat{\xi}_n - \xi] \Rightarrow N(0, \tau^2) \tag{5}$$

as $n \to \infty$, where the asymptotic variance in (5) satisfies

$$\tau^2 = \psi^2 / f^2(\xi). \tag{6}$$

The numerator ψ^2, which is the asymptotic variance in the CLT for $\hat{F}_n(\xi)$ in (4), depends on the Monte Carlo method employed to estimate the CDF, but the denominator $f^2(\xi)$ in (6) is the same for all Monte Carlo methods, as long as the technique first estimates the CDF and then inverts it to get a quantile estimator. Thus, we want VRTs that reduce the asymptotic variance ψ^2 of the estimator $\hat{F}_n(\xi)$ of $F(\xi)$.

We can also exploit the Bahadur representation (2) with $\sqrt{n}R_n \Rightarrow 0$ and the quantile estimator CLT (5) to construct various confidence intervals for ξ. If we have a *consistent* estimator $\hat{\tau}_n^2$ of the asymptotic variance τ^2 in (6) (i.e., $\hat{\tau}_n^2 \Rightarrow \tau^2$ as $n \to \infty$), then we can unfold the CLT in (5) to obtain an asymptotic γ-level CI for ξ as

$$J_n = (\hat{\xi}_n \pm z_\gamma \hat{\tau}_n / \sqrt{n}), \tag{7}$$

where the constant $z_\gamma = \Phi^{-1}(1 - (1-\gamma)/2)$ is the upper $(1-(1-\gamma)/2)$-critical point of a standard (mean 0 and variance 1) normal (e.g., $z_{0.95} = 1.96$), and Φ is the $N(0,1)$ CDF. The CI J_n, which is asymptotically valid in the sense that its coverage $P(\xi \in J_n) \to \gamma$ as $n \to \infty$, critically depends on $\hat{\tau}_n^2$ consistently estimating τ^2. While this can be accomplished by constructing separate consistent estimators for the numerator and denominator in (6) (e.g., see Bloch and Gastwirth 1968, Chu and Nakayama 2012, and Dong and Nakayama 2017b), devising such consistent estimators in practice can be nontrivial.

We can avoid the complications of consistently estimating τ^2 in (6) to construct a CI for ξ by instead applying *batching* (also known as subsampling) or *sectioning*. In each method, the basic idea is to divide the overall sample of size n into $b \geq 2$ (nonoverlapping) batches (e.g., $b = 10$), each of size $m = n/b$. For each batch $r = 1, 2, \ldots, b$, let $\hat{F}_{r,m}$ be the estimator of CDF F from batch r, and compute the corresponding p-quantile estimator $\hat{\xi}_{r,m} = \hat{F}_{r,m}^{-1}(p)$. Also, let $\bar{\xi}_{b,m} = (1/b)\sum_{r=1}^{b} \hat{\xi}_{r,m}$ and $S_{b,m}^2 = (1/(b-1))\sum_{r=1}^{b}(\hat{\xi}_{r,m} - \bar{\xi}_{b,m})^2$ be the sample average and sample variance of the $\hat{\xi}_{r,m}$, $1 \leq r \leq b$. Then we obtain

$$J_{b,n} = (\bar{\xi}_{b,m} \pm t_{b-1,\gamma} S_{b,m} / \sqrt{b}) \tag{8}$$

as an asymptotic γ-level batching CI, where the critical point $t_{b-1,\gamma}$ satisfies $P(T_{b-1} \leq t_{b-1,\gamma}) = 1 - (1-\gamma)/2$ for T_{b-1} a Student-t random variable with $b-1$ degrees of freedom.

While the batching CI in (8) is asymptotically valid ($\lim_{n\to\infty} P(\xi \in J_{b,n}) = \gamma$), it may have poor coverage when the overall sample size n is not large; i.e., $P(\xi \in J_{b,n})$ may differ substantially from the nominal confidence level γ. The problem arises because quantile estimators are generally biased: $E[\hat{\xi}_n] \neq \xi$ for fixed n. Although the bias of the batching point estimator $\bar{\xi}_{b,m}$ vanishes as $n \to \infty$, it can be substantial for fixed n as it is determined by the batch size $m = n/b$, which is small when the overall sample size n is not large. Hence, the batching CI in (8) is not properly centered on average, which can cause poor coverage.

Sectioning, originally proposed in Section III.5a of Asmussen and Glynn (2007) for SRS and extended to some VRTs by Nakayama (2014a), addresses this shortcoming by instead centering the CI at a point

estimator that is typically less biased than $\bar{\xi}_{b,m}$. Let $\tilde{F}_{b,m}(y) = (1/b)\sum_{r=1}^{b} \hat{F}_{r,m}(y)$ be the sectioning CDF estimator constructed from all $n = bm$ sampled outputs across the b batches, and compute the sectioning p-quantile estimator as $\tilde{\xi}_{b,m} = \tilde{F}_{b,m}^{-1}(p)$. Many Monte Carlo methods have $\tilde{F}_{b,m}(y) = \hat{F}_n(y)$ and $\tilde{\xi}_{b,m} = \hat{\xi}_n$, but this is not always the case, e.g., as for LHS. Then the sectioning CI is

$$\tilde{J}_{b,n} = (\tilde{\xi}_{b,m} \pm t_{b,\gamma} S_{b,m}/\sqrt{b}), \tag{9}$$

which can be shown to be asymptotically valid when the Bahadur representation (2) holds with $\sqrt{n}R_n \Rightarrow 0$ as $n \to \infty$. The key point is that because $\tilde{J}_{b,n}$ is better centered on average than $J_{b,n}$ in (8), the sectioning CI often has coverage closer to γ than does the batching CI when the overall sample size n is not very large.

3 SIMPLE RANDOM SAMPLING

Chapter 2 of Serfling (1980) provides a thorough treatment of SRS quantile estimation, which we now review for our context when the output random variable Y is defined by (1). We first generate n IID copies of $Y \sim F$ as follows. Let $U_{i,j}$, $1 \le i \le n$, $1 \le j \le d$, be IID $U[0,1]$ random numbers, which we arrange in an $n \times d$ grid. We then apply the function c_Y from (1) to each row to obtain the sample $Y_1, Y_2, \ldots, Y_n$, with

$$\begin{array}{rcccccc} Y_1 & = & c_Y(U_{1,1}, & U_{1,2}, & \ldots, & U_{1,d}), \\ Y_2 & = & c_Y(U_{2,1}, & U_{2,2}, & \ldots, & U_{2,d}), \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_n & = & c_Y(U_{n,1}, & U_{n,2}, & \ldots, & U_{n,d}). \end{array} \tag{10}$$

Each $Y_i \sim F$ by (1) because row i of (10) has d IID uniform random numbers $U_{i,j}$, $1 \le j \le d$. Moreover, the n rows of uniforms in (10) are independent, so $Y_1, Y_2, \ldots, Y_n$, are independent.

The CDF F of Y satisfies $F(y) = P(Y \le y) = E[I(Y \le y)]$, where $I(\cdot)$ denotes the indicator function, which equals 1 (resp., 0) when its argument is true (resp., false). Thus, we can obtain an unbiased estimator of $F(y)$ by averaging identically distributed copies of $I(Y \le y)$. When SRS is used, the CDF estimator (at y) is $\hat{F}_{\mathrm{SRS},n}(y) = (1/n)\sum_{i=1}^{n} I(Y_i \le y)$. The SRS p-quantile estimator is then $\hat{\xi}_{\mathrm{SRS},n} = \hat{F}_{\mathrm{SRS},n}^{-1}(p)$. An equivalent way to compute $\hat{\xi}_{\mathrm{SRS},n}$ first sorts the sample to get $Y_{1:n} \le Y_{2:n} \le \cdots \le Y_{n:n}$, and then sets $\hat{\xi}_{\mathrm{SRS},n} = Y_{\lceil np \rceil:n}$, where $\lceil \cdot \rceil$ is the ceiling function. (For simplicity, we consider only the quantile estimator $\hat{\xi}_{\mathrm{SRS},n}$, but variants can also be constructed, e.g., by interpolating the CDF estimator; see Avramidis and Wilson 1998.)

If $f(\xi) > 0$, the SRS p-quantile estimator $\hat{\xi}_{\mathrm{SRS},n}$ obeys the CLT in (5) (Section 2.3.3 of Serfling 1980). The numerator ψ^2_{SRS} of the asymptotic variance τ^2_{SRS} in (6) of the SRS quantile estimator appears as the asymptotic variance in the CLT for $\hat{F}_n(\xi) = \hat{F}_{\mathrm{SRS},n}(\xi)$ in (4), which averages IID $I(Y_i \le \xi)$, $i = 1, 2, \ldots, n$. Hence, we see that

$$\psi^2_{\mathrm{SRS}} = \mathrm{Var}[I(Y \le \xi)] = p(1-p), \tag{11}$$

and the asymptotic variance in (6) of the SRS CLT for the p-quantile estimator $\hat{\xi}_{\mathrm{SRS},n}$ is $\tau^2_{\mathrm{SRS}} = p(1-p)/f^2(\xi)$.

Bloch and Gastwirth (1968) apply a finite difference to consistently estimate $1/f(\xi) = \frac{d}{dp}F^{-1}(p)$, which can then be used to build the CI in (7). For the batching and sectioning CIs, let batch $r = 1, 2, \ldots, b$, consist of outputs $Y_{(r-1)m+i}$, $1 \le i \le m = n/b$. Then $\hat{F}_{\mathrm{SRS},r,m}(y) = (1/m)\sum_{i=1}^{m} I(Y_{(r-1)m+i} \le y)$ is the CDF estimator at y for batch r, and $\hat{\xi}_{\mathrm{SRS},r,m} = \hat{F}_{\mathrm{SRS},r,m}^{-1}(p)$ is the corresponding p-quantile estimator. Then (8) and (9) with $\hat{\xi}_{r,m} = \hat{\xi}_{\mathrm{SRS},r,m}$ and $\tilde{\xi}_{b,m} = \hat{\xi}_{\mathrm{SRS},n}$ give the batching and sectioning CIs, respectively, for SRS.

4 LATIN HYPERCUBE SAMPLING

McKay, Beckman, and Conover (1979) devised LHS, which separately stratifies each input coordinate of (1). Reducing variance by producing negatively correlated outputs, the method can be thought of as

an efficient way of extending stratified sampling to high dimensions. Stein (1987) further analyzes the technique, and Owen (1992) proves a CLT for the LHS estimator of the mean of bounded outputs.

We implement LHS to generate a dependent sample of size n as follows. For each input coordinate $j = 1,2,\ldots,d$, of (1), let $\pi_j = (\pi_j(1),\pi_j(2),\ldots,\pi_j(n))$ be a random permutation of $(1,2,\ldots,n)$. In other words, each of the $n!$ permutations is equally likely, and $\pi_j(i)$ is the integer to which i is mapped in permutation π_j. Then for each $1 \le i \le n$ and $1 \le j \le d$, define

$$V_{i,j} = \frac{\pi_j(i) - 1 + U_{i,j}}{n} \tag{12}$$

for $U_{i,j}$ in (10). Next arrange the $V_{i,j}$ into an $n \times d$ grid, and apply function c_Y from (1) to each row to get

$$\begin{array}{rcccccc} Y_1' &=& c_Y(V_{1,1}, & V_{1,2}, & \ldots, & V_{1,d}), \\ Y_2' &=& c_Y(V_{2,1}, & V_{2,2}, & \ldots, & V_{2,d}), \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_n' &=& c_Y(V_{n,1}, & V_{n,2}, & \ldots, & V_{n,d}). \end{array} \tag{13}$$

It is straightforward to show that in each row i of (13), the d entries $V_{i,j}$, $1 \le j \le d$, are IID $U[0,1]$, so $Y_i' \sim F$ by (1). But as all of the n entries $V_{i,j}$, $1 \le i \le n$, in each column j of (13) share the same permutation π_j, the rows of (13) are dependent, making $Y_1', Y_2', \ldots, Y_n'$ *dependent*. As seen from (12), each column j in (13) forms a stratified sample of size n of the unit interval, i.e., with probability 1,

$$\text{for each } 1 \le j \le d \text{ and } 1 \le k \le n, \text{ exactly one } V_{i,j},\ 1 \le i \le n, \text{ lies in subinterval } [(k-1)/n, k/n]. \tag{14}$$

By (12), the one $V_{i,j}$ from column j lying in $[(k-1)/n, k/n]$ is uniformly distributed within that subinterval.

The LHS estimator of the CDF $F(y)$ is

$$\hat{F}_{\mathrm{LHS},n}(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i' \le y), \tag{15}$$

which, because each $Y_i' \sim F$, is an unbiased estimator of $F(y)$, despite $Y_1', Y_2', \ldots, Y_n'$ being dependent. We then obtain the LHS p-quantile estimator as $\hat{\xi}_{\mathrm{LHS},n} = \hat{F}^{-1}_{\mathrm{LHS},n}(p)$. Avramidis and Wilson (1998) prove that the LHS quantile estimator $\hat{\xi}_{\mathrm{LHS},n}$ obeys the CLT in (5), and the next paragraph will give an expression for the numerator ψ^2_{LHS} in the LHS CLT's asymptotic variance (6). Dong and Nakayama (2014) construct the LHS batching and sectioning CIs in (8) and (9), respectively, for ξ when applying replicated LHS (Iman 1981), in which the $b \ge 2$ batches are independent LHS samples, each of size $m = n/b$. Dong and Nakayama (2017b) provide consistent estimators of ψ^2_{LHS} and τ^2_{LHS} from a single LHS sample of size n by extending a variance estimator of Owen (1992) to estimate the numerator ψ^2_{LHS} of (6) and using a finite difference to handle $1/f(\xi) = \frac{d}{dp}F^{-1}(p)$ from the denominator, enabling the construction of the CI in (7).

We now give an expression for the numerator $\psi^2 = \psi^2_{\mathrm{LHS}}$ in (6); recall that the denominator $f^2(\xi)$ in (6) remains the same for each simulation method. Also recall that ψ^2_{LHS} comes from the CLT for $\hat{F}_n(\xi) = \hat{F}_{\mathrm{LHS},n}(\xi)$ in (4), which holds by the LHS CLT of Owen (1992) for the average of bounded responses. Deriving an expression for ψ^2_{LHS} is complicated by the fact that $\hat{F}_{\mathrm{LHS},n}(\xi)$ in (15) with $y = \xi$ averages *dependent* copies of the SRS response $I(Y \le \xi)$. By (1), we can write the SRS response as

$$I(Y \le \xi) = I(c_Y(U_1, U_2, \ldots, U_d) \le \xi) \equiv w(U_1, U_2, \ldots, U_d) \tag{16}$$

for $U_1, U_2, \ldots, U_d$ IID $U[0,1]$, and $E[w(U_1, U_2, \ldots, U_d)] = E[I(Y \le \xi)] = F(\xi)$. Applying an analysis-of-variance (ANOVA) decomposition as in Stein (1987) and Avramidis and Wilson (1998), we express $w(U_1, U_2, \ldots, U_d)$ using an *additive* (or *separable*) approximation

$$w(U_1, U_2, \ldots, U_d) = F(\xi) + \sum_{j=1}^{d} w_j(U_j) + \varepsilon(U_1, U_2, \ldots, U_d), \tag{17}$$

where $w_j(U_j) \equiv E[w(U_1,U_2,\ldots,U_d)\,|\,U_j] - F(\xi)$ is a function of only U_j as the other U_l, $l \neq j$, have been integrated out through the conditional expectation, and the residual $\varepsilon(U_1,U_2,\ldots,U_d)$ is defined so that the equality in (17) holds. Then Avramidis and Wilson (1998) show that

$$\psi^2_{\text{LHS}} = \text{Var}[\varepsilon(U_1,U_2,\ldots,U_d)] = \psi^2_{\text{SRS}} - \sum_{j=1}^{d} \text{Var}[w_j(U_j)], \tag{18}$$

with $\psi^2_{\text{SRS}} = p(1-p)$ by (11). Thus, LHS removes the variance from the additive part of the SRS response.

To gain additional insight into why (18) holds, we adapt to our setting the following heuristic argument from Section 10.3 of Owen (2013). We want to determine the asymptotic variance ψ^2_{LHS} of $\hat{F}_{\text{LHS},n}(\xi)$, so the ith summand in (15) for $y = \xi$ becomes $I(Y'_i \leq \xi) = w(V_{i,1},V_{i,2},\ldots,V_{i,d})$ by (13) and (16). Putting (17) with each U_j replaced with $V_{i,j}$ into $\hat{F}_{\text{LHS},n}(\xi)$ in (15) then leads to

$$\hat{F}_{\text{LHS},n}(\xi) = \frac{1}{n}\sum_{i=1}^{n} w(V_{i,1},V_{i,2},\ldots,V_{i,d}) = F(\xi) + \sum_{j=1}^{d}\frac{1}{n}\sum_{i=1}^{n} w_j(V_{i,j}) + \frac{1}{n}\sum_{i=1}^{n}\varepsilon(V_{i,1},V_{i,2},\ldots,V_{i,d}). \tag{19}$$

On the right side of (19), the middle term $\bar{w}_{j,n} \equiv (1/n)\sum_{i=1}^{n} w_j(V_{i,j})$ for each j estimates the 1-dimensional integral $\zeta_j \equiv \int_0^1 w_j(u)\,du$. But by (14), LHS stratifies the unit interval of each input coordinate j, so $\bar{w}_{j,n}$ is a (random) Riemann approximation to the integral ζ_j. When the integrand w_j is sufficiently smooth, a 1-dimensional Riemann sum converges much faster than the ordinary Monte Carlo rate of $n^{-1/2}$. On the other hand, the last term in (19), which is the average of the residuals, converges at rate $n^{-1/2}$. Thus, the asymptotic variance of $\hat{F}_{\text{LHS},n}(\xi)$ is determined by the variance of the residual, as in (18).

As a consequence, when the response is nearly additive, in the sense that the residual $\varepsilon(U_1,U_2,\ldots,U_d)$ in (17) has small variance, LHS can substantially lower variance. But because the response in (16) is an indicator function, the additive approximation in (17) is typically a poor fit, leading LHS by itself to not reduce variance by much. To address this, we next introduce conditional Monte Carlo, which replaces the response in (16) with its conditional expectation, which can be a smoother function of its inputs.

5 CONDITIONAL MONTE CARLO

CMC is a VRT that reduces variance by analytically integrating out some of the variability through iterated expectations; e.g., see Section V.4 of Asmussen and Glynn (2007) for an overview of CMC for estimating a mean. We now review CMC quantile estimation, as developed in Nakayama (2014b). Let X be an auxiliary random vector that is generated along with the output Y in a simulation run. (We later give in (27) the assumed exact form of X as it relates to Y defined by (1).) Using iterated expectations (e.g., p. 448 of Billingsley 1995) by conditioning on X, we express the CDF F of Y at each y as

$$F(y) = E[I(Y \leq y)] = E[E[I(Y \leq y)\,|\,X]] = E[q(X,y)], \tag{20}$$

$$\text{where} \quad q(x,y) = E[I(Y \leq y)\,|\,X = x] = P(Y \leq y\,|\,X = x). \tag{21}$$

Hence, (20) implies that averaging IID copies of $q(X,y)$ produces an unbiased estimator of $F(y)$. Specifically, for IID copies $X_1,X_2,\ldots,X_n$ of conditioning vector X, the CMC estimator of the CDF F of Y is given by

$$\hat{F}_{\text{CMC},n}(y) = \frac{1}{n}\sum_{i=1}^{n} q(X_i,y), \tag{22}$$

which assumes that the conditional expectation $q(x,y)$ in (21) can be computed (analytically or numerically) for each possible x and y. We finally invert $\hat{F}_{\text{CMC},n}$ to obtain the CMC p-quantile estimator

$$\hat{\xi}_{\text{CMC},n} = \hat{F}^{-1}_{\text{CMC},n}(p). \tag{23}$$

Computing $\hat{\xi}_{\mathrm{CMC},n}$ typically requires applying an iterative root-finding algorithm, such as the Newton, secant, and false-position methods; see Chapter 7 of Ortega and Rheinboldt (1987).

The benefit of CMC quantile estimation becomes apparent by examining the asymptotic variance τ^2_{CMC} in (6) of $\hat{\xi}_{\mathrm{CMC},n}$. When CMC is applied, the numerator ψ^2_{CMC} of (6) is the asymptotic variance in the CLT in (4) for $\hat{F}_{\mathrm{CMC},n}(\xi)$. Because $\hat{F}_{\mathrm{CMC},n}(\xi)$ in (22) averages IID $q(X_i,\xi)$, $1 \le i \le n$, we see that

$$\psi^2_{\mathrm{CMC}} = \mathrm{Var}[q(X,\xi)] = \mathrm{Var}[E[I(Y \le \xi)\,|\,X]] \tag{24}$$

$$\le \mathrm{Var}[E[I(Y \le \xi)\,|\,X]] + E[\mathrm{Var}[I(Y \le \xi)\,|\,X]] = \mathrm{Var}[I(Y \le \xi)] = \psi^2_{\mathrm{SRS}} \tag{25}$$

from a variance decomposition (e.g., p. 456 of Billingsley 1995) and (11). As a consequence, we have

$$\tau^2_{\mathrm{CMC}} = \psi^2_{\mathrm{CMC}}/f^2(\xi) \le \psi^2_{\mathrm{SRS}}/f^2(\xi) = \tau^2_{\mathrm{SRS}}.$$

In other words, the asymptotic variance in (6) of the CLT for the CMC p-quantile estimator $\hat{\xi}_{\mathrm{CMC},n}$ is no greater than that of the SRS quantile estimator $\hat{\xi}_{\mathrm{SRS},n}$.

But as noted earlier, applying CMC critically relies on being able to compute (analytically or numerically) the conditional expectation $q(x,y)$ in (21) for each x and y, which can be done in certain settings. A simple example when this may be possible arises in structural reliability (Section 1.4.2 of Melchers 1999). Let C and L be independent $\Re$-valued random variables denoting the capacity and load, respectively, of a structure (e.g., a bridge), and $Y = C - L$ is the structure's safety margin. Let $X = L$ be the (scalar) conditioning vector, and assume that the CDF G of C is (analytically or numerically) computable. Then (21) becomes

$$q(x,y) = P(C - L \le y\,|\,L = x) = P(C \le x + y\,|\,L = x) = G(x+y) \tag{26}$$

by the independence of C and L, so $q(x,y)$ is computable.

Returning now to the general setting, we next explain our assumed form of the conditioning vector X in (20) for the framework of (1). Assume that X is $\Re^{d'_X}$-valued, where $d'_X \ge 1$, and recall that the function c_Y in (1) uses d IID $U[0,1]$ inputs $U_1, U_2, \ldots, U_d$ to generate $Y \in \Re$. Then we suppose that there exist two more deterministic functions $c_* : [0,1]^d \to \Re^{1+d'_X}$ and $c_X : [0,1]^{d_X} \to \Re^{d'_X}$ with $1 \le d_X \le d$ such that

$$(Y,X) = c_*(U_1,U_2,\ldots,U_d) = (c_Y(U_1,U_2,\ldots,U_d), c_X(U_1,U_2,\ldots,U_{d_X})). \tag{27}$$

Thus, the function c_* produces both the original output variable $Y \in \Re$ and the conditioning vector $X \in \Re^{d'_X}$ from the *same* d IID uniforms, so Y and X are dependent. Moreover, the $\Re^{1+d'_X}$-valued function c_* decomposes into two functions c_Y and c_X, where the $\Re^1$-valued function c_Y in (1) computes Y from all d uniforms, and the $\Re^{d'_X}$-valued function c_X generates X using only the first $d_X \le d$ of them, which may require relabeling the inputs. Avramidis and Wilson (1996) make a similar assumption in their study of combined VRTs to estimate a mean.

Now for the framework of (27), we explain how to construct the CMC quantile estimator $\hat{\xi}_{\mathrm{CMC},n}$ in (23). Because the CMC CDF estimator $\hat{F}_{\mathrm{CMC},n}$ in (22) averages IID copies of $q(X,y)$, we need to generate IID copies of only the conditioning vector X and not the original output Y, which has been (partially) integrated out through the conditional expectation in (21). Thus, to generate n IID copies of X, (27) implies that we can use just the first d_X columns of the grid of IID $U[0,1]$ numbers $U_{i,j}$ in (10), and apply function c_X from (27) to each row to obtain the sample $X_1, X_2, \ldots, X_n$, with

$$\begin{array}{rcccccc} X_1 &=& c_X(U_{1,1}, & U_{1,2}, & \ldots, & U_{1,d_X}), \\ X_2 &=& c_X(U_{2,1}, & U_{2,2}, & \ldots, & U_{2,d_X}), \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ X_n &=& c_X(U_{n,1}, & U_{n,2}, & \ldots, & U_{n,d_X}). \end{array} \tag{28}$$

The first d_X of the $U_{i,j}$ in each row i of (10) are IID $U[0,1]$, so $X_i \stackrel{\mathscr{D}}{=} X$ by (27), where $\stackrel{\mathscr{D}}{=}$ denotes equality in distribution. Further, the independence of the rows of $U_{i,j}$ in (10) ensures that $X_1, X_2, \ldots, X_n$ are independent. We then put $X_1, X_2, \ldots, X_n$ in (22) to compute $\hat{F}_{\mathrm{CMC},n}$, which is inverted to obtain $\hat{\xi}_{\mathrm{CMC},n}$ in (23).

As for a CMC CI for ξ, estimating the variance ψ^2_{CMC} in (24) is complicated by the fact that $q(X,\xi)$ is not directly observable because ξ is unknown. Although it is possible to construct consistent estimators of ψ^2_{CMC} and $f(\xi)$ to build the CI in (7) for ξ when applying CMC, we can avoid the issue by instead using batching or sectioning. Specifically, divide the sample $X_1, X_2, \ldots, X_n$, into $b \geq 2$ batches, where batch $r = 1, 2, \ldots, b$, comprises $X_{(r-1)m+i}$, $1 \leq i \leq m = n/b$. For batch r, we compute $\hat{F}_{\mathrm{CMC},r,m}(y) = (1/m)\sum_{i=1}^{m} q(X_{(r-1)m+i}, y)$ as the CDF estimator, and $\hat{\xi}_{\mathrm{CMC},r,m} = \hat{F}^{-1}_{\mathrm{CMC},r,m}(p)$ is the corresponding p-quantile estimator. Then (8) and (9) with $\hat{\xi}_{r,m} = \hat{\xi}_{\mathrm{CMC},r,m}$ and $\tilde{\xi}_{b,m} = \hat{\xi}_{\mathrm{CMC},n}$ give the CMC batching and sectioning CIs, respectively, for ξ.

6 COMBINING CMC AND LHS

We now combine CMC with LHS to estimate ξ. To do this, (20) implies we can obtain an unbiased estimator of $F(y)$ by averaging identically distributed but dependent observations of $q(X,y)$. To generate n (dependent) observations of X using LHS, we apply the function c_X from (27) to the first d_X columns of the LHS grid of $V_{i,j}$ in (13) to get the LHS sample $X'_1, X'_2, \ldots, X'_n$, with

$$\begin{array}{rcccccc} X'_1 & = & c_X(V_{1,1}, & V_{1,2}, & \ldots, & V_{1,d_X}), \\ X'_2 & = & c_X(V_{2,1}, & V_{2,2}, & \ldots, & V_{2,d_X}), \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ X'_n & = & c_X(V_{n,1}, & V_{n,2}, & \ldots, & V_{n,d_X}). \end{array} \tag{29}$$

As the first d_X values of $V_{i,j}$ in each row i of (13) are IID $U[0,1]$, we have that $X'_i \stackrel{\mathscr{D}}{=} X$ by (27). But as before, the rows of $V_{i,j}$ in (13) are dependent, making X'_i, $1 \leq i \leq n$, in (29) dependent.

The CMC+LHS estimators of the CDF F of Y at y and the p-quantile $\xi = F^{-1}(p)$ are respectively

$$\hat{F}_{\mathrm{CMC+LHS},n}(y) = \frac{1}{n}\sum_{i=1}^{n} q(X'_i, y) \quad \text{and} \quad \hat{\xi}_{\mathrm{CMC+LHS},n} = \hat{F}^{-1}_{\mathrm{CMC+LHS},n}(p). \tag{30}$$

The numerator $\psi^2_{\mathrm{CMC+LHS}}$ of the asymptotic variance $\tau^2_{\mathrm{CMC+LHS}}$ in (6) of the CMC+LHS p-quantile estimator $\hat{\xi}_{\mathrm{CMC+LHS},n}$ is the asymptotic variance in the CLT for $\hat{F}_{\mathrm{CMC+LHS},n}(\xi)$ in (4). Deriving an expression for $\psi^2_{\mathrm{CMC+LHS}}$ is complicated by the fact that $\hat{F}_{\mathrm{CMC+LHS},n}(\xi)$ in (30) averages *dependent* copies of the CMC response $q(X,\xi)$ defined in (21) with $y = \xi$. By (27), we can write the CMC response as

$$q(X,\xi) = q(c_X(U_1, U_2, \ldots, U_{d_X}), \xi) \equiv w'(U_1, U_2, \ldots, U_{d_X}) \tag{31}$$

for IID $U[0,1]$ numbers $U_1, U_2, \ldots, U_{d_X}$, and $E[w'(U_1, U_2, \ldots, U_{d_X})] = E[q(X,\xi)] = F(\xi)$ by (20) with $y = \xi$. For each coordinate $1 \leq j \leq d_X$, let $w'_j(U_j) = E[w'(U_1, U_2, \ldots, U_{d_X}) \,|\, U_j] - F(\xi)$. Applying an ANOVA decomposition as in Owen (1992), we express $w'(U_1, U_2, \ldots, U_{d_X})$ using an additive approximation

$$w'(U_1, U_2, \ldots, U_{d_X}) = F(\xi) + \sum_{j=1}^{d_X} w'_j(U_j) + \varepsilon'(U_1, U_2, \ldots, U_{d_X}), \tag{32}$$

where the residual $\varepsilon'(U_1, U_2, \ldots, U_{d_X})$ is defined so that the equality in (32) holds. Then we have that

$$\psi^2_{\mathrm{CMC+LHS}} = \mathrm{Var}[\varepsilon'(U_1, U_2, \ldots, U_{d_X})] = \psi^2_{\mathrm{CMC}} - \sum_{j=1}^{d_X} \mathrm{Var}[w'_j(U_j)], \tag{33}$$

with $\psi^2_{\mathrm{CMC}} = \mathrm{Var}[q(X,\xi)]$ as in (24). Hence, CMC+LHS removes the variance from the additive part of the CMC response. The following result is proven in Dong and Nakayama (2017a).

Theorem 1 When (1) and (27) hold, we have that $\tau^2_{\text{CMC+LHS}} \leq \tau^2_{\text{CMC}} \leq \tau^2_{\text{SRS}}$ and $\tau^2_{\text{CMC+LHS}} \leq \tau^2_{\text{LHS}} \leq \tau^2_{\text{SRS}}$.

Theorem 1 shows that the asymptotic variance of the CMC+LHS quantile estimator is no larger than those of both the LHS and CMC quantile estimators, which in turn are no greater than the asymptotic variance of the SRS quantile estimator. But the key point about this result becomes clear from comparing (33) to (18). As Section 4 noted, (18) shows that LHS removes the variability of the additive part of the SRS response $I(Y \leq \xi)$, corresponding to $F(\xi) = E[I(Y \leq \xi)]$. Because the SRS response is an indicator function, the additive approximation (17) is typically a poor fit, so LHS does not reduce variance by much compared to SRS. In contrast, (33) implies that CMC+LHS removes the variability of the additive part of the CMC response $q(X,\xi)$, arising from $F(\xi) = E[q(X,\xi)]$. By (21) the CMC response is a conditional expectation, which we can think of as a *smoothing* operation, so the additive approximation in (32) can be much better than that in (17). Because of this, CMC+LHS can substantially reduce variance compared to CMC, which has smaller variance than SRS by (24). The numerical results in Section 7 demonstrate this.

We build the CMC+LHS batching and sectioning CIs for ξ as follows. Create $b \geq 2$ independent LHS samples as in (29) but each with $m = n/b$ rows instead of n, and treat each LHS sample $r = 1,2,\ldots,b$, as a batch. For each batch r, let $\hat{F}_{\text{CMC+LHS},r,m}$ be the estimator of CDF F, computed as in (30) but for sample size m rather than n, and compute the corresponding p-quantile estimator $\hat{\xi}_{\text{CMC+LHS},r,m} = \hat{F}^{-1}_{\text{CMC+LHS},r,m}(p)$. The batching CI is then in (8). The sectioning CDF estimator $\tilde{F}_{\text{CMC+LHS},b,m}$ combines the outputs (29) from all b batches, i.e., $\tilde{F}_{\text{CMC+LHS},b,m}(y) = (1/b)\sum_{r=1}^{b} \hat{F}_{\text{CMC+LHS},r,m}(y)$. We then compute the sectioning p-quantile estimator as $\tilde{\xi}_{\text{CMC+LHS},b,m} = \tilde{F}^{-1}_{\text{CMC+LHS},b,m}(p)$, leading to the sectioning CI in (9).

7 NUMERICAL RESULTS

We ran numerical experiments with (Y,X) as bivariate normal (so the function c_X in (27) is $\Re^1$-valued, i.e., $d'_X = 1$) to study the performance of quantile estimation via SRS (Section 3), LHS (Section 4), CMC (Section 5), and CMC+LHS (Section 6). To evaluate d-dimensional integrals, Monte Carlo methods become attractive alternatives to quadrature as d grows (see Chapter 7 of Owen 2013), so we inflate our problem's dimension by generating (Y,X) as follows. In (27), let $(Y,X) = c_*(U_1,U_2,\ldots,U_d) = (\sum_{j=1}^{d} \Phi^{-1}(U_j), \sum_{j=1}^{d_X} \Phi^{-1}(U_j))$ for $U_1,U_2,\ldots,U_d$ IID $U[0,1]$ and $d_X \leq d$. Thus, Y is the sum of d IID standard normals, and X sums the first d_X of them. The bivariate normal (Y,X) has marginal means 0, marginal variances $\text{Var}[Y] = \sigma_Y^2 = d$ and $\text{Var}[X] = \sigma_X^2 = d_X$, and correlation $\rho = \text{Cov}[Y,X]/(\sigma_Y\sigma_X) = \sqrt{d_X/d}$. In our experiments, we chose $d = 30$ and varied $d_X = 5$, 10, 20, to study the impact of ρ. The experiments estimated the p-quantile $\xi = \sigma_Y\Phi^{-1}(p)$ of Y for $p = 0.5$, 0.8, 0.95. CMC and CMC+LHS use X as the (scalar) conditioning vector, and we can compute the conditional expectation in (21) as $q(x,y) = \Phi([y - \rho(\sigma_Y/\sigma_X)x]/[\sigma_Y(1-\rho^2)^{1/2}])$; e.g., see p. 68 of McNeil, Frey, and Embrechts (2005).

We constructed confidence intervals having nominal confidence level $\gamma = 0.9$ through batching (8) and sectioning (9) with $b = 10$ batches and batch size $m = 160$. Table 1 presents the CI coverage estimated from 10^4 independent experiments. For SRS and LHS, the sectioning CI attains better coverage than batching, especially for $p = 0.95$, which agrees with the discussion in the last paragraph of Section 2.

The variance-reduction factor (VRF) of a method x is the ratio of the sample variance of SRS p-quantile estimates across the 10^4 experiments to that of method x. The VRF of LHS is not very large, and it decreases as p approaches 1 for fixed d_X. Thus, LHS is not too effective by itself and becomes even less so when estimating extreme quantiles. In contrast, the VRF for CMC improves as $p \to 1$ for fixed d_X. To see why, for fixed correlation ρ, a bivariate normal becomes asymptotically independent in the tails (see Example 5.32 of McNeil, Frey, and Embrechts 2005), so CMC integrates out more of the variability as $p \to 1$; see (25). On the other hand, for fixed p, the VRF for CMC decreases as d_X grows, i.e., as $\rho = \sqrt{d_X/d}$ increases. In this case, X and Y become more dependent as $\rho \to 1$, so conditioning on X integrates out less variance.

But combining CMC+LHS can lead to substantially smaller variance than either method by itself. Indeed, the VRF for CMC+LHS often far exceeds the product of the individual VRFs for CMC and LHS, demonstrating a synergy of the two applied together. For example, for $d_X = 5$ and $p = 0.5$, the VRFs for

Table 1: The results compare p-quantile estimation via SRS, LHS, CMC, and CMC+LHS for a bivariate normal using 10^4 independent experiments, each with $b = 10$ batches and batch size $m = 160$. We used batching (columns labeled "batch") and sectioning (columns labeled "section") to construct confidence intervals having nominal confidence level $\gamma = 0.9$. We vary d_X to study the effect of the correlation ρ.

			SRS		LHS		CMC		CMC+LHS	
d_X (ρ)	p		batch	section	batch	section	batch	section	batch	section
5	0.5	coverage	0.891	0.902	0.866	0.902	0.898	0.898	0.896	0.896
(0.41)		VRF	1.0e+00	1.0e+00	2.8e+00	2.7e+00	9.4e+00	9.6e+00	1.7e+03	1.7e+03
5	0.8	coverage	0.886	0.904	0.875	0.904	0.893	0.893	0.902	0.902
(0.41)		VRF	1.0e+00	1.0e+00	2.0e+00	2.0e+00	1.1e+01	1.2e+01	2.4e+02	2.5e+02
5	0.95	coverage	0.826	0.897	0.827	0.907	0.893	0.893	0.904	0.905
(0.41)		VRF	1.0e+00	1.0e+00	1.3e+00	1.3e+00	2.0e+01	2.2e+01	1.3e+02	1.4e+02
10	0.5	coverage	0.891	0.902	0.870	0.900	0.900	0.900	0.899	0.899
(0.58)		VRF	1.0e+00	1.0e+00	2.7e+00	2.7e+00	4.6e+00	4.7e+00	2.3e+02	2.4e+02
10	0.8	coverage	0.880	0.900	0.875	0.901	0.901	0.901	0.902	0.901
(0.58)		VRF	1.0e+00	1.0e+00	2.0e+00	2.0e+00	5.4e+00	5.5e+00	5.2e+01	5.3e+01
10	0.95	coverage	0.831	0.903	0.832	0.904	0.900	0.903	0.903	0.903
(0.58)		VRF	1.0e+00	1.0e+00	1.3e+00	1.3e+00	8.5e+00	9.0e+00	2.7e+01	2.9e+01
20	0.5	coverage	0.894	0.903	0.868	0.906	0.899	0.898	0.904	0.905
(0.82)		VRF	1.0e+00	1.0e+00	2.7e+00	2.7e+00	2.1e+00	2.2e+00	2.4e+01	2.5e+01
20	0.8	coverage	0.884	0.904	0.879	0.905	0.897	0.898	0.902	0.903
(0.82)		VRF	1.0e+00	1.0e+00	2.0e+00	2.0e+00	2.3e+00	2.4e+00	1.0e+01	1.0e+01
20	0.95	coverage	0.827	0.902	0.830	0.901	0.896	0.899	0.901	0.900
(0.82)		VRF	1.0e+00	1.0e+00	1.3e+00	1.3e+00	2.9e+00	3.0e+00	5.4e+00	5.8e+00

LHS and CMC are about 2.7 and 8.8, respectively, but the VRF for CMC+LHS is roughly 1700. The VRFs of CMC+LHS in the other cases are not nearly as dramatic, but they still are impressive. As explained in the penultimate paragraph of Section 6, the CMC response $q(X,\xi) = E[I(Y \le \xi) \,|\, X]$ is a smoother function of the uniform inputs than the SRS response $I(Y \le \xi)$ is, leading to a better additive fit in (32) for CMC than the corresponding SRS approximation in (17). As LHS removes the variability from the additive part of the response, CMC+LHS can work extremely well when the CMC response is nearly additive.

8 SUMMARY AND CONCLUDING REMARKS

We combined conditional Monte Carlo with Latin hypercube sampling to estimate a quantile ξ of a random variable Y, and we explained how to use sectioning to construct an asymptotic confidence interval for ξ. The application of LHS requires that Y has the form in (1), so an observation of Y can be generated from a fixed number of independent uniform random numbers. This precludes models for which sampling Y needs an unbounded number of uniforms, e.g., if Y is a functional of a compound Poisson process, even over a finite time horizon. As noted in Section 5, applying CMC critically relies on being able to compute (analytically or numerically) the conditional expectation $q(x,y)$ in (21). This often necessitates that the stochastic model has some level of analytic tractability, which places practical restrictions on when CMC can be employed. We provided a simple example when this can be done in (26).

Our numerical experiments in Section 7 demonstrate that the combination of the CMC and LHS can synergistically reduce variance compared to each method by itself. Section 6 also provides insight into why the combination can be so effective for quantile estimation.

ACKNOWLEDGMENTS

This work has been supported in part by the National Science Foundation under Grant No. CMMI-1537322. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Alban, A., H. Darji, A. Imamura, and M. K. Nakayama. 2017. "Efficient Monte Carlo methods for estimating failure probabilities". *Reliability Engineering and System Safety* 165:376–394.

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

Avramidis, A. N., and J. R. Wilson. 1996. "Integrated Variance Reduction Strategies for Simulation". *Operations Research* 44:327–346.

Avramidis, A. N., and 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.

Basel Committee on Banking Supervision 2004. "Basel II: International Convergence of Capital Measurement and Capital Standards: a Revised Framework". Technical report, Bank for International Settlements, Basel, Switzerland.

Billingsley, P. 1995. *Probability and Measure*. Third ed. New York: John Wiley and Sons.

Bloch, D. A., and J. L. Gastwirth. 1968. "On a Simple Estimate of the Reciprocal of the Density Function". *Annals of Mathematical Statistics* 39:1083–1085.

Chu, F., and M. K. Nakayama. 2012. "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., and M. K. Nakayama. 2014. "Constructing Confidence Intervals for a Quantile Using Batching and Sectioning When Applying Latin Hypercube Sampling". In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk, S. D. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, 640–651: Institute of Electrical and Electronics Engineers.

Dong, H., and M. K. Nakayama. 2017a. "Efficient Monte Carlo Estimation of Quantiles Using Conditional Monte Carlo and Latin Hypercube Sampling". In preparation.

Dong, H., and M. K. Nakayama. 2017b. "Quantile Estimation With Latin Hypercube Sampling". *Operations Research*. to appear.

Ghosh, J. K. 1971. "A New Proof of the Bahadur Representation of Quantiles and an Application". *Annals of Mathematical Statistics* 42:1957–1961.

Hess, S., J. Gaertner, J. Gabor, L. Shanley, L. Enos-Sylla, S. Prasad, and S. Hollingsworth. 2009. "Framework for Risk-Informed Safety Margin Characterization". Technical Report 1019206, Electric Power Research Institute, Palo Alto, California.

Hong, L. J., Z. Hu, and G. Liu. 2014. "Monte Carlo Methods for Value-at-Risk and Conditional Value-at-Risk: A Review". *ACM Transactions on Modeling and Computer Simulataion* 24:Article 22 (37 pages).

Iman, R. L. 1981. "Statistical Methods for Including Uncertainties Associated With the Geologic Isolation of Radioactive Waste Which Allow for a Comparison With Licensing Criteria.". In *Proceedings of the Symposium on Uncertainties Associated with the Regulation of the Geologic Disposal of High-Level Radioactive Waste*, edited by D. C. Kocher, 145–157. Washington, DC: US Nuclear Regulatory Commission, Directorate of Technical Information and Document Control.

McKay, M. D., R. J. Beckman, and W. J. Conover. 1979. "A Comparison of Three Methods for Selecting Input Variables in the Analysis of Output from a Computer Code". *Technometrics* 21:239–245.

McNeil, A. J., R. Frey, and P. Embrechts. 2005. *Quantitative Risk Management: Concepts, Techniques, Tools*. Princeton, NJ: Princeton University Press.

Melchers, R. E. 1999. *Structural Reliability Analysis and Prediction*. Second ed. Chichester, UK: John Wiley & Sons.

Nakayama, M. K. 2014a. "Confidence Intervals Using Sectioning for Quantiles When Applying Variance-Reduction Techniques". *ACM Transactions on Modeling and Computer Simulation* 24:Article 19.

Nakayama, M. K. 2014b. "Quantile Estimation When Applying Conditional Monte Carlo". In *SIMULTECH 2014 Proceedings*, 280–285.

Nakayama, M. K. 2015. "Estimating a Failure Probability Using a Combination of Variance-Reduction Tecniques". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 621–632: IEEE.

Ortega, J. M., and W. C. Rheinboldt. 1987. *terative Solution of Nonlinear Equations in Several Variables*. SIAM.

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. 2013. *Monte Carlo theory, methods and examples*. Draft. In preparation.

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

Stein, M. 1987. "Large Sample Properties of Simulations Using Latin Hypercube Sampling". *Technometrics* 29:143–151. Correction 32:367.

Sun, L., and 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 2005. "Final Safety Evaluation For WCAP-16009-P, Revision 0, "Realistic Large Break LOCA Evaluation Methodology Using Automated Statistical Treatment Of Uncertainty Method (ASTRUM)" (TAC No. MB9483)". Technical report, U.S. Nuclear Regulatory Commission, Washington, DC.

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 Section 50.46 (10CFR50.46), U.S. Nuclear Regulatory Commission, 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.

AUTHOR BIOGRAPHIES

HUI DONG is currently a research scientist at Amazon.com Corporate LLC, Seattle. She received her Ph.D. in Management from Rutgers University. Her research interests include simulation, optimization and their applications in risk evaluation, supply chain and general operations management. Her papers have been accepted in journals including *Operations Research*, *European Journal of Operational Research*, etc. One of her co-authored papers won the Best Theoretical Paper Award for the 2014 Winter Simulation Conference. She has served as a reviewer for *Annals of Operations Research* and the Winter Simulation Conference. Her work at Amazon includes applying optimization and simulation to complex e-commerce logistics system. She recently extended her research interest into deep learning and general machine learning. Her email address is huidong@amazon.com.

MARVIN K. NAKAYAMA is a computer science professor at the New Jersey Institute of Technology. He received a Ph.D. in operations research from Stanford University and a B.A. in mathematics-computer science from U.C. San Diego. He is a recipient of a CAREER Award from the National Science Foundation, and a paper he co-authored received the Best Theoretical Paper Award for the 2014 Winter Simulation Conference. He served as the simulation area editor for the *INFORMS Journal on Computing* from 2007–2016, and is an associate editor for *ACM Transactions on Modeling and Computer Simulation*. His research interests include simulation, modeling, statistics, risk analysis, and energy. His email address is marvin@njit.edu.