Journal of Simulation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tjsm20

A pooled percentile estimator for parallel simulations

Qiong Zhang , Bo Wang & Wei Xie

To cite this article: Qiong Zhang , Bo Wang & Wei Xie (2020): A pooled percentile estimator for parallel simulations, Journal of Simulation, DOI: 10.1080/17477778.2020.1758597

To link to this article: https://doi.org/10.1080/17477778.2020.1758597

Published online: 21 May 2020.

JOURNAL OF SIMULATION
https://doi.org/10.1080/17477778.2020.1758597

TECHNICAL NOTE

A pooled percentile estimator for parallel simulations

Qiong Zhang [a], Bo Wang [b] and Wei Xie [b]

[a]School of Mathematical and Statistical Sciences, Clemson University, Clemson, SC, USA; [b]Department of Mechanical and Industrial Engineering, Northeastern University, Boston, MA, USA

ABSTRACT
Percentile is an important risk measure quantifying the stochastic system random behaviours. This paper studies a pooled percentile estimator, which is the sample percentile of detailed simulation outputs after directly pooling independent sample paths together. We derive the asymptotic representation of the pooled percentile estimator and further prove its normality. By comparing with the classical percentile estimator used in stochastic simulation, both theoretical and empirical studies demonstrate the advantages of the proposal under the context of parallel simulation.

ARTICLE HISTORY
Received 8 April 2019
Accepted 16 April 2020

KEYWORDS
Percentile estimation; parallel computing; stochastic simulation; system risk measure

1. Introduction

Discrete-event simulation is often used to assess the performance of complex stochastic systems, especially in the situations where the direct analytical solution and physical experiments are infeasible or prohibitive (Banks et al., 2010). Advanced computer architectures have made parallel computing available and popular in many engineering and scientific areas. Nelson (2016) raises new research questions about how to exploit this computing advantage in estimating simulation system performance, especially risk measures, such as percentiles. As mentioned in Nelson (2016), a fundamental challenge is how to efficiently utilise all available parallel computing processors to improve estimation accuracy.

In this paper, we consider the steady-state system performance. A single run of simulation generates a sample path of detailed outputs with a given run-length. For various existing system performance (e.g., mean and risk) estimation approaches, both variance and bias of their estimators can be reduced by increasing the run-length (Nelson, 2016). Since the detailed outputs in a simulation sample path are generated sequentially, it could be challenging to chop a dependent sample path into chunks and run separately in parallel from multiple processors. As a result, the run-length becomes the key bottleneck in improving the computational efficiency with parallel simulation.

For the classical percentile estimation approach, we typically calculate the sample percentile of outputs from each simulation run and then take the average of the percentile estimators from multiple replications. The asymptotic properties of this estimator have been well-studied in the literature (e.g., Sun and Hong (2010)). As noted in Heidelberger and Lewis (1984), accurate estimators of tail percentile measures greatly rely on a sufficiently large run-length, which could be a luxury for complex and fast-evolving stochastic systems. For example, we are interested in the 95% percentile of waiting time in a queueing service system. Since the system is required to adapt fast to the evolving demand, we need to assess the system performance and make decisions under a certain tight time deadline. Notice that the run-length greatly relates to the simulation running time before the deadline. Therefore, the parallel processors can increase the number of replications and utilise the available parallel processor under an urgent deadline. However, the classical percentile estimators may not make efficient use of the detailed output sample paths. This could impact the percentile estimation accuracy, especially when the decision-making time is tight.

In this paper, we study a percentile estimator which is computed by directly pooling the detailed simulation outputs from various replications to obtain a sample percentile estimator, which is named as *the pooled sample percentile estimator*. By pooling the dependent (within each replication) and independent (cross different replications) simulation outputs, the resulted sample percentile is introduced to accurately estimate the percentile. Given tight decision time, this result can leverage the prevalent parallel computing power. The pooled percentile estimator has been investigated under independent and identically distributed observations, such as Asmussen and Glynn (2007) and Nakayama (2014), and recently been used to construct the confidence interval of percentile estimators based on a sequence of dependent observations

CONTACT Wei Xie w.xie@northeastern.edu Department of Mechanical and Industrial Engineering, Northeastern University, 334 SN, 360 Huntington Avenue, Boston, MA 02115, USA

in Alexopoulos et al. (2019). In this paper, we develop the asymptotic results of the pooled percentile estimator based on the framework in Sen (1972). Our asymptotic results show that the proposed estimator has better performance especially under the situations with multiple processors and urgent time deadlines, i.e., the job request is too urgent to produce sufficiently longer sample paths. We highlight our contribution as follows.

- We illustrate how the pooled percentile estimator can be used to improve the accuracy of the classical average percentile estimator under the context of parallel simulation.
- We provide the proof of the asymptotic results of the pooled percentile estimator that is generated from multiple replications of dependent sequences.

2. A pooled percentile estimator for parallel simulations

Our goal is to estimate the percentile of the stationary distribution for steady-state simulation. Let X be a random variable representing a single entry in a detailed simulation output sample path. We denote the marginal cumulative distribution function (CDF) by $F(x) = P(X \leq x)$, and denote the α-level percentile by $\xi_\alpha \equiv F^{-1}(\alpha)$, where $0 < \alpha < 1$.

We consider the detailed outputs generated from the steady-state stochastic simulation model,

$$\{X_{ji}; j = 1, 2, \cdots, R; i = 1, 2, \cdots, L\} \tag{1}$$

with R independent sample paths each with run-length L. Particularly, X_{ji} denotes the i-th element of the j-th sample path output. Notice that different sample paths are independent with each other, but entries within each sample path are element-wise dependent. The pooled outputs in (1) contain $N = L \times R$ entries, which construct an empirical CDF:

$$F_N(x) = \frac{1}{N}\sum_{j=1}^{R}\sum_{i=1}^{L}\mathbb{I}\left(X_{ji} \leq x\right), \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Let $X_{(1)} \leq \cdots \leq X_{(N)}$ denote the order statistics over all the entries in (1). We define the α-level pooled percentile estimator obtained by combining all the entries by

$$\hat{\xi}_\alpha^{(P)} = X_{(N\alpha)}, \tag{3}$$

where a denotes the smallest integer greater than or equal to a.

As noted earlier in Section 1, the classical percentile estimator (see Bekki et al. (2009); Chen and Kelton (2006) for examples) is to obtain the α-level sample percentile from each replication, $\hat{\xi}_{j,\alpha} = X_{j,(L\alpha)}$ with $X_{j,(1)} \leq \cdots \leq X_{j,(L)}$ for $j = 1, \ldots, R$. The final estimator is

$$\hat{\xi}_\alpha^{(A)} = \frac{1}{R}\sum_{j=1}^{R}\hat{\xi}_{j,\alpha}, \tag{4}$$

which is referred to as the *average percentile estimator* in this paper. The asymptotic properties of each individual $\hat{\xi}_{j,\alpha}$ have been studied by a vast collection of papers; see, for example, Sen (1972). Thus, the corresponding asymptotic properties of the average percentile estimator can be directly developed based on the mutual independence among $\hat{\xi}_{j,\alpha}$ for $j = 1, 2, \ldots, R$.

The pooled percentile estimator has been investigated under independent and identically distributed (i.i.d.) observations, such as Asmussen and Glynn (2007), Nakayama (2014), and Alexopoulos et al. (2019), the comparison of the asymptotic properties between the pooled percentile estimator and the average percentile estimator based on i.i.d. observations have been stated. Also, given a single replication, the pooled estimator obtained from dependent sequences is used to adjusting the confidence interval of percentile estimators. However, the comparison of asymptotic properties between the pooled percentile estimator and the average percentile estimator based on multiple replications of dependent sequences has not been formally stated in the literature. To fill this gap, we provide the theoretical comparison of the pooled percentile estimator and the average percentile estimator. Our theory is developed based on Assumptions of 2.1–2.4.

Assumption 2.1. *For each replication* j $(= 1, \cdots, R)$, $\{X_{ji}; -\infty < i < \infty\}$ *is a stationary sequence of* ϕ*-mixing random variables, i.e., for the* $\sigma -$ *fields* $F_{-\infty}^{k}$ *and* F_{k+n}^{∞} *generated by* $\{X_{ji}; i \leq k\}$ *and* $\{X_{ji}; i \geq k + n\}$ *respectively, we have*

$$\left|P(E_2|E_1) - P(E_2)\right| \leq \phi(n), \text{for} - \infty < k < \infty, \text{and} n \geq 1 \tag{5}$$

where $E_1 \in F_{-\infty}^{k}$ *and* $E_2 \in F_{k+n}^{\infty}$, *and* $1 \geq \phi(1) \geq \phi(2) \geq \cdots \geq 0$ *with* $\lim_{n\to\infty}\phi(n) = 0$, *and*

$$\sum_{n=1}^{\infty} e^{tn}\phi(n) < \infty \text{ for some } t > 0. \tag{6}$$

Assumption 2.2. $R = o(L)$ *as* $L \to \infty$.

Assumption 2.3. $F'(x) = f(x)$ *is continuous and positive in the neighbourhood of* ξ_α.

Assumption 2.4. $f'(x) = \frac{d}{dx}f(x)$ *is positive and bounded in the neighbourhood of* ξ_α.

Assumption 2.1 is called the ϕ-mixing condition, which is commonly adopted in the steady-state simulation output analysis (Chen & Kelton, 2000; Steiger &

Wilson, 2001). It states that serial dependency decreases as the lag increases. Bradley (2005) and Bradley (2010) provided the results of theoretically verifying the ϕ-mixing condition for some popular examples of dependent sequences, such as Markov chains, stationary Gaussian processes, and etc. Assumption 2.2 is that we normally consider the run-length far larger than the number of replications, and it also matches the situation in parallel computing where the number of available processors is often less than the run-length for the steady-state simulation. Assumptions 2.3 and 2.4 are the common assumptions for developing asymptotic representations of percentile estimators.

3. The asymptotic representation of the pooled percentile estimator

The asymptotic representation of sample percentile has been investigated under both independent data and dependent sequence data. For the pooled percentile estimator of simulation outputs from independent replications, asymptotic characterisation is still missing in the literature. We aim to fill this gap and develop the asymptotic representation for the percentile estimator with pooled sample paths and provide the theoretical insights on how to deploy multiple processors to improve the estimation of system percentile response. We first provide the asymptotic representation of the pooled percentile estimator through Theorem 3.1, and then Theorem 3.2 gives its asymptotic distribution. The proofs of these Theorems follows a similar logic as in Sen (1972), and extend their results in incorporate multiple replications of dependent sample paths.

Theorem 3.1. *Consider a small neighbourhood around the true α-level percentile ξ_α, denoted by $I_N = \{x : |x - \xi_\alpha| \le N^{-1/2}\log L\}$. Under Assumptions 2.1–2.3, as $L \to \infty$,*

$$\sup_{x\in I_N} |[F_N(x) - F_N(\xi_\alpha)] - [F(x) - F(\xi_\alpha)]| = O(N^{-3/4}\log L) \tag{7}$$

almost surely. Further, under Assumption 2.4, we have,

$$|[\alpha - F_N(\xi_\alpha)] - (\hat{\xi}_\alpha^{(P)} - \xi_\alpha)f(\xi_\alpha)| = O(N^{-3/4}\log L), \tag{8}$$

almost surely.

Proof. Let $\eta_{r,N} = \xi_\alpha + rN^{-3/4}\log L$, where $r = 0, \pm1, \ldots, \pm b_N$, and $b_N = N^{1/4}$. Then for all $x \in J_{r,N} = [\eta_{r,N}, \eta_{r+1,N}]$, we have that

$$\begin{aligned}\sup_{x\in I_N} &|[F_N(x) - F_N(\xi_\alpha)] - [F(x) - F(\xi_\alpha)]| \\ &\le \max_{-b_N\le r\le b_N} |[F_N(\eta_{r,N}) - F_N(\xi_\alpha)] \\ &\quad - [F(\eta_{r,N}) - F(\xi_\alpha)]| \\ &\quad + \max_{-b_N\le r\le b_N-1} |F(\eta_{r+1,N}) - F(\eta_{r,N})|.\end{aligned}$$

Since $\eta_{r+1,N} - \eta_{r,N} = N^{-3/4}\log L$, by the Mean-Value Theorem,

$$\begin{aligned}|F(\eta_{r+1,N}) - F(\eta_{r,N})| &\le |\sup_{x\in J_{r,N}} f(x)|(\eta_{r+1,N} - \eta_{r,N}) \\ &= O(N^{-3/4}\log L),\end{aligned}$$

and $\max_{-b_N\le r\le b_N-1} |F(\eta_{r+1,N}) - F(\eta_{r,N})| = O(N^{-3/4}\log L)$almost surely.

For $r = 1, 2, \ldots, b_N$, let $U_{ji}^{(r)} = \mathbb{I}(X_{ji} \le \eta_{r,N}) - \mathbb{I}(X_{ji} \le \xi_\alpha)$. Notice that $U_{ji}^{(r)}$ is 0–1 valued, and such that,

$$F_N(\eta_{r,N}) - F_N(\xi_\alpha) = \frac{1}{N}\sum_{j=1}^{R}\sum_{i=1}^{L} U_{ji}^{(r)}$$

and $F(\eta_{r,N}) - F(\xi_\alpha) = P(U_{ji}^{(r)} = 1) =: \alpha_N^{(r)}$.

According to the Mean-Value Theorem and the definition of b_N, $K_1N^{-3/4}\log L \le p_N^{(r)} \le K_2N^{-1/2}\log L$. By directly applying Lemma A.2 (see Appendix A), we have that, as $L \to \infty$,

$$P\left\{|[F_N(\eta_{r,N}) - F_N(\xi_\alpha)] - [F(\eta_{r,N}) - F(\xi_\alpha)]| > CN^{-3/4}\log L\right\} \le C_2L^{-2}$$

if let $s = 2$ in Lemma A.2.

For $r = -b_N, \ldots, -1$, let $U_{ji}^{(r)} = \mathbb{I}(X_{ji} \le \xi_\alpha) - \mathbb{I}(X_{ji} \le \eta_{r,N})$, and we could derive the same results. According to the Bonferroni Inequality,

$$\begin{aligned}\mathrm{P}\{&\max_{-b_N\le r\le b_N} |[F_N(\eta_{r,N}) - F_N(\xi_\alpha)] - [F(\eta_{r,N}) - F(\xi_\alpha)]| \\ &> CN^{-3/4}\log L\} \le C_2 \cdot 2b_N \cdot L^{-2} = O(L^{-3/2}).\end{aligned}$$

Then, by Borel–Cantelli Lemma Serfling (2009),

$$\max_{-b_N\le r\le b_N} |[F_N(\eta_{r,N}) - F_N(\xi_\alpha)] - [F(\eta_{r,N}) - F(\xi_\alpha)]| = O(N^{-3/4}\log L)$$

almost surely. Then (7) holds.

We now prove (8). Let $k = N\alpha$,

$$\begin{aligned}&\mathrm{P}\left(\hat{\xi}_\alpha^{(P)} < \xi_\alpha - N^{-1/2}\log L\right) \\ &= \mathrm{P}\left\{\sum_{j=1}^{R}\sum_{i=1}^{L}\mathbb{I}\left(X_{ji} \le \xi_\alpha - N^{-1/2}\log L\right) \ge k\right\} \\ &= \mathrm{P}\left\{\frac{1}{N}\sum_{j=1}^{R}\sum_{i=1}^{L} W_{ji} - F\left(\xi_\alpha - N^{-1/2}\log L\right) \ge \frac{k}{N} - F\left(\xi_\alpha - N^{-1/2}\log L\right)\right\}\end{aligned}$$

where $W_{ji} = \mathbb{I}\left(X_{ji} \le \xi_\alpha - N^{-1/2}\log L\right)$, and $\mathrm{P}\{W_{ji} = 1\} = F\left(\xi_\alpha - N^{-1/2}\log L\right)$. Since as $L \to \infty$,

$$\frac{k}{N} - F\left(\xi_\alpha - N^{-1/2}\log L\right) = f(\xi_\alpha)N^{-1/2}\log L[1+o(1)].$$

From Lemma A.1 (see Appendix A), as $L \to \infty$,

$$\frac{1}{N}\sum_{j=1}^{R}\sum_{i=1}^{L} W_{ji} - \mathrm{P}\{W_{ji}=1\} \le \left(\frac{2}{t}\right)N^{-1/2}\log L,$$

almost surely. Thus, we have that

$$\hat{\xi}_\alpha^{(P)} \ge \xi_\alpha - N^{-1/2}\log L \tag{9}$$

almost surely. Similarly,

$$\begin{aligned}&\mathrm{P}\left(\hat{\xi}_\alpha^{(P)} > \xi_\alpha + N^{-1/2}\log L\right)\\ &= \mathrm{P}\left\{\frac{1}{N}\sum_{j=1}^{R}\sum_{i=1}^{L}\mathbb{I}\left(X_{ji} \le \xi_\alpha + N^{-1/2}\log L\right) < \frac{k}{N}\right\}.\end{aligned}$$

By the monotonicity of $F(\cdot)$, as $L \to \infty$,

$$\begin{aligned}&\frac{1}{N}\sum_{j=1}^{R}\sum_{i=1}^{L}\mathbb{I}\left(X_{ji} \le \xi_\alpha + N^{-1/2}\log L\right)\\ &\quad\to F(\xi_\alpha + N^{-1/2}\log L),\end{aligned}$$

and $\frac{k}{N} \to F(\xi_\alpha)$. Thus, we have that,

$$\hat{\xi}_\alpha^{(P)} \le \xi_\alpha + N^{-1/2}\log L \tag{10}$$

as $L \to \infty$ almost surely. Therefore, under Assumption 2.4, the conclusion holds by setting $x = \hat{\xi}_\alpha^{(P)}$ in (7).

Notice that Equation (8) can be rewritten as,

$$\hat{\xi}_\alpha^{(P)} - \xi_\alpha = \frac{\alpha - F_N(\xi_\alpha)}{f(\xi_\alpha)} + O(N^{-3/4}\log L), \tag{11}$$

which gives the Bahadur representation of sample percentile of the pooled sample paths. Now we consider the asymptotic distribution of the estimator $\hat{\xi}_\alpha^{(P)}$. Let $F_j(x) = \frac{1}{L}\sum_{i=1}^{L}\mathbb{I}(X_{ji} \le x)$ for $j = 1, \ldots, R$, and $F_N(x) = \frac{1}{R}\sum_{j=1}^{R}F_j(x)$. Following the general definition in literature (e.g., Sen (1972)), we denote

$$\nu^2 = \nu_0 + 2\sum_{h=1}^{\infty}\nu_h, \tag{12}$$

where $\nu_h = \mathrm{E}\left[\mathbb{I}(X_{j,1} \le \xi_\alpha)\mathbb{I}(X_{j,1+h} \le \xi_\alpha)\right] - \alpha^2$, which is the same for all replications with $j = 1, 2, \ldots, R$. Under the setting of pooled sample paths in this paper, we obtain that

$$\begin{aligned}\lim_{L\to\infty}\{N\cdot\mathrm{Var}[F_N(\xi_\alpha)]\} &= \lim_{L\to\infty}\left\{L\cdot\mathrm{Var}[F_j(\xi_\alpha)]\right\}\\ &= \nu^2.\end{aligned}$$

Theorem 3.2. *Under Assumptions 2.1–2.3, and* $\sigma^2 := \nu^2/[f(\xi_\alpha)]^2$, $0 < \sigma^2 < \infty$,

$$\frac{N^{1/2}(\hat{\xi}_\alpha^{(P)} - \xi_\alpha)}{\sigma} \xrightarrow{d} \mathcal{N}(0,1). \tag{13}$$

Proof. According to Theorem 3.1, we have

$$N^{1/2}[F_N(\hat{\xi}_\alpha^{(P)}) - F(\hat{\xi}_\alpha^{(P)})] \xrightarrow{p} N^{1/2}[F_N(\xi_\alpha) - \alpha] \tag{14}$$

as $L \to \infty$. By the central limit theorem for $\phi-$mixing variables (see Sen (1972) for example), $L^{1/2}[F_j(\xi_\alpha) - \alpha]/\nu \xrightarrow{d} \mathcal{N}(0,1)$ and by the independence between replications,

$$\frac{N^{1/2}[F_N(\xi_\alpha) - \alpha]}{\nu} \xrightarrow{d} \mathcal{N}(0,1). \tag{15}$$

On the other hand, $F_N(\hat{\xi}_\alpha^{(P)}) = k/N = \alpha + O(N^{-1}) = F(\xi_\alpha) + O(N^{-1})$. As $L \to \infty$,

$$\begin{aligned}&N^{1/2}[F_N(\hat{\xi}_\alpha^{(P)}) - F(\hat{\xi}_\alpha^{(P)})]\\ &= N^{1/2}[F(\xi_\alpha) - F(\hat{\xi}_\alpha^{(P)})] + O(N^{-1/2})\\ &= \left[N^{1/2}(\xi_\alpha - \hat{\xi}_\alpha^{(P)})f(\xi_\alpha)\right]\frac{f(\theta\hat{\xi}_\alpha^{(P)} + (1-\theta)\xi_\alpha)}{f(\xi_\alpha)}\\ &\quad + O(N^{-1/2})\end{aligned} \tag{16}$$

where $\theta \in [0,1]$. Since $f(x)$ is continuous in some neighbourhood of ξ_α, $0 < f(\xi_\alpha) < \infty$, and from Theorem 3.1, $|\xi_\alpha - \hat{\xi}_\alpha^{(P)}| \le N^{-1/2}\log L$, then as $L \to \infty$, $f(\theta\hat{\xi}_\alpha^{(P)} + (1-\theta)\xi_\alpha)/f(\xi_\alpha) \xrightarrow{p} 1$. By applying (14), (15) and (16), and the Slutsky's Theorem, (13) holds. □

According to (11) and (13), for the sample α-percentile $\hat{\xi}_\alpha^{(P)}$ given by (3), we have the following asymptotic bias and variance:

$$\begin{aligned}&\mathrm{Bias}(\hat{\xi}_\alpha^{(P)}) = O(N^{-3/4}\log L), \text{and}\,\mathrm{Var}(\hat{\xi}_\alpha^{(P)})\\ &\quad = \frac{\nu^2}{N[f(\xi_\alpha)]^2}.\end{aligned} \tag{17}$$

On the other hand, for the classical sample percentile $\hat{\xi}_\alpha^{(A)}$ given by (4), we can directly apply the results from Sen (1972) to obtain

$$\begin{aligned}\mathrm{Bias}(\hat{\xi}_\alpha^{(A)}) &= O(L^{-3/4}\log L)\text{and}\,\mathrm{Var}(\hat{\xi}_\alpha^{(A)})\\ &= \frac{\nu^2}{N[f(\xi_\alpha)]^2},\end{aligned} \tag{18}$$

by the mutual independence among different replications. *Asymptotically,* $\hat{\xi}_\alpha^{(P)}$ *achieves the same variance as* $\hat{\xi}_\alpha^{(A)}$*, while the bias of* $\hat{\xi}_\alpha^{(P)}$ *is in a smaller order than the bias of* $\hat{\xi}_\alpha^{(A)}$. We see that the bias of $\hat{\xi}_\alpha^{(A)}$ decreases as we increase the run-length, whereas stays the same as we increase the number of replications. Different from the classical percentile estimator, the bias of $\hat{\xi}_\alpha^{(P)}$ decreases as we increase either the run-length or the number of replications. Comparing these two estimators using the mean squared error (MSE), *the pooled percentile estimator can achieve smaller MSE than the classical average percentile estimator.*

As discussed earlier, multiple replications could be assigned to parallel processors. Thus, given a tight

decision time, different replications can be allocated to parallel processors to improve the percentile estimation under a tight time constraint. With that said, the fixed decision time implies that the run-length is fixed to be L, while the number of replications R can be increased depending on how many parallel processors are available. Under this situation, the bias of $\hat{\xi}_\alpha^{(A)}$ stays at the same order no matter how many processors have been adopted to enlarge the number of replications. Different from $\hat{\xi}_\alpha^{(A)}$, the bias of $\hat{\xi}_\alpha^{(P)}$ decreases as we increase the number of parallel processors. In Section 4, we use an empirical example to demonstrate that the pooled percentile estimator outperforms the classical average percentile estimator when the run-length is insufficient due to an urgent deadline.

4. Numerical study

In this section, we provide an empirical example to illustrate the performance of the proposed approach under the parallel computing setting. We use MSE to demonstrate the performance of different estimators. In the following examples, the MSE is computed with 100 micro-replications:

$$\widehat{\mathrm{MSE}}(\hat{\xi}_\alpha) = \frac{1}{100}\sum_{m=1}^{100}\left(\hat{\xi}_\alpha^{(m)} - \xi_\alpha\right)^2$$

where $\hat{\xi}_\alpha^{(m)}$ is the estimator for α-level percentile from the pooled or the classical average method at the m-th micro-replication. Two examples AR(1) and M/M/1 queue are used to generate the dependent sequences. We consider that R parallel processors are available to use, and the simulation running in each processor generates one replication of the dependent sequence. Two situations are considered: (1) there is an urgent deadline to provide a percentile estimator, so we only have time to generate dependent sequences with run-length $L = 1000$; and (2) the deadline to provide a percentile estimator is not urgent, and we are able to generate dependent sequences with sufficient run-length $L = 10000$.

Example 1: AR(1) We consider an AR(1) process. For one replication of the dependent sequence, the outputs are given by

$$X_i = \mu + \phi X_{i-1} + \varepsilon_i,$$

where ε_i is a white noise process with zero mean and variance σ^2. We fix the mean $\mu = 0$ and variance $\sigma^2 = 1$. The correlation parameter ϕ is varying from 0.3, 0.5, to 0.9. We estimate the percentiles ξ_α of the outputs with $\alpha = 50\%$, 95%, and compare the performance of the pooled estimator in (3) with the classical average percentile estimator in (4). The results of

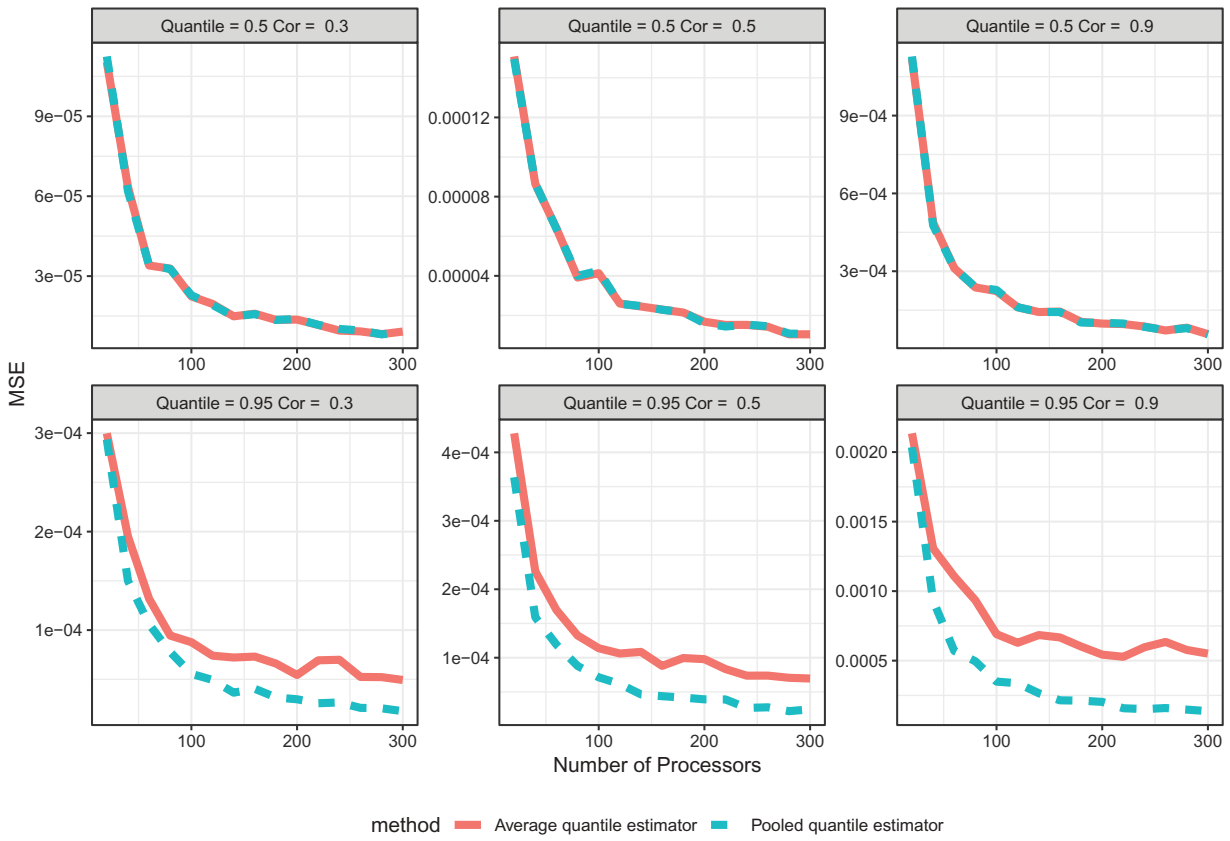

Figure 1. Percentile estimates under an urgent deadline ($L = 1000$) of the AR(1) example. This figure demonstrates the situation that the experimenter encounters an urgent deadline to provide a percentile estimator, and the time constraint does not allow the run-length on each processor to be sufficient.

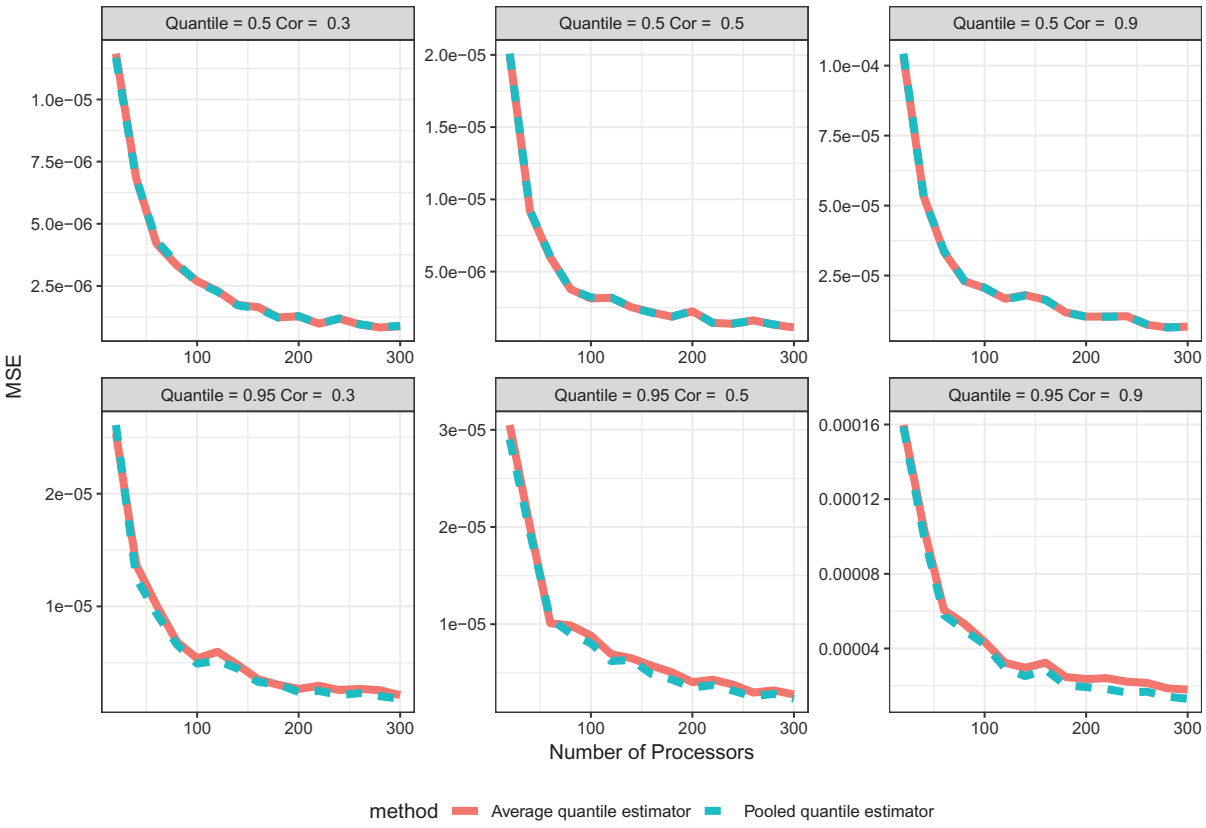

Figure 2. Percentile estimates under non-urgent deadline ($L = 10000$) of the AR(1) example. This figure demonstrates that there is no time constraint to generate the percentile estimator, and the run-length can be sufficient to guarantee the accuracy.

MSEs under urgent deadline and non-urgent deadline are given in Figures 1 and 2, respectively.

Example 2: M/M/1 Queue We consider the steady-state $M/M/1$ Queueing system. We fix the arrival rate to be 1, and vary the utilisation (traffic intensity) to be 0.7 or 0.9. We estimate the percentiles ξ_α of time staying in the system with $\alpha = 50\%$, 95%, and compare the performance of the pooled estimator in (3) with the classical average percentile estimator in (4). The results of MSEs under urgent deadline and non-urgent deadline are given in Figures 3 and 4, respectively.

The results shown in Figure 3 represent the performances of the percentile estimators under an urgent deadline (i.e., run-length $L = 1000$), whereas the results shown in Figure 4 represent the performances of the percentile estimators under a non-urgent deadline (i.e., run-length $L = 10000$). The y-axis represents the estimated MSE and the x-axis is the number of processors R, and different scenarios are labelled on top of each sub-figure. For the cases representing an urgent deadline, the pooled percentile estimator gives smaller or competitive MSEs compared to the average percentile estimator. This demonstrates the benefits of using the pooled percentile estimator when there is not sufficient time to generate a lengthy sequence. For the cases representing a non-urgent deadline, we have sufficient time to generate dependent sequences with a relatively large run-length on each processor. As also demonstrated in the theoretical comparison, the MSEs from the pooled estimator significantly outperform the average percentile estimator if the run-length is sufficient.

5. Discussion

As a summary, we study the pooled percentile estimator, which takes the sample percentile by pooling independently generated sample paths together. We develop the asymptotic representation of the proposed percentile estimator generated from multiple replications of dependent sequences. Compared with the classical average percentile estimator, the pooled percentile estimator demonstrates better asymptotic and finite-sample performance under the context of parallel simulations. The pooled percentile estimator can be advanced by combining various existing variance reduction techniques in the literature. Hence, as a promising future direction, the accuracy of the pooled percentile estimator can be further improved by incorporating existing techniques, such as control variates (i.e., Hesterberg and Nelson (1998); Hsu and Nelson (1990)), importance sampling and stratified sampling (i.e., Glasserman et al. (2000), Glynn

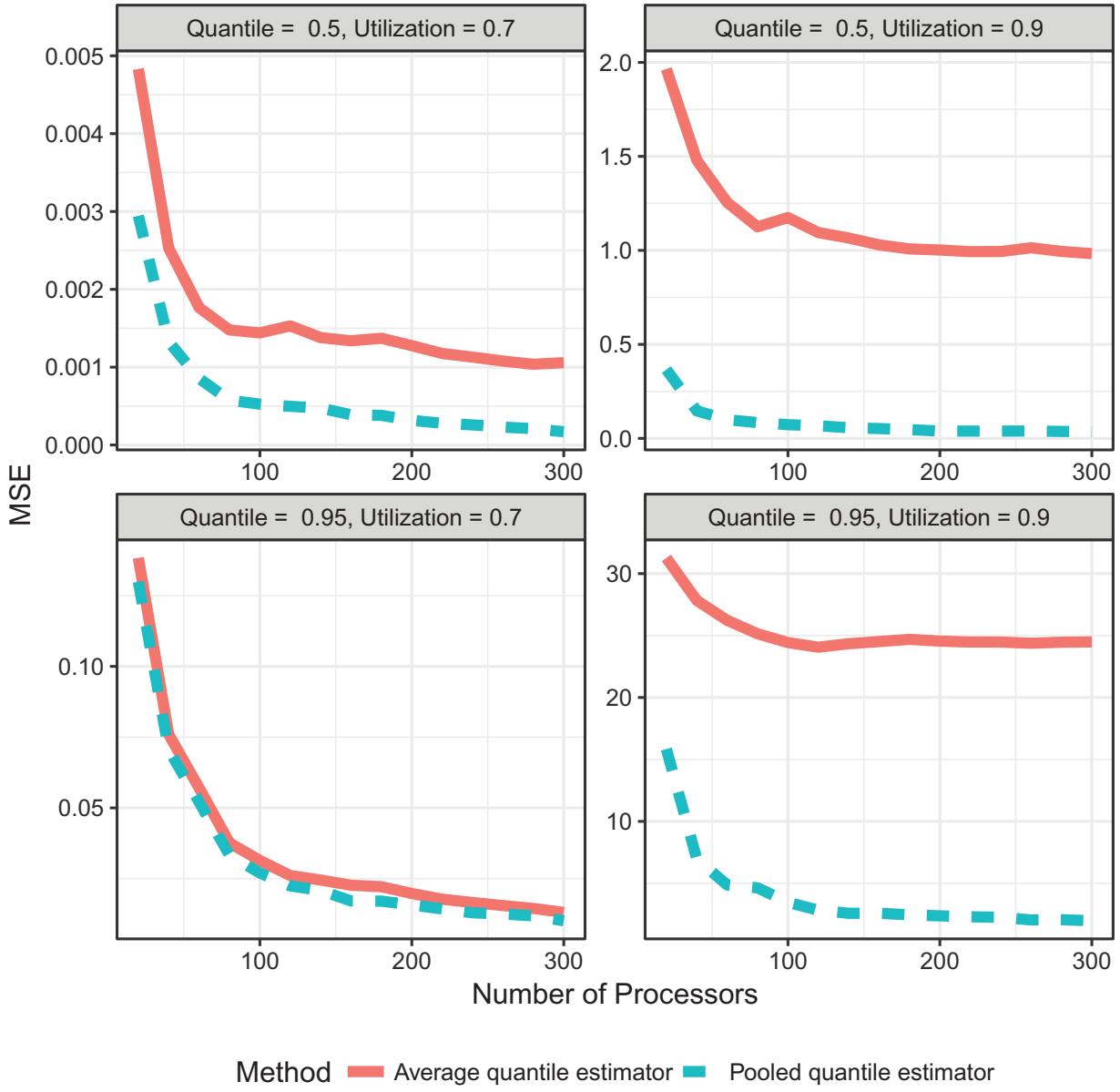

Figure 3. Percentile estimates under an urgent deadline ($L = 1000$) of the M/M/1 example. This figure demonstrates the situation that the experimenter encounters an urgent deadline to provide a percentile estimator, and the time constraint does not allow the run-length on each processor to be sufficient.

(1996), and Sun and Hong (2010)) antithetic variates and Latin hypercube sampling (i.e., Avramidis and Wilson (1998); Jin et al. (2003)), as well as bias correction techniques (i.e., Gomes and Figueiredo (2006), Gomes and Pestana (2007), and Matthys et al. (2004)).

Although the results of theoretically verification of the ϕ-mixing condition for some popular examples of dependent sequences can be found in Bradley (2005), it is known that the ϕ-mixing condition can be difficult to check in general as mentioned in Alexopoulos et al. (2019). It is a promising future direction to extend the asymptotic properties in this paper based on the milder conditions in Wu (2005) by following the development of Alexopoulos et al. (2019).

The trade-off between one single replication with multiple replications is an important issue to address for future study. The effort spending on the warm-up simulation procedure to achieve steady-state should also be considered. For the case with multiple replications, the warm-up procedure to generate steady-state outputs may not be negligible. A preliminary idea is that, we can formulate the accuracy of an estimator as an objective function of R and L, and the computational cost can be specified as $N = (R + W)L$, where W is the warm-up runs, which is assumed to be a fixed number. Therefore, the optimal choice of run length R and replication L could be a solution of maximising the accuracy under the constraint that $N = (R + W)L$.

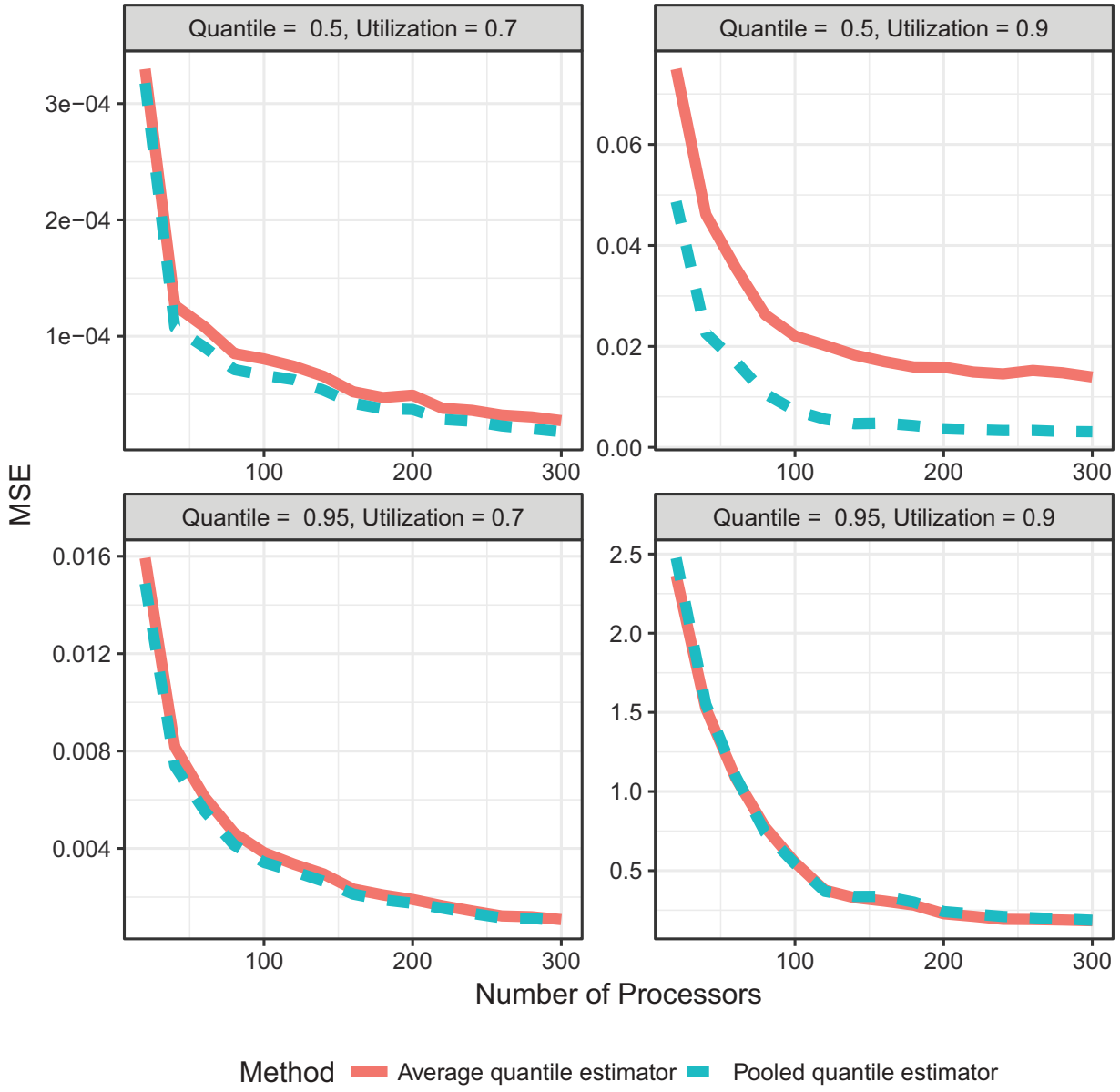

Figure 4. Percentile estimates under non-urgent deadline ($L = 10000$) of the M/M/1 example. This figure demonstrates that there is no time constraint to generate the percentile estimator, and the run-length can be sufficient to guarantee the accuracy.

Along this line, it is critically important to further investigate how to balance the trade-off between one single replication with multiple replications.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Qiong Zhang http://orcid.org/0000-0003-1995-2127

References

Alexopoulos, C., Goldsman, D., Mokashi, A. C., Tien, K.-W., & Wilson, J. R. (2019). Sequest: A sequential procedure for estimating quantiles in steady-state simulations. *Operations Research, 67*(4), 1162–1183. https://doi.org/10.1287/opre.2018.1829

Asmussen, S., & Glynn, P. W. (2007). *Stochastic simulation: Algorithms and analysis* (Vol. 57). Springer Science & Business Media.

Avramidis, A. N., & Wilson, J. R. (1998). Correlation-induction techniques for estimating quantiles in simulation experiments. *Operations Research*, *46*(4), 574–591. https://doi.org/10.1287/opre.46.4.574

Banks, J., Carson, J., & Nelson, B. (2010). *Discrete-event system simulation*. Prentice Hall.

Bekki, J. M., Fowler, J. W., Mackulak, G. T., & Nelson, B. L. (2009). Indirect cycle time quantile estimation using the cornish–fisher expansion. *IIE Transactions*, *42*(1), 31–44. https://doi.org/10.1080/07408170903019135

Bradley, R. C. (2005). Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, *2*, 107–144. https://doi.org/10.1214/154957805100000104

Bradley, R. C. (2007). *Introduction to strong mixing conditions*. Kendrick Press.

Chen, E. J., & Kelton, W. D. (2000). Batching methods for simulation output analysis: A stopping procedure based on phi-mixing conditions. In *Proceedings of the 32nd conference on winter simulation* (pp. 617–626), Orlando, FL.

Chen, E. J., & Kelton, W. D. (2006). Quantile and tolerance-interval estimation in simulation. *European Journal of Operational Research*, *168*(2), 520–540. https://doi.org/10.1016/j.ejor.2004.04.040

Glasserman, P., Heidelberger, P., & Shahabuddin, P. (2000). Variance reduction techniques for estimating value-at-risk. *Management Science*, *46*(10), 1349–1364. https://doi.org/10.1287/mnsc.46.10.1349.12274

Glynn, P. W. (1996). Importance sampling for monte carlo estimation of quantiles. In *Mathematical methods in stochastic simulation and experimental design: Proceedings of the 2nd st. petersburg workshop on simulation* (pp. 180–185).

Gomes, M. I., & Figueiredo, F. (2006). Bias reduction in risk modelling: Semi-parametric quantile estimation. *Test*, *15* (2), 375–396. https://doi.org/10.1007/BF02607058

Gomes, M. I., & Pestana, D. (2007). A sturdy reduced-bias extreme quantile (var) estimator. *Journal of the American Statistical Association*, *102*(477), 280–292. https://doi.org/10.1198/016214506000000799

Heidelberger, P., & Lewis, P. A. (1984). Quantile estimation in dependent sequences. *Operations Research*, *32*(1), 185–209. https://doi.org/10.1287/opre.32.1.185

Hesterberg, T. C., & Nelson, B. L. (1998). Control variates for probability and quantile estimation. *Management Science*, *44* (9), 1295–1312. https://doi.org/10.1287/mnsc.44.9.1295

Hsu, J. C., & Nelson, B. L. (1990). Control variates for quantile estimation. *Management Science*, *36*(7), 835–851. https://doi.org/10.1287/mnsc.36.7.835

Jin, X., Fu, M. C., & Xiong, X. (2003). Probabilistic error bounds for simulation quantile estimators. *Management Science*, *49* (2), 230–246. https://doi.org/10.1287/mnsc.49.2.230.12743

Matthys, G., Delafosse, E., Guillou, A., & Beirlant, J. (2004). Estimating catastrophic quantile levels for heavy-tailed distributions. *Insurance: Mathematics and Economics*, *34* (3), 517–537. https://doi.org/10.1016/j.insmatheco.2004.03.004

Nakayama, M. K. (2014). Confidence intervals for quantiles using sectioning when applying variance-reduction techniques. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, *24*(4), 19. https://doi.org/10.1145/2558328

Nelson, B. L. (2016). 'some tactical problems in digital simulation' for the next 10 years. *Journal of Simulation*, *10*(1), 2–11. https://doi.org/10.1057/jos.2015.22

Sen, P. K. (1972). On the bahadur representation of sample quantiles for sequences of ϕ-mixing random variables. *Journal of Multivariate Analysis*, *2*(1), 77–95. https://doi.org/10.1016/0047-259X(72)90011-5

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics* (Vol. 162). John Wiley & Sons.

Steiger, N. M., & Wilson, J. R. (2001). Convergence properties of the batch means method for simulation output analysis. *INFORMS Journal on Computing*, *13*(4), 277–293. https://doi.org/10.1287/ijoc.13.4.277.9737

Sun, L., & Hong, L. J. (2010). Asymptotic representations for importance-sampling estimators of value-at-risk and conditional value-at-risk. *Operations Research Letters*, *38* (4), 246–251. https://doi.org/10.1016/j.orl.2010.02.007

Wu, W. B. (2005). On the bahadur representation of sample quantiles for dependent sequences. *The Annals of Statistics*, *33*(4), 1934–1963. https://doi.org/10.1214/009053605000000291

Appendix A. Some Useful Lemmas

We first consider that $\{Y_{ji} : j = 1, 2, \ldots, R; i = 1, 2, \ldots, L\}$ are series of 0–1 valued random variables which satisfies the same mixing condition as $\{X_{ji}\}$ given through (5) and (6), and sharing the same marginal distribution $P(Y_{ji} = 1) = 1 - P(Y_{ji} = 0) = \alpha$. Assume that $S_j = \sum_{i=1}^{L} Y_{ji}$ and $S_N = \sum_{j=1}^{R} S_j$, then we can have the following lemma by extending the results in Section 4 of Sen (1972).

Lemma A.1. *For a positive t ($t < 3$) that the ϕ − mixing condition holds, as $L \to \infty$ and $R = o(L)$,*

$$S_N \le N\alpha + \left(\frac{2}{t}\right) N^{1/2} \log L, \ \ w.p.1 \tag{A1}$$

Proof. By Markov inequality, we would have,

$$\mathrm{P}\left(S_N > N\alpha + \left(\frac{2}{t}\right) N^{1/2} \log L\right) \le \inf_{h>0} \left\{ \exp\left[-hN\alpha - \left(\frac{2}{t}\right) hN^{1/2} \log L \right] \mathrm{E}\left[\exp(hS_N)\right] \right\}$$

And we can rewrite $S_j = S_j^{(1)} + S_j^{(2)} + \cdots + S_j^{(k_L)}$ with $S_j^{(\ell)} = Y_{j\ell} + Y_{j,\ell+n_L} + \cdots + Y_{j,\ell+m_L^{(\ell)} n_L}$, and choose integer $n_L = \left(\frac{2}{t}\right) \log L$, $1 \le \ell \le n_L$, and $m_L^{(\ell)}$ be the largest positive integer s.t. $\ell + m_L^{(\ell)} n_L \le L$, notice that $m_L^{(\ell)} \le m_L^{(1)} \le L/n_L - 1$. Then from the independence between replications and inequality between arithmetic and geometric means, we have,

$$\mathrm{E}\left[\exp(hS_N)\right] = \prod_{j=1}^{R} \mathrm{E}\left[\exp(hS_j)\right] = \prod_{j=1}^{R} \mathrm{E}\left[\prod_{\ell=1}^{n_L} \exp(hS_j^{(\ell)}) \right]$$

$$\le \prod_{j=1}^{R} \mathrm{E}\left[\left(\frac{\sum_{\ell=1}^{n_L} \exp(hS_j^{(\ell)})}{n_L} \right)^{n_L} \right] \le \left\{ \mathrm{E}\left[\exp(hn_L S_1^{(1)})\right] \right\}^R$$

According to the ϕ − mixing condition (5), for every i, $P(Y_{j,i+n_L} = 1 | F_{-\infty}^{i}) \le \alpha + \phi(n_L)$ and from condition (6) and the choice of n_L, we have,

$$\phi(n_L) = o(e^{-tn_L}) = o[\exp(-2\log L)] = o(L^{-2}), \ \ as \ L \to \infty.$$

Then, for $m = 1, \cdots, m_L^{(1)}$,

$$\begin{aligned} &\mathrm{E}\left[\exp(hn_L Y_{j,1+mn_L}) \middle| F_{-\infty}^{1+(m-1)n_L}\right] = 1 + \mathrm{P}\left(Y_{j,1+mn_L} = 1 \middle| F_{-\infty}^{1+(m-1)n_L}\right)[\exp(hn_L) - 1] \\ &\le 1 + [\alpha + o(L^{-2})][\exp(hn_L) - 1], \end{aligned} \tag{A2}$$

and $\mathrm{E}[\exp(hn_L Y_{j1})] = 1 + \alpha[\exp(hn_L) - 1]$. Applying those recursively yields:

$$\begin{aligned} &\mathrm{E}\left[\exp(hn_L S_1^{(1)})\right] = \mathrm{E}\left\{ \mathrm{E}\left[\exp(hn_L \sum_{m=0}^{m_L^{(1)}} Y_{1,1+mn_L}) \middle| F_{-\infty}^{1+(m-1)n_L} \right] \right\} \le \mathrm{E}\left[\exp(hn_L \sum_{m=0}^{m_L^{(1)}-1} Y_{1,1+mn_L}) \right] \{1 + [\alpha + o(L^{-2})][\exp(hn_L) - 1]\} \\ &\le \cdots \le \{1 + [\alpha + o(L^{-2})][\exp(hn_L) - 1]\}^{m_L^{(1)}+1} \\ &\le \exp\left\{ \frac{L}{n_L} \log\{1 + [\alpha + o(L^{-2})][\exp(hn_L) - 1]\} \right\} \\ &\le \exp\left\{ \frac{tL}{2\log L} \log\{1 + [\alpha + o(L^{-2})][\exp(hn_L) - 1]\} \right\} \end{aligned} \tag{A3}$$

Then, for any $h > 0$,

$$\mathrm{P}\left(S_N > N\alpha + (\frac{2}{t}) N^{1/2} \log L\right) \le \exp\left\{ -hN\alpha - (\frac{2}{t}) hN^{1/2} \log L + \frac{tRL}{2\log L} \log\{1 + [\alpha + o(L^{-2})][\exp(hn_L) - 1]\} \right\}. \tag{A4}$$

By selecting $h = \frac{t}{2N^{1/2}\alpha(1-\alpha)}$, we can have,

$$\begin{aligned} &\mathrm{P}\left(S_N > N\alpha + (\frac{2}{t}) N^{1/2} \log L\right) \le \exp\left\{ -\frac{tN^{1/2}}{2(1-\alpha)} - \frac{\log L}{\alpha(1-\alpha)} + \frac{tN}{2\log L} \log\{1 + [\alpha + o(L^{-2})][\exp(\frac{\log L}{N^{1/2}\alpha(1-\alpha)}) - 1]\} \right\} \\ &= \exp\left\{ -\frac{(1 - t/4)\log L}{\alpha(1-\alpha)} + o(1) \right\} = O(L^{-r}) \end{aligned}$$

where the equality is obtained by applying Taylor's expansion on $\exp(\cdot)$ and $\log(1 + \cdot)$, and requires $R = o(L)$. Notice that $\alpha(1 - \alpha) \le 1/4$, for $t < 3$, we have $r > 1$, so (19) can directly follow Borel-Cantelli Lemma (Serfling, 2009). □

We now consider sequence of 0–1 valued random variables U_{ji} s.t. $U_{ji} = U(X_{ji})$ and $P(U_{ji} = 1) = 1 - P(U_{ji} = 0) = \alpha_N$ for all j and i. We can define S_j, S_N similar as before by replacing Y_{ji} with U_{ji}.

Lemma A.2. *If there exists positive K_1 and K_2 such that $K_1 N^{-3/4} \log L \le \alpha_N \le K_2 N^{-1/2} \log L$, for every positive C and s, there exists positive $C_s < \infty$ and $L_0(s)$, such that for $L \ge L_0(s)$,*

$$P\left\{\frac{S_N}{N}-\alpha_N>CN^{-3/4}\log L\right\}\le C_sL^{-s} \tag{A5}$$

Proof. For simplicity, let $C=1$, and then $\forall\varepsilon>0$, follow the similar procedure in the proof of Lemma A.1,

$$\begin{aligned}&\mathrm{P}\left\{\frac{S_N}{N}-\alpha_N>N^{-(3/4-\varepsilon)}\log L\right\}=\mathrm{P}\left\{S_N>N\alpha_N+N^{1/4+\varepsilon}\log L\right\}\\&\le\inf_{h>0}\left\{\exp\left[-hN\alpha_N-hN^{1/4+\varepsilon}\log L\right]\left(\mathrm{E}\left[\exp(hS_1^{(1)})\right]\right)^R\right\}\end{aligned}$$

where $S_j^{(\ell)}$ is also defined similarly as in Lemma A.1, by replacing Y_{ji} with U_{ji}. Again we choose $n_L=2t^{-1}\log L$, then (A2), (A3) still hold. for any $h>0$,

$$\begin{aligned}&\mathrm{P}\left\{\frac{S_N}{N}-\alpha_N>N^{-(3/4-\varepsilon)}\log L\right\}\le\exp\left\{-\left(\frac{2}{t}\right)hN^{1/4+\varepsilon}\log L\right.\\&\left.-hN\alpha_N+\frac{tRL}{2\log L}\log\{1+[\alpha_N+o(L^{-2})][\exp(hn_L)-1]\}\right\}.\end{aligned}$$

By selecting $h=\frac{s}{N^{1/4+\varepsilon}}$, we can have,

$$\begin{aligned}&\mathrm{P}\left\{\frac{S_N}{N}-\alpha_N>N^{-(3/4-\varepsilon)}\log L\right\}\le\exp\left\{-stN^{3/4-\varepsilon}\alpha_N-s\log L+\frac{tN}{2\log L}\log\{1+[\alpha_N+o(L^{-2})][\exp\left(\frac{2s\log L}{tN^{1/4+\varepsilon}}\right)-1]\}\right\}\\&=\exp\left\{-s\log L+\frac{s^2\alpha_N(1-\alpha_N)\log L}{t}N^{1/2-2\varepsilon}+o(1)\right\}\end{aligned}$$

where the last equality is obtained by applying Taylor's expansion on $\exp(\cdot)$ and $\log(1+\cdot)$, and requires $R=o(L)$. Since $\alpha_N\le K_2N^{-1/2}\log L$, $t^{-1}s^2\alpha_N(1-\alpha_N)\log LN^{1/2-2\varepsilon}\le O(N^{-2\varepsilon}(\log L)^2)=o(1)$, we have that

$$\mathrm{P}\left\{\frac{S_N}{N}-\alpha_N>N^{-(3/4-\varepsilon)}\log L\right\}\le\exp\{-s\log L+o(1)\}. \tag{A6}$$

Let $\varepsilon\to0$, then (A5) directly follows from (A6). □