QUANTILE SENSITIVITY ESTIMATION FOR DEPENDENT SEQUENCES

GUANGXIN JIANG,* *City University of Hong Kong*

MICHAEL C. FU,** *University of Maryland, College Park*

Abstract

In this paper, we estimate quantile sensitivities for dependent sequences via infinitesimal perturbation analysis (IPA), and prove asymptotic unbiasedness, weak consistency, and a central limit theorem for the estimators under some mild conditions. Two common cases, the regenerative setting and ϕ-mixing, are analyzed further, and a new batched estimator is constructed based on regenerative cycles for regenerative processes. Two numerical examples, the G/G/1 queue and the Ornstein-Uhlenbeck process, are given to show the effectiveness of the estimator.

Keywords: quantile; Monte Carlo simulation; sensitivity analysis; regenerative processes; ϕ-mixing

2010 Mathematics Subject Classification: Primary 62H12

Secondary 65C05

1. Introduction

Tail performance measures such as quantiles and tail probabilities are more appropriate than classical performance measure such as mean and variance when characterizing the extreme properties of a stochastic system. For example, Value-at-Risk (VaR), which is a quantile, is often used as a measure of risk in the management of a financial portfolio (Glasserman et al. [13]). Quantiles as performance measures have been studied widely in other fields like inventory control (Gelinas et al. [10]), queueing (Heidelberger and

* Postal address: Department of Economics and Finance, City University of Hong Kong, Kowloon, Hong Kong

** Postal address: Robert H. Smith School of Business and Institute for Systems Research, University of Maryland, College Park, Maryland, 20742, USA

Lewis [14]) and reliability (Nair and Sankaran [25]).

For complex stochastic systems, simulation is commonly used to estimate performance, for which Serfling [29] provides an overview of quantile estimation for independent and identically distributed (i.i.d.) data; recent work includes Ghosh and Pasupathy [11]. For dependent sequences, Sen [28] provides a representation of the quantile by order statistics and a confidence interval (CI) for the estimator. Other methods, such as interpolation (Iglehart [19]), batching (Seila [26]), maximum transformation (Heidelberger and Lewis [14]), and the piecewise-parabolic algorithm (Jain and Chlamtac [20]) have been developed. For some recent work, see Bekki et al. [4] and Alexopoulos and Wilson [1].

Our work addresses quantile sensitivity estimation. Previous work goes back to Hong [18], who introduced infinitesimal perturbation analysis (IPA) to estimate the quantile sensitivity based on analysis of a probability sensitivity. Jiang and Fu [21] presented an alternative more direct derivation of the IPA estimators for both batched and unbatched estimators. Fu, Hong and Hu [9] applied smoothed perturbation analysis (SPA) to derive a more general estimator with wider applicability and improved convergence rate. Other gradient estimators, such as a kernel estimator (cf. Liu and Hong [23]) and weak derivative estimator (cf. Heidergott and Volk-Makarewicz [15]) have also been introduced.

Previous work has primarily treated the i.i.d. setting. In this paper, we consider dependent processes, with a focus on estimating the quantile sensitivity of steady-state performance. A particular case is regenerative processes, which have many applications in inventory control and queueing theory, e.g., analyzing the steady-state waiting time. For more details about regenerative processes, refer to Asmussen [2] and Serfozo [30]. More generally, a strong mixing condition is introduced to capture asymptotic properties of dependent random variables (Bradley [5]), and a quantile sensitivity estimator derived for this setting.

The most closely related works are Hong [18], which provided a numerical example to illustrate that the IPA estimator can be used to estimate quantile sensitivities of the waiting time in an M/M/1 queue, and Liu and Hong [23], which considered stationary ϕ-mixing sequences for the kernel estimator. **The main contribution of this paper** is a rigorous theoretical analysis of the IPA estimator for more general

dependent sequences that are not necessarily stationary but which have a limiting distribution. For regenerative processes, we derive an alternative batched estimator based on regenerative cycles, which shows better performance than the original batched estimator. The rest of the paper is organized as follows: In Section 2, we briefly review the IPA estimator for quantile sensitivities in the i.i.d. setting and adapt this framework to dependent sequences, proving asymptotic unbiasedness and consistency of the estimator under some mild conditions. In Section 3, we consider two special cases, regenerative processes and ϕ-mixing processes, and provide their statistical properties. Numerical examples are given in Section 4 to show the effectiveness of the estimators for a G/G/1 queue and an Ornstein-Uhlenbeck (OU) process. Section 5 concludes.

2. IPA quantile sensitivity estimation

IPA has been used for estimating quantile sensitivities of i.i.d sequences (Hong [18] and Jiang and Fu [21]). In the setting of those papers, the quantile sensitivity estimator is obtained from the IPA derivatives of the corresponding order statistics. Since we consider sequences where each element may depend on previous elements and may not be identically distributed, such as waiting times in a queueing system, the results in the previous papers cannot be applied directly, which motivates us to provide an alternative approach. In this section, we first consider stationary processes, and then extend the results to non-stationary processes. Finally, we provide a batched estimator which is asymptotically unbiased and weakly consistent.

2.1. IPA estimator for quantile sensitivity in stationary processes

Let $\{X_t, t = 0, 1, 2, \ldots\}$ be a discrete-time stationary stochastic process with marginal cumulative distribution function (c.d.f.) $F(x;\theta) = \Pr\{X_t \leq x\}$, where $\theta \in \Theta$ is the parameter of interest, and let q_α denote the α-quantile of X_t ($\alpha \in (0,1)$), which is a number satisfying $\Pr\{X \leq q_\alpha\} \geq \alpha$ and $\Pr\{X \geq q_\alpha\} \geq 1-\alpha$. If X_t is a continuous random variable, then $\Pr\{X_t \leq q_\alpha\} = F(q_\alpha;\theta) = \alpha$, where $F(\cdot,\theta)$ has a continuous support. We make the following assumption:

A1. In a neighbourhood of $x = q_\alpha$, $F(x;\theta)$ is continuously differentiable with respect to (w.r.t.) both arguments. The density $\partial_1 F(x;\theta)$ is strictly positive for each

$\theta \in \Theta$.

Let $q'_\alpha \triangleq dq_\alpha/d\theta$ denote the α-quantile sensitivity of X_t w.r.t. θ. The setting of interest is where $F(x;\theta)$ is unknown, but X_t can be represented as a function of some other random variables that have known c.d.f.'s. For example, if X_t is the waiting time of a G/G/1 queue, it is a function of interarrival times $\{A_1, \ldots, A_t\}$ and service times $\{S_1, \ldots, S_t\}$, with c.d.f.'s assumed known. Since $\alpha = F(q_\alpha;\theta)$, differentiating both sides w.r.t. θ leads to

$$q'_\alpha = -\frac{\partial_2 F(q_\alpha;\theta)}{\partial_1 F(q_\alpha;\theta)}, \tag{1}$$

where ∂_i denotes the partial differentiation w.r.t. ith argument of F, and A1 guarantees the RHS of Equation (1) exists.

If $F(x;\theta)$ is known, we can first estimate the quantile q_α, then substitute into Equation (1). However, in general, $F(x;\theta)$ cannot be computed exactly, and we have to estimate q'_α by another approach. The order statistics of sequence $\{X_t, t = 1, \ldots, n\}$ are given by

$$X_{(1)} \leq X_{(2)} \cdots \leq X_{(\lceil \alpha n \rceil)} \leq \cdots \leq X_{(n)}. \tag{2}$$

Then we can define the sample quantile based on the corresponding order statistic

$$\hat{q}^n_\alpha = X_{(\lceil \alpha n \rceil)}, \tag{3}$$

where $\lceil x \rceil$ is the smallest integer greater than x. Then, we need the following additional assumption:

A2. $\hat{q}^n_\alpha \xrightarrow{d} q_\alpha$ as $n \to \infty$, where "$\xrightarrow{d}$" denotes convergence in distribution.

Remark 1. A1 is a typical assumption used in quantile sensitivity estimation (cf. Hong [18] and Jiang and Fu [21]). For i.i.d. sequences, $\hat{q}^n_\alpha$ converges to q_α w.p.1, which is the key step for proving the asymptotic unbiasedness of the quantile sensitivity estimator (Jiang and Fu [21]). Therefore, for dependent sequences, if it still holds that the quantile estimator converges to the true quantile, we can follow the framework in Jiang and Fu [21] with some technical adjustments. A2 holds for many classes of dependent sequences. For example, in regenerative processes, if the derivative of the density of the steady-state c.d.f. $\partial_1^2 F(x;\theta)$ exists and is bounded by M in a neighbourhood of q_α, then A2 holds (Iglehart [19]). We will discuss this case in detail in the next section.

Let $dX/d\theta$ denote the ordinary IPA estimator of X w.r.t. θ. For example, $X = A + \theta S$, where A is a random variable with parameter θ and S does not contain θ. If $dA/d\theta$ is well defined, then $dX/d\theta = dA/d\theta + S$.

For the sequence (2), the corresponding derivatives w.r.t. θ are given by

$$\frac{dX_{(1)}}{d\theta}, \frac{dX_{(2)}}{d\theta}, \cdots, \frac{dX_{(\lceil \alpha n \rceil)}}{d\theta}, \cdots, \frac{dX_{(n)}}{d\theta},$$

and define the IPA quantile sensitivity estimator by

$$I_n = \frac{dX_{(\lceil \alpha n \rceil)}}{d\theta}. \tag{4}$$

Theorem 1 in Jiang and Fu [21] is introduced to establish asymptotic unbiasedness of the IPA estimator for dependent sequences.

Lemma 1. (Jiang and Fu [21].) *Assume that X is differentiable w.r.t. $\theta \in \Theta$ w.p.1, and A1 is satisfied at the point y. Then,*

$$\mathbb{E}\left[\frac{dX}{d\theta} \middle| X = y \right] = -\frac{\partial_2 F(y;\theta)}{\partial_1 F(y;\theta)}, \tag{5}$$

where $F(\cdot;\theta)$ is the c.d.f. of X.

Equation (5) connects the IPA estimator $dX/d\theta$ with the c.d.f. of X through conditional expectation. By Fu [8], if this parameter is *location, scale* or *generalized scale* parameter, the conditional expectation can be removed. For example, consider the generalized scale parameter $dX/d\theta = (X - \bar{\theta})/\theta$, then $\mathbb{E}[dX/d\theta|X] = \mathbb{E}[(X - \bar{\theta})/\theta|X] = (X - \bar{\theta})/\theta$. Lemma 1 can be regarded as an extension of the classical results of the IPA estimator in Suri and Zazanis [32], where the random variable is simple and the derivative can be represented by the quotient of partial derivatives of c.d.f.'s. However, if X_t is a complicated function of other random variables, the IPA estimator cannot be expressed by the quotient form. For example, $X_t = \theta X_1 + X_2$, where X_1 and X_2 are independent standard normal random variables. Then, $X_t \sim N(0, \theta^2 + 1)$, and $-\partial_2 F(X_t;\theta)/\partial_1 F(X_t;\theta) = \theta X_t/(\theta^2 + 1)$. By taking partial derivatives directly, we obtain $dX_t/d\theta = X_1$, which does not equal $\theta X_t/(\theta^2 + 1)$. As in the i.i.d. case, the IPA quantile sensitivity estimator given by (4) is asymptotically unbiased.

Theorem 1. *Suppose that* $\sup_n \mathbb{E}[I_n^2] < \infty$. *If A1 and A2 hold, and $dX/d\theta$ exists, then $\mathbb{E}[I_n] \to q'_\alpha$ as $n \to \infty$.*

Proof. By Lemma 1,

$$\mathbb{E}\left[\frac{dX_t}{d\theta}\middle| X_t = x\right] = -\frac{\partial_2 F(x;\theta)}{\partial_1 F(x;\theta)}.$$

Substituting $x = q_\alpha$ into the equation above, and by Equation (1),

$$q'_\alpha = \mathbb{E}\left[\frac{dX_t}{d\theta}\middle| X_t = q_\alpha\right] = -\frac{\partial_2 F(q_\alpha;\theta)}{\partial_1 F(q_\alpha;\theta)}.$$

By A2, $\hat{q}^n_\alpha \xrightarrow{d} q_\alpha$ as $n \to \infty$, and by A1, for $i = 1, 2$,

$$\partial_i F(\hat{q}^n_\alpha, \theta) \xrightarrow{d} \partial_i F(q_\alpha, \theta),$$

so with the second part of A1,

$$-\frac{\partial_2 F(\hat{q}^n_\alpha,\theta)}{\partial_1 F(\hat{q}^n_\alpha;\theta)} \xrightarrow{d} -\frac{\partial_2 F(q_\alpha,\theta)}{\partial_1 F(q_\alpha;\theta)} \quad \text{as} \quad n \to \infty.$$

Note that

$$\begin{aligned}
\mathbb{E}\left[I_n\right] &= \mathbb{E}\left[\frac{dX_{(\lceil \alpha n \rceil)}}{d\theta}\right] = \int_{-\infty}^{\infty} \mathbb{E}\left[\frac{dX_{(\lceil \alpha n \rceil)}}{d\theta}\middle| X_{(\lceil \alpha n \rceil)} = x\right] dF_{X_{(\lceil \alpha n \rceil)}}(x) \qquad (6)\\
&= \int_{-\infty}^{\infty} -\frac{\partial_2 F(x;\theta)}{\partial_1 F(x;\theta)} dF_{X_{(\lceil \alpha n \rceil)}}(x)\\
&= \mathbb{E}\left[-\frac{\partial_2 F(\hat{q}^n_\alpha;\theta)}{\partial_1 F(\hat{q}^n_\alpha;\theta)}\right],
\end{aligned}$$

where $F_{X_{(\lceil \alpha n \rceil)}}(x)$ is the c.d.f. of $X_{(\lceil \alpha n \rceil)}$, and the last equality holds since we define $\hat{q}^n_\alpha = X_{(\lceil \alpha n \rceil)}$ in (3). In the following, it suffices to prove $\partial_2 F(\hat{q}^n_\alpha;\theta)/\partial_1 F(\hat{q}^n_\alpha;\theta)$ is uniformly integrable.

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\partial_2 F(\hat{q}^n_\alpha;\theta)}{\partial_1 F(\hat{q}^n_\alpha;\theta)}\right)^2\right] &= \int_{-\infty}^{\infty} \left(\mathbb{E}\left[\frac{dX_t}{d\theta}\middle| X_t = y\right]\right)^2 dF_{X_{(\lceil \alpha n \rceil)}}(y)\\
&\le \int_{-\infty}^{\infty} \mathbb{E}\left[\left(\frac{dX_t}{d\theta}\right)^2\middle| X_t = y\right] dF_{X_{(\lceil \alpha n \rceil)}}(y) = \mathbb{E}[I_n^2], \qquad (7)
\end{aligned}$$

where the inequality follows by Jensen's inequality for conditional expectation.

Since $\sup_n \mathbb{E}[I_n^2] < \infty$, $\partial_2 F(\hat{q}^n_\alpha;\theta)/\partial_1 F(\hat{q}^n_\alpha;\theta)$ is uniformly integrable. Therefore,

$$\mathbb{E}[I_n] = \mathbb{E}\left[-\frac{\partial_2 F(\hat{q}^n_\alpha;\theta)}{\partial_1 F(\hat{q}^n_\alpha;\theta)}\right] \to \mathbb{E}\left[-\frac{\partial_2 F(q_\alpha;\theta)}{\partial_1 F(q_\alpha;\theta)}\right] = q'_\alpha \text{ as } n \to \infty.$$

□

The proof is almost identical to the i.i.d. case, since the process is stationary, which means each random variable has the same c.d.f. However, if the process is not stationary but has a limiting distribution, e.g., the waiting time of a stable G/G/1 queue, we can still establish that the IPA estimator is asymptotically unbiased under some mild conditions.

2.2. IPA estimator for quantile sensitivity in non-stationary processes

Let $\{X_t, t = 0, 1, 2, \ldots\}$ be a real-valued discrete stochastic process (not necessarily stationary), such that there exists a steady-state random variable X satisfying $X_t \xrightarrow{d} X$ as $t \to \infty$. Stationary processes are the special case where $X_t \stackrel{d}{=} X$. Let $F(x;\theta)$ denote the c.d.f. of X, and q_α denote the α-quantile of X. The goal in this subsection is to estimate the quantile sensitivity of X, i.e., $dq_\alpha/d\theta$. Let $\{F_t(x;\theta), t = 0, 1, \ldots\}$ denote the c.d.f. of $\{X_t, t = 0, 1, 2, \ldots\}$, and we have $F_t(x;\theta) \to F(x;\theta)$ as $t \to \infty$. Since we need the regularity conditions for all the c.d.f.'s, A1 is replaced by A1$'$:

A1$'$. In a neighbourhood of $x = q_\alpha$, $\{F_t(x;\theta), t = 0, 1, 2, \ldots\}$ and $F(x;\theta)$ are continuously differentiable w.r.t. both arguments. The densities $\{\partial_1 F_t(x;\theta), t = 0, 1, 2, \ldots\}$ and $\partial_1 F(x;\theta)$ are strictly positive for each $\theta \in \Theta$.

Because each $X_{(t)}$ has a different c.d.f., the conditional expectation of $dX_{(t)}/d\theta$ has different expressions, which is one of the difficulties in proving the statistical properties. Moreover, $F_t(x;\theta)$ converging to $F(x;\theta)$ does not necessarily imply the partial derivatives of $F_t(x;\theta)$ converge to the partial derivatives of $F(x;\theta)$, so the following lemma is needed.

Lemma 2. *If A1$'$ is satisfied, and $F_t(x;\theta)$ satisfies the following conditions:*
(i) $F_t(x;\theta)$ converges uniformly to $F(x;\theta)$ on $[a,b] \times \Theta$;
(ii) $\partial_1 F_t(x,\theta)$ and $\partial_2 F_t(x,\theta)$ are uniformly convergent on $[a,b] \times \Theta$, where a, b can take value $\pm\infty$.
Then

$$\partial_1 F(x,\theta) = \lim_{t\to\infty} \partial_1 F_t(x,\theta),$$

$$\partial_2 F(x,\theta) = \lim_{i\to\infty} \partial_2 F_t(x,\theta).$$

Proof. Since $\partial_1 F_t(x,\theta)$ is a uniformly convergent sequence of continuous functions,

we have

$$\varphi(x,\theta) = \lim_{t\to\infty} \partial_1 F_t(x,\theta),$$

where the limit function $\varphi(x,\theta)$ is continuous on $[a,b]$ w.r.t. x for fixed θ. Then, it remains to show that $\varphi(x,\theta) = \partial_1 F(x,\theta)$. Because

$$\int_a^x \varphi(\eta,\theta)d\eta = \int_a^x \lim_{t\to\infty} \partial_1 F_t(\eta,\theta)d\eta = \lim_{t\to\infty} \int_a^x \partial_1 F_t(\eta,\theta)d\eta,$$

then

$$\begin{aligned}\int_a^x \varphi(\eta,\theta)d\eta &= \lim_{t\to\infty} [F_t(x,\theta) - F_t(a,\theta)] \\ &= F(x,\theta) - F(a,\theta),\end{aligned}$$

the second equality holds because of the uniform convergence of the original sequence $F_t(x,\theta)$. Differentiating both sides w.r.t. x leads to $\varphi(x,\theta) = \partial_1 F(x,\theta)$, i.e., $\partial_1 F(x,\theta) = \lim_{t\to\infty} \partial_1 F_t(x,\theta)$. Similarly, $\partial_2 F(x,\theta) = \lim_{t\to\infty} \partial_2 F_t(x,\theta)$. □

Remark 2. If $X_t \xrightarrow{d} X$, by A1′, $F(x;\theta)$ is continuous w.r.t. x, $\sup_{x\in[a,b]} |F_t(x;\theta) - F(x;\theta)| \to 0$, i.e., $F_t(x;\theta)$ converges uniformly to $F(x;\theta)$ w.r.t. x. Moreover, when Θ is a compact set, pointwise convergence is equivalent to uniform convergence.

For ease of notation, let $F_{(t)}$ denote the c.d.f. from which $X_{(t)}$ has been sampled, and note that $F_{(t)}$ is NOT the order statistics distribution. For example, suppose that $\{x_t, t = 0,1,\ldots,5\}$ is a sequence of observations of $\{X_t, t = 0,1,\ldots,5\}$ with c.d.f.'s $\{F_t, t = 0,1,\ldots,5\}$, which satisfies $x_0 < x_5 < x_3 < x_2 < x_1 < x_4$. So for this realization, the 4th order statistic observation $x_{(4)}$ corresponds to x_2, which is sampled from X_2 with c.d.f. F_2, so $F_{(4)}$ corresponds to F_2 for this sample. Similar to the i.i.d. case, the estimator (4) is asymptotically unbiased under some mild conditions.

Theorem 2. *Suppose that* $\sup_n \mathbb{E}[I_n^2] < \infty$ *and* $F_{(\lceil \alpha n \rceil)}(x;\theta)$ *satisfies conditions (i) and (ii) in Lemma 2. If A1′ and A2 hold, and* $dX/d\theta$ *exists, then* $\mathbb{E}[I_n] \to q'_\alpha$ *as* $n \to \infty$.

Proof. Similar to the proof in Theorem 1, the quantile sensitivity of X can be obtained by

$$q'_\alpha = \mathbb{E}\left[\frac{dX}{d\theta}\middle| X = q_\alpha\right] = -\frac{\partial_2 F(q_\alpha;\theta)}{\partial_1 F(q_\alpha;\theta)}. \tag{8}$$

Equation (8) represents a conditional expectation of the derivative of the random variable X as a quotient of partial derivatives of its c.d.f., i.e., $-\partial_2 F(X;\theta)/\partial_1 F(X;\theta)$.

Analogously, for the random variables $\{X_t, t = 0, 1, 2, \ldots\}$,

$$\mathbb{E}\left[\frac{dX_t}{d\theta}\middle| X_t = x\right] = -\frac{\partial_2 F_t(x;\theta)}{\partial_1 F_t(x;\theta)}, t = 0, 1, 2, \ldots. \tag{9}$$

Now consider the expectation of $dX_{(\lceil \alpha n \rceil)}/d\theta$. By the definition of $F_{(\lceil \alpha n \rceil)}(x;\theta)$, it is one of $\{F_t(x;\theta), t = 0, 1, 2, \ldots\}$, so it satisfies the conditions in Lemma 1 with A1 replaced by A1′, so

$$\mathbb{E}\left[\frac{dX_{(\lceil \alpha n \rceil)}}{d\theta}\middle| X_{(\lceil \alpha n \rceil)} = x\right] = -\frac{\partial_2 F_{(\lceil \alpha n \rceil)}(x;\theta)}{\partial_1 F_{(\lceil \alpha n \rceil)}(x;\theta)}. \tag{10}$$

Then, similar to Equation (6),

$$\mathbb{E}[I_n] = \mathbb{E}\left[-\frac{\partial_2 F_{(\lceil \alpha n \rceil)}(\hat{q}^n_\alpha;\theta)}{\partial_1 F_{(\lceil \alpha n \rceil)}(\hat{q}^n_\alpha;\theta)}\right].$$

Next, we will prove $|\partial_i F_{(\lceil \alpha n \rceil)}(\hat{q}^n_\alpha;\theta) - \partial_i F(q_\alpha,\theta)| \to 0$ in distribution for $i = 1, 2$ as $n \to \infty$.

$$|\partial_i F_{(\lceil \alpha n \rceil)}(\hat{q}^n_\alpha;\theta) - \partial_i F(q_\alpha,\theta)| \le |\partial_i F_{(\lceil \alpha n \rceil)}(\hat{q}^n_\alpha;\theta) - \partial_i F(\hat{q}^n_\alpha,\theta)| + |\partial_i F(\hat{q}^n_\alpha,\theta) - \partial_i F(q_\alpha,\theta)|.$$

By Lemma 2, $\partial_i F_{(\lceil \alpha n \rceil)}(x,\theta)$ is uniformly convergent to $\partial_i F(x,\theta)$ for $i = 1, 2$. So for any $\epsilon > 0$, there exists N' such that for all $n \ge N'$, $|\partial_i F_{(\lceil \alpha n \rceil)}(\hat{q}^n_\alpha;\theta) - \partial_i F(\hat{q}^n_\alpha,\theta)| < \epsilon/2$ a.s., for all $\theta \in \Theta$. By A1′, $F(x,\theta)$ is continuously differentiable w.r.t. both arguments, i.e., $\partial_i F(x,\theta)$ is continuous w.r.t both x and θ for $i = 1, 2$. Then by A2, for any $\epsilon > 0$, there exists N'' such that for all $n \ge N''$, $|\partial_i F(\hat{q}^n_\alpha,\theta) - \partial_i F(q_\alpha,\theta)| < \epsilon/2$ in distribution. So it is easy to prove that $|\partial_i F_{(\lceil \alpha n \rceil)}(\hat{q}^n_\alpha;\theta) - \partial_i F(q_\alpha,\theta)| \to 0$ in distribution for $i = 1, 2$ as $n \to \infty$, i.e., $\partial_i F_{(\lceil \alpha n \rceil)}(\hat{q}^n_\alpha;\theta) \to \partial_i F(q_\alpha,\theta)$ in distribution for $i = 1, 2$ as $n \to \infty$. By A1′,

$$\frac{\partial_2 F_{(\lceil \alpha n \rceil)}(\hat{q}^n_\alpha;\theta)}{\partial_1 F_{(\lceil \alpha n \rceil)}(\hat{q}^n_\alpha;\theta)} \xrightarrow{d} \frac{\partial_2 F(q_\alpha;\theta)}{\partial_1 F(q_\alpha;\theta)}.$$

Similar to Equation (7), $\partial_2 F_{(\lceil \alpha n \rceil)}(x;\theta)/\partial_1 F_{(\lceil \alpha n \rceil)}(x;\theta)$ is uniformly integrable. Therefore, $\mathbb{E}[I_n] \to q'_\alpha$ as $n \to \infty$. □

Remark 3. (i) If all $\{F_t(x;\theta), t = 0, 1, \ldots\}$ satisfy (i) and (ii) in Lemma 2, $F_{(\lceil \alpha n \rceil)}(x;\theta)$ satisfies the conditions automatically. (ii) Although $X_t \xrightarrow{d} X$, i.e., $F_t(x;\theta) \to F(x;\theta)$ as $t \to \infty$, it does not directly imply $F_{(\lceil \alpha n \rceil)}(x;\theta) \to F(x;\theta)$ as $n \to \infty$. However, after an appropriate "warm-up" period, all the c.d.f's can be regarded as the same as the steady-state c.d.f. $F(x;\theta)$. Technically, as the length of the warm-up period

$n' \to \infty$, then $F_{n'+t}(x;\theta) \to F(x;\theta)$. Since $F_{(\lceil \alpha n \rceil)}(x;\theta)$ is one of $\{F_{n'+t}(x;\theta), t = 0,1,2,\ldots\}$, $F_{(\lceil \alpha n \rceil)}(x;\theta) \to F(x;\theta)$. We use this idea in the next section. Consider a long run simulation sequence, and divide this sequence into k batches of batch size $m_i, i = 1,2,\ldots,k$. When $m_1 \to \infty$, all the c.d.f's of X_t in the ith batch ($i \geq 2$) converge to the steady-state c.d.f. $F(x;\theta)$. Hence, in each batch, the corresponding $F_{(\lceil \alpha m_i \rceil)}(x;\theta) \to F(x;\theta)$.

2.3. Batched estimator

In general, the IPA estimator I_n is asymptotically unbiased but not necessarily consistent (Hong [18] and Jiang and Fu [21]). To estimate the quantile sensitivity, we divide this long-run sequence into batches, and take the mean of the IPA estimators derived from each batch, i.e., the nonoverlapping batch means method, widely used simulation method for analyzing the steady-state mean of a stochastic process (Steiger and Wilson [31]). By Theorem 1 and 2, I_n is asymptotically unbiased, i.e., $\mathbb{E}[I_n] \to q'_\alpha$, so this batched estimator may also be regarded as approximating the expected value of I_n via the Strong Law of Large Numbers (SLLN).

Specifically, for a sequence of length n $\{X_t, t = 0,1,2,\ldots,n-1\}$, take k batches of length m_i, $i = 1,2,\ldots,k$, where $n = \sum_{i=1}^{k} m_i$, so the ith batch is based on samples

$$X_{\sum_{j=1}^{i-1} m_j}, X_{\sum_{j=1}^{i-1} m_j+1}, \ldots, X_{\sum_{j=1}^{i-1} m_j+m_i-1}.$$

Let I_{i,m_i} denote the IPA estimator derived from the ith batch with batch size m_i, then the batched estimator is given by

$$\hat{q}'^{m}_{\alpha,k} = \frac{1}{k}\sum_{i=1}^{k} I_{i,m_i}. \tag{11}$$

In practice, we often take m_i to be a constant, but in some cases the m_i may be different or even random variables. For ease of notation, denote $m = \min_i m_i$, so when letting $m \to \infty$, it means all $m_i \to \infty$.

Next, we study the consistency of the batched estimators. Consider a regenerative process, where choosing the batch size based on regenerative cycles eliminates the dependence among $\{I_{i,m_i}, i = 1,2,\ldots,k\}$, so $Var(\hat{q}'^{m}_{\alpha,k}) = 1/k^2 \sum_{i=1}^{k} Var(I_{i,m_i})$. Suppose that $\max_i Var(I_{i,m_i}) < \infty$, when $k \to \infty$, $Var(\hat{q}'^{m}_{\alpha,k}) < 1/k \max_i Var(I_{i,m_i}) \to 0$, and we can apply Chebyshev's inequality to prove weak consistency of the batched

estimator. More generally, we have the following theorem for stationary processes.

Theorem 3. *Suppose that for each i, $\sup_{m_i} \mathbb{E}[I_{i,m_i}^2] < \infty$ and $Var(\hat{q}'^m_{\alpha,k}) \to 0$ as $k \to \infty$. If A1 and A2 hold, and $dX/d\theta$ exists, then $\hat{q}'^m_{\alpha,k} \to q'_\alpha$ in probability as $m \to \infty$ and $k \to \infty$.*

Proof. Let $m \to \infty$, by Theorem 1, for each i, $\mathbb{E}[I_{i,m_i}] \to q'_\alpha$, then

$$\mathbb{E}[\hat{q}'^m_{\alpha,k}] = \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k} I_{i,m_i}\right] = \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[I_{i,m_i}\right] \to q'_\alpha.$$

By Chebyshev's inequality, for any $\epsilon > 0$,

$$\Pr\{|\hat{q'}^m_{\alpha,k} - \mathbb{E}[\hat{q}'^m_{\alpha,k}]| \geq \epsilon\} \leq \frac{Var[\hat{q'}^m_{\alpha,k}]}{\epsilon^2}.$$

Since $\sup_{m_i} \mathbb{E}[I_{i,m_i}^2] < \infty$ for each i, when $k \to \infty$, $Var(\hat{q}'^m_{\alpha,k}) \to 0$, then $\hat{q}'^m_{\alpha,k} \to \mathbb{E}[\hat{q}'^m_{\alpha,k}]$ in probability, so $\lim_{m\to\infty}\lim_{k\to\infty} \hat{q}'^m_{\alpha,k} = q'_\alpha$ in probability.

Similarly, when $k \to \infty$ first, $\hat{q}'^m_{\alpha,k} \to \mathbb{E}[\hat{q}'^m_{\alpha,k}]$ in probability. Then, let $k \to \infty$, and it is easy to prove $\lim_{k\to\infty}\lim_{m\to\infty} \hat{q}'^m_{\alpha,k} = q'_\alpha$ in probability. □

For non-stationary processes in subsection 2.2, we can still prove that the batched estimator (11) is weakly consistent. According to Remark 3, if $m_1 \to \infty$, the c.d.f.'s from the second batch onwards follow the steady-state c.d.f., where the first batch can be regarded as the warm-up period. Then letting the number of batches $k \to \infty$ yields consistency of the batched estimator as formalized in the following theorem.

Theorem 4. *Suppose that for each i, $\sup_{m_i} \mathbb{E}[I_{i,m_i}^2] < \infty$ and $Var(\hat{q}'^m_{\alpha,k}) \to 0$ as $k \to \infty$. Additionally, $\{F_t(x;\theta), t = 0, 1, 2, \ldots\}$ satisfies condition (ii) in Lemma 2. If A1′ and A2 hold, and $dX/d\theta$ exists, then $\hat{q}'^m_{\alpha,k} \to q'_\alpha$ in probability as $m \to \infty$ and $k \to \infty$.*

Proof. Separate $\hat{q}'^m_{\alpha,k}$ into two parts, the first batch and the following batches, i.e.,

$$\hat{q}'^m_{\alpha,k} = \frac{1}{k}\sum_{i=1}^{k} I_{i,m_i} = \frac{1}{k} I_{1,m_1} + \frac{1}{k}\sum_{i=2}^{k} I_{i,m_i},$$

then,

$$\mathbb{E}[\hat{q}'^m_{\alpha,k}] = \frac{1}{k}\mathbb{E}\left[I_{1,m_1}\right] + \frac{1}{k}\sum_{i=2}^{k}\mathbb{E}\left[I_{i,m_i}\right].$$

When $m \to \infty$, by Remark 3, we know $\mathbb{E}[I_{i,m_i}] \to q'_\alpha$, for $i = 2, \ldots, k$, then $\mathbb{E}[\hat{q}'^m_{\alpha,k}] \to (1/k)\gamma + (1-1/k)q'_\alpha$, where $\gamma = \lim_{m_1 \to \infty} \mathbb{E}[I_{1,m_1}] < \infty$, since for each i, $\sup_{m_i} \mathbb{E}[I^2_{i,m_i}] < \infty$. By Chebyshev's inequality, for any $\epsilon > 0$,

$$\Pr\{|\hat{q'}^m_{\alpha,k} - \mathbb{E}[\hat{q}'^m_{\alpha,k}]| \geq \epsilon\} \leq \frac{Var[\hat{q'}^m_{\alpha,k}]}{\epsilon^2}.$$

By assumption when $k \to \infty$, $Var(\hat{q}'^m_{\alpha,k}) \to 0$, and $(1/k)\gamma + (1-1/k)q'_\alpha \to q'_\alpha$. Therefore $\lim_{m\to\infty} \lim_{k\to\infty} \hat{q}'^m_{\alpha,k} = q'_\alpha$ in probability.

Similarly, when $k \to \infty$, first, $\hat{q}'^m_{\alpha,k} \to \mathbb{E}[\hat{q}'^m_{\alpha,k}]$ in probability. Then, let $k \to \infty$, and it is easy to prove $\mathbb{E}[\hat{q}'^m_{\alpha,k}] \to q'_\alpha$. □

Remark 4. (i) Note that in Theorem 4, we do NOT need $F_{(\lceil \alpha n \rceil)}(x;\theta)$ to satisfy condition (i) in Lemma 2, because, $F_{(\lceil \alpha n \rceil)}(x;\theta)$ satisfies this condition automatically after the second batch, since the effect of the bias of the first batch is eliminated by letting $k \to \infty$. (ii) For i.i.d. sequences, the SLLN can be used to prove the strong consistency of the batched estimator, but for nonstationary dependent sequences, neither the SLLN nor the Ergodic Theorem can be directly applied.

In the i.i.d. case, the batched estimator is the average of the IPA estimators, for which a central limit theorem is available directly. However, in the dependent case, any central limit theorem requires additional special conditions, such as the ϕ-mixing condition. In the next section, we consider two common dependent conditions, for which central limit theorems can be established.

3. Regenerative processes and mixing processes

Regenerative processes are used to model stochastic phenomena in which an event occurs repeatedly over time, and the times between occurrences are i.i.d. Alternatively, mixing conditions are used to measure dependence in stochastic processes (cf. Doukhan [7]), and a common mixing condition is the ϕ-mixing condition. In this section, we study these two types of dependent processes.

3.1. Quantile sensitivity for regenerative processes

Let $\{X_t, t = 0, 1, 2, \ldots\}$ be a real-valued discrete regenerative stochastic process s.t. there exists a sequence $0 = T_0 \leq T_1 \leq T_2 \leq \ldots$ called regenerative points, which makes

the process $\{X_t, t \geq T_i\}, i = 1, 2, \ldots$ equal in distribution to the process $\{X_t, t \geq T_1\}$. The segment of the process $\{X_t, T_{i-1} \leq t \leq T_i, i = 1, 2, \ldots\}$ is called the ith *cycle*, and the set of regenerative cycles form an i.i.d sequence. Classical examples of discrete regenerative processes are ergodic Markov chains in discrete time. In this paper, we consider the non-delayed regenerative processes, i.e., the first cycle $\{X_t, T_0 \leq t \leq T_1\}$ has the same c.d.f. as other cycles. Denote $N_i = T_i - T_{i-1}$ as the length of the ith cycle, and $\mu = \mathbb{E}[N_i]$ as the expected length of cycles. By Theorem 1.2 of Chapter 4 in Asmussen [2], if μ is finite, X_t converges to a random variable X in distribution (the steady-state distribution). For example, in a G/G/1 queue, denote W_i as the waiting time of the ith customer, and W_i is a regenerative process where a regeneration occurs when the system is idle. Assuming that the mean service time is less than the mean interarrival time, then $\mathbb{E}[N_i]$ is finite, and W_i converges to a steady-state random variable W as $i \to \infty$. As in the last section, $F_t(x;\theta)$ and $F(x;\theta)$ denote the c.d.f's. of X_t and X, respectively. Using the special structure of regenerative processes, we select the batch size based on cycles.

Suppose that the simulation runs $\tilde{n}$ cycles, generating $X_0, X_1, \ldots, X_{T_{\tilde{n}}-1}$ with total number of samples $T_{\tilde{n}}$. Then the order statistics of this sequence and their corresponding IPA estimator are given by

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(\lceil \alpha T_{\tilde{n}} \rceil)} \leq \cdots \leq X_{(T_{\tilde{n}})},$$

$$\frac{dX_{(1)}}{d\theta}, \frac{dX_{(2)}}{d\theta}, \ldots, \frac{dX_{(\lceil \alpha T_{\tilde{n}} \rceil)}}{d\theta}, \ldots, \frac{dX_{(T_{\tilde{n}})}}{d\theta}.$$

Denote $\hat{q}_\alpha^{\tilde{n}} = X_{(\lceil \alpha T_{\tilde{n}} \rceil)}$ and $I_{\tilde{n}} = dX_{(\lceil \alpha T_{\tilde{n}} \rceil)}/d\theta$. By Iglehart [19], we know in a neighbourhood of q_α, if the derivative of density of X exists and bounded, then

$$\frac{\hat{q}_\alpha^{\tilde{n}} - q_\alpha}{\sigma_\alpha / (\mu \partial_1 F(q_\alpha;\theta) \tilde{n}^{\frac{1}{2}})} \xrightarrow{d} \mathcal{N}(0,1), \text{ as } \tilde{n} \to \infty, \tag{12}$$

where σ_α is the standard deviation of $Y_i(q_\alpha) - N_i\alpha$, and $Y_i(q_\alpha)$ is the number of observations in the ith cycle which are less than or equal to q_α. Equation (12) also implies that $\hat{q}_\alpha^{\tilde{n}} \to q_\alpha$ in distribution as $\tilde{n} \to \infty$. Then Theorem 1 holds with assumption A2 replaced by A2′:

A2′. In a neighbourhood of q_α, $\partial_1^2 F(x;\theta)$ exists and is bounded by some $K < \infty$ for all x.

Corollary 1. *Suppose that* $\sup_{\tilde{n}} \mathbb{E}[I_{\tilde{n}}^2] < \infty$ *and* $F_{(\lceil \alpha T_{\tilde{n}} \rceil)}(x;\theta)$ *satisfies conditions (i) and (ii) in Lemma 2. If A1′ and A2′ hold, and* $dX/d\theta$ *exists, then* $\mathbb{E}[I_{\tilde{n}}] \to q'_\alpha$ *as* $\tilde{n} \to \infty$,

We now provide an alternative batched estimator based on regenerative cycles, analogous to the estimator used for estimating quantiles in Seila [26]. Suppose that these $\tilde{n}$ cycles are divided into $\tilde{k}$ batches with each batch containing $\tilde{m}$ cycles. The ith batch has sample size $M_i(\tilde{m}) = T_{i\tilde{m}} - T_{(i-1)\tilde{m}} = N_{(i-1)\tilde{m}+1} + \ldots + N_{i\tilde{m}}$, i.e., it contains samples $X_{T_{(i-1)\tilde{m}}}, X_{T_{(i-1)\tilde{m}}+1}, \ldots, X_{T_{i\tilde{m}}-1}$. The lth-order statistic derived from the ith batch will be denoted by the subscript (l,i), i.e., for the ith batch and sample size $M_i(\tilde{m})$,

$$X_{(1,i)} \le X_{(2,i)} \le \cdots \le X_{(\lceil \alpha M_i(\tilde{m}) \rceil, i)} \le \cdots \le X_{(M_i(\tilde{m}),i)}.$$

The corresponding derivatives w.r.t. θ are given by

$$\frac{dX_{(1,i)}}{d\theta}, \ \frac{dX_{(2,i)}}{d\theta}, \ \cdots, \frac{dX_{(\lceil \alpha M_i(\tilde{m}) \rceil, i)}}{d\theta}, \ \cdots, \frac{dX_{(M_i(\tilde{m}),i)}}{d\theta}.$$

We batch the $\tilde{k}$ estimators, and get the batched estimator

$$\hat{q}'^{\tilde{m}}_{\alpha,\tilde{k}} = \frac{1}{\tilde{k}} \sum_{i=1}^{\tilde{k}} I_{i,\tilde{m}}. \tag{13}$$

Although $\{X_t, t = 0, 1, \ldots\}$ is regenerative, establishing that their derivatives $\{dX_t/d\theta, t = 0, 1, \ldots\}$ are regenerative requires additional special structure. One condition is to assume $dX_t/d\theta = \phi(X_t;\theta)$, i.e., the derivative of X_t is only a function of X_t and does not contain other random variables. By A1, we know $\phi(X_t;\theta) = -\partial_2 F(X_t;\theta)/\partial_1 F(X_t;\theta)$ is continuous w.r.t the first argument. Since X_t is regenerative, then $dX_t/d\theta = \phi(X_t;\theta)$ is still regenerative. Glasserman [12] considered more general conditions for regenerative processes whose derivatives are also regenerative. We summarize the results in the following lemma.

Lemma 3. (Glasserman [12].) *Denote* $X_t := X_t(\theta, U_t(\theta))$*, where the input* $U_t(\theta), t = 0, 1, 2, \ldots$ *is an i.i.d. sequence. Suppose that the following conditions are satisfied:*
(i) Each X_t *and* U_t*,* $t \ge 0$*, is continuous in a neighbourhood of some point* θ *and differentiable at* θ *a.s.;*
(ii) $\{X_t, t = 0, 1, 2, \ldots\}$ *is regenerative such that for any open recurrent set* B *and*

some open $A_i, i = 1, \ldots, r$, *if* $X_t(\theta) \in B$ *and* $U_{n+i}(\theta) \in A_i, i = 1, \ldots, r$, *then* $X_{t+r}(\theta) = h(U_n(\theta), \ldots, U_{n+r}(\theta))$, *where* h *is a measurable function:* $A_1 \times \ldots \times A_r \to B$; *(iii)* $\Pr\{h$ *is continuously differentiable at* $(U_1(\theta), \ldots, U_r(\theta)) | U_i(\theta) \in A_i, i = 1, \ldots, r\} = 1$.

Then $\{dX_t/d\theta, t = 0, 1, \ldots\}$ *is regenerative.*

As an example, consider the waiting time of an M/M/1 queue, where $\{U_i(\theta)\}$ are the input exponential random variables and X_n is the waiting time. If $X_n = 0$, X_{n+r} can be expressed in terms of $U_{n+i}, i = 1, \ldots, r$.

As the IPA estimators are derived from different regenerative cycles, and each batch has the same number of cycles, they are automatically i.i.d. If $\sup_{\tilde{m}} \mathbb{E}[I^2_{\tilde{m}}] < \infty$, then $Var(I_{\tilde{m}}) < \infty$ and $Var(\hat{q}'^{\tilde{m}}_{\alpha,\tilde{k}}) = 1/\tilde{k} Var(I_{\tilde{m}}) \to 0$ as $k \to \infty$, which satisfies the conditions in Theorem 3. The following corollary states weak consistency of the batched estimator for regenerative processes.

Corollary 2. *Under the assumptions in Lemma 3, suppose that* $\sup_{\tilde{m}} \mathbb{E}[I^2_{\tilde{m}}] < \infty$. *If A1′ and A2′ are satisfied, and* $dX/d\theta$ *exists, then* $\hat{q}'^{\tilde{m}}_{\alpha,\tilde{k}} \to q'_\alpha$ *in probability as* $\tilde{m} \to \infty$ *and* $\tilde{k} \to \infty$.

In the i.i.d. setting, $\mathbb{E}[\hat{q}^m_\alpha - q_\alpha] = O(m^{-1})$ and $\mathbb{E}[(\hat{q}^m_\alpha - q_\alpha)^2] = O(m^{-1})$, where m is the sequence sample size, but the convergence rate of the quantile estimator in the dependent setting is not well studied, and clearly depends on the type of dependence, so we assume respective rates $O(\tilde{m}^{-\gamma_1})$ and $O(\tilde{m}^{-\gamma_2})$, $\gamma_1, \gamma_2 > 0$, where γ_1 and γ_2 are used in the following central limit theorem for the batched estimator (13) similar to the i.i.d. case. Although charactering the convergence rate is beyond the scope of this paper, in the first numerical example, γ_1 and γ_2 lie in $[0.5, 1]$.

Theorem 5. *Assume* $\varphi(x) = -\partial_2 F(x;\theta)/\partial_1 F(x;\theta)$ *is twice differentiable and* $|\varphi''(x)| < K$, $\mathbb{E}[\hat{q}^{\tilde{m}}_\alpha - q_\alpha] = O(\tilde{m}^{-\gamma_1})$ *and* $\mathbb{E}[(\hat{q}^{\tilde{m}}_\alpha - q_\alpha)^2] = O(\tilde{m}^{-\gamma_2})$. *Under the assumptions in Corollary 2, if* $\inf_{\tilde{m}} Var(I_{\tilde{m}}) > 0$, *then*

$$\frac{\hat{q}'^{\tilde{m}}_{\alpha,\tilde{k}} - q'_\alpha}{\sigma(\hat{q}'^{\tilde{m}}_{\alpha,\tilde{k}})} \xrightarrow{d} \mathcal{N}(0,1) \quad as \quad \tilde{m} \to \infty,\ \tilde{k} \to \infty \text{ and } \frac{\tilde{k}^{\frac{1}{2}}}{\tilde{m}^{\gamma}} \to 0, \tag{14}$$

where $\sigma(\hat{q}'^{\tilde{m}}_{\alpha,\tilde{k}})$ *is the standard deviation of* $\hat{q}'^{\tilde{m}}_{\alpha,\tilde{k}}$ *and* $\gamma = \min\{\gamma_1, \gamma_2\}$.

Proof. First, we get the convergence rate of the quantile sensitivity estimator $I_{\tilde{m}}$. By the proof of Theorem 2, when $\tilde{m} \to \infty$,

$$\mathbb{E}[I_{\tilde{m}}] - q'_\alpha = \mathbb{E}[\varphi(\hat{q}_\alpha^{\tilde{m}}) - \varphi(q_\alpha)].$$

By Taylor's theorem,

$$\varphi(\hat{q}_\alpha^{\tilde{m}}) - \varphi(q_\alpha) = \varphi'(q_\alpha)(\hat{q}_\alpha^{\tilde{m}} - q_\alpha) + \varphi''(\xi)(\hat{q}_\alpha^{\tilde{m}} - q_\alpha)^2,$$

where ξ is between $\hat{q}_\alpha^{\tilde{m}}$ and q_α. Then

$$|\mathbb{E}[I_{\tilde{m}}] - q'_\alpha| = |\mathbb{E}[\varphi(\hat{q}_\alpha^{\tilde{m}}) - \varphi(q_\alpha)]| \leq |\varphi'(q_\alpha)||\mathbb{E}[\hat{q}_\alpha^{\tilde{m}} - q_\alpha]| + |\varphi''(\xi)||\mathbb{E}[(\hat{q}_\alpha^{\tilde{m}} - q_\alpha)^2]|.$$

Since $\varphi''(x)$ is bounded by K, the convergence rate is determined by the slowest of $\mathbb{E}[\hat{q}_\alpha^{\tilde{m}} - q_\alpha]$ and $\mathbb{E}[(\hat{q}_\alpha^{\tilde{m}} - q_\alpha)^2]$, so $\mathbb{E}[I_{\tilde{m}}] - q'_\alpha$ is $O(\tilde{m}^{-\gamma})$.

Next the left-hand side of Equation (14) can be written as

$$\frac{\hat{q'}_{\alpha,\tilde{k}}^{\tilde{m}} - q'_\alpha}{\sigma(\hat{q'}_{\alpha,\tilde{k}}^{\tilde{m}})} = \frac{\hat{q'}_{\alpha,\tilde{k}}^{\tilde{m}} - \mathbb{E}\left[\hat{q'}_{\alpha,\tilde{k}}^{\tilde{m}}\right]}{\sigma(\hat{q'}_{\alpha,\tilde{k}}^{\tilde{m}})} + \frac{\mathbb{E}\left[\hat{q'}_{\alpha,\tilde{k}}^{\tilde{m}}\right] - q'_\alpha}{\sigma(\hat{q'}_{\alpha,\tilde{k}}^{\tilde{m}})}.$$

By definition, $\hat{q'}_{\alpha,\tilde{k}}^{\tilde{m}}$ is the sample average of i.i.d replications of $I_{\tilde{m}}$, and by the Lindeberg-Levy central limit theorem, the first term on the right-hand side of (14) satisfies

$$\frac{\hat{q'}_{\alpha,\tilde{k}}^{\tilde{m}} - \mathbb{E}\left[\hat{q'}_{\alpha,\tilde{k}}^{\tilde{m}}\right]}{\sigma(\hat{q'}_{\alpha,\tilde{k}}^{\tilde{m}})} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{as} \quad \tilde{k} \to \infty.$$

By condition $\mathbb{E}[I_{\tilde{m}}] - q'_\alpha = O(\tilde{m}^{-\gamma})$, $|\mathbb{E}[I_{\tilde{m}}] - q'_\alpha|$ can be bounded by $L/(\tilde{m}^\gamma)$ for some constant L. Since $Var(\hat{q}'^{\tilde{m}}_{\alpha,\tilde{k}}) = 1/\tilde{k} Var(I_{\tilde{m}})$ and $\inf_{\tilde{m}} Var(I_{\tilde{m}}) > 0$, we can obtain

$$\left|\frac{\mathbb{E}\left[\hat{q}'^{\tilde{m}}_{\alpha,\tilde{k}}\right] - q'_\alpha}{\sigma(\hat{q}'^{\tilde{m}}_{\alpha,\tilde{k}})}\right| \leq \frac{L}{(Var[I_{\tilde{m}}])^{\frac{1}{2}}} \frac{\tilde{k}^{\frac{1}{2}}}{\tilde{m}^\gamma} \to 0 \text{ as } \frac{\tilde{k}^{\frac{1}{2}}}{\tilde{m}^\gamma} \to 0.$$

The theorem is proved. □

The batched estimator for regenerative processes is different from the previous section. To distinguish these two batched estimators, we call the batched estimator (11) based on deterministic batch sizes the deterministic batched estimator (DET), and the batched estimator in this subsection based on a fixed number of regenerative cycles the regenerative batched estimator (REG).

At the end of this section, we compare the different two batched estimators. Intuitively, the DET is easy to implement, but it does not exploit the structure of regenerative sequences. In contrast, the REG takes full advantage of the regenerative properties. Moreover, by Meketon and Heidelberger [24], the last cycle contains significantly more information than a typical cycle. Hong [18] used the DET to estimate the quantile sensitivities of the waiting time of a G/G/1 queue, and we will compare these two batched estimators for the same numerical example in the next section.

3.2. Quantile sensitivity for other classes of dependent processes

In this subsection, we consider other classes of dependent processes. As discussed above, the key for the validity of the IPA estimator is the quantile estimator converging to the true quantile. Therefore, for other classes of dependent segments that satisfy this condition, we can still use the previous method to estimate the quantile sensitivity. Some other dependent conditions that have been considered include m-dependent processes and ϕ-mixing processes (Sen [27], Sen [28], and Heidelberger and Lewis [14]), short-range dependent (SRD) linear processes (Hesse [16]), and long-range dependent (LRD) linear processes (Ho and Hsing [17]). In this subsection, we consider dependent processes satisfying a ϕ-mixing condition, which was considered in the stationary setting by Liu and Hong [23] for a kernel estimator.

All the notations are the same as in Section 2. First, we introduce the ϕ-mixing condition. Let $\mathcal{F}_0^k$ and $\mathcal{F}_{k+n}^\infty$ denote the σ-field generated by $\{X_t, t \le k\}$ and $\{X_t, t \ge k+n\}$, respectively. If $A \in \mathcal{F}_0^k$ and $B \in \mathcal{F}_{k+n}^\infty$, then for all k and $n \ge 1$,

$$|\Pr(B|A) - \Pr(B)| \le \phi(n), \phi(n) \ge 0,$$

where $\phi(n)$ is non-increasing w.r.t. n and $\lim_{n\to\infty} \phi(n) = 0$. Sen [28] studied the Bahadur representation of sample quantile for ϕ-mixing sequences. Denote

$$A_0(\phi) = \sum_{n=1}^{\infty} (\phi(n))^{\frac{1}{2}},$$

and

$$v_i = \mathbb{E}\left[\mathbf{1}\{q_\alpha \ge X_0\}\mathbf{1}\{q_\alpha \ge X_i\}\right], i = 0, 1, 2, \ldots.$$

Let $v^2 = v_0 + 2\sum_{i=1}^{\infty} v_i$. By Sen [28], $A_0(\phi) < \infty$ implies $v < \infty$. The main theorem in Sen [28] shows if $A_0(\phi) < \infty$, $0 < \partial_1 F(x;\theta) < \infty$ and $\partial_1^2 F(x;\theta)$ exists and is bounded,

then

$$\frac{\hat{q}_\alpha^n - q_\alpha}{v/(\partial_1 F(q_\alpha;\theta)n^{\frac{1}{2}})} \xrightarrow{d} \mathcal{N}(0,1), \text{ as } n \to \infty.$$

Therefore, A2 can be replaced by A2$''$:

A2$''$. $A_0(\phi) < \infty$.

Corollary 3. *Suppose that* $\sup_n \mathbb{E}[I_n^2] < \infty$ *and* $F_{(\lceil \alpha n \rceil)}(x;\theta)$ *satisfies conditions (i) and (ii) in Lemma 2. If A1 and A2*$''$ *are satisfied, and* $dX/d\theta$ *exists, then* $\mathbb{E}[I_n] \to q'_\alpha$ *as* $n \to \infty$.

Now we take all the batch sizes to be the same, i.e., $m_i \equiv m$ for all i, and the batched estimator is given by $\hat{q}'^m_{\alpha,k} = 1/k \sum_{i=1}^k I_{i,m}$. Suppose that $\{(X_t, dX_t/d\theta), t \geq 0\}$ satisfies the ϕ-mixing condition, the same condition required in Liu and Hong [23]. Under certain conditions, $\{X_t, t \geq 0\}$ ϕ-mixing implies that $\{dX_t/d\theta, t \geq 0\}$ is ϕ-mixing. For example, if there is a function f satisfying some mild conditions, e.g., it is a linear function, such that $\{dX_t/d\theta = f(X_t), t \geq 0\}$, then $dX_t/d\theta$ satisfies a ϕ-mixing condition. When $m \to \infty$ for all $i \neq j$, $Cov(I_{i,m}, I_{j,m}) \to 0$. Then

$$Var(\hat{q}'^m_{\alpha,k}) = \frac{1}{k^2}\left(\sum_{i=1}^k Var(I_{i,m}) + 2\sum_{i\neq j} Cov(I_{i,m}, I_{j,m})\right) \to 0 \text{ as } m, k \to \infty,$$

and by Theorem 3, we obtain the following corollary.

Corollary 4. *Suppose that for each* i, $\sup_m \mathbb{E}[I_{i,m}^2] < \infty$, *and the sequence* $\{(X_t, dX_t/d\theta), t \geq 0\}$ *satisfies the* ϕ*-mixing condition. If A1 and A2*$''$ *are satisfied, then* $\hat{q}'^m_{\alpha,k} \to q'_\alpha$ *in probability as* $m \to \infty$ *and* $k \to \infty$.

The sequence $I_{i,m}, i = 1, 2, \ldots$ is derived from $dX_t/d\theta$, so it satisfies the ϕ-mixing condition. By Deo [6], if A2$''$ holds, $\sup_m \mathbb{E}[I_{i,m}^2] < \infty$ for each i, and $\sigma^2 = \mathbb{E}[I_{i,m}^2] + 2\sum_{i\neq j}^\infty \mathbb{E}[I_{i,m}I_{j,m}] - (\mathbb{E}[I_{i,m}])^2 > 0$, then

$$\frac{\hat{q'}^m_{\alpha,k} - \mathbb{E}\left[\hat{q'}^m_{\alpha,k}\right]}{\sigma/\sqrt{k}} \xrightarrow{d} \mathcal{N}(0,1) \text{ as } k \to \infty.$$

Therefore, similar to regenerative processes, we have a central limit theorem for the batched estimator, with the proof being nearly identical as for regenerative processes.

Corollary 5. *Assume* $\varphi(x) = -\partial_2 F(x;\theta)/\partial_1 F(x;\theta)$ *is twice differentiable and* $|\varphi''(x)| < K$, $\mathbb{E}[\hat{q}_\alpha^m - q_\alpha] = O(m^{-\gamma_1})$ *and* $\mathbb{E}[(\hat{q}_\alpha^m - q_\alpha)^2] = O(m^{-\gamma_2})$, *where* $\gamma = \min\{\gamma_1, \gamma_2\}$.

Under the assumptions in Corollary 4, if $\sigma^2 = \mathbb{E}[I_{i,m}^2] + 2\sum_{i\neq j}^{\infty} \mathbb{E}[I_{i,m}I_{j,m}] - (\mathbb{E}[I_m]) > 0$*, then*

$$\frac{\hat{q'}_{\alpha,k}^{m} - q'_{\alpha}}{\sigma/\sqrt{k}} \xrightarrow{d} \mathcal{N}(0,1) \quad as \quad m \to \infty,\ k \to \infty \ and \ \frac{k^{\frac{1}{2}}}{m^{\gamma}} \to 0.$$

In this subsection, we used ϕ-mixing processes as an example. However, the framework can be adapted to other dependent processes as long as the appropriate quantile estimator converges to the true quantile.

4. Numerical Experiments

In this section, we illustrate the performance of the estimators on two numerical examples: the G/G/1 queue and the discrete Ornstein-Uhlenbeck process, i.e., AR(1) process. For the G/G/1 queue, we consider the waiting (queueing) time, which is a regenerative process, and compare the REG estimator with the DET estimator considered in Hong [18]. Then, the discrete Ornstein-Uhlenbeck process, which is a m-dependent sequence, is considered. Under some mild conditions, the discrete Ornstein-Uhlenbeck process is stationary Gaussian. By Bradley [5], ϕ-mixing is equivalent to m-dependence for stationary Gaussian sequence.

4.1. G/G/1 Queue

Let W_i denote the waiting time of the ith customer in a FCFS G/G/1 queue, which by the Lindley equation, satisfies the recursion $W_{i+1} = (W_i + X_i)^+$, where $X_i = S_i - A_i$, S_i is the service time of the ith customer, and A_i is interarrival time between the ith and $(i+1)$st customer. Under appropriate stability conditions, W_i is a regenerative process w.r.t. indices of customers who initiate busy periods. The IPA estimator of W_i is given by direct differentiation:

$$\frac{dW_{i+1}}{d\theta} = \left(\frac{dW_i}{d\theta} + \frac{dX_i}{d\theta}\right) \mathbb{1}_{\{W_i + X_i \geq 0\}}.$$

Similar to Section 8 in Hong [18], we consider the setting where A_i and S_i are i.i.d. exponentially distributed random variables with means μ_A and μ_S, respectively. When $\mu_A > \mu_S$, the queue is stable, and the steady-state waiting time W is exponentially distributed with rate $(1/\mu_S) - (1/\mu_A)$. Then

$$\frac{\partial q_\alpha}{\partial \mu_A} = \left(\frac{\mu_S}{\mu_A - \mu_S}\right)^2 \log(1-\alpha),$$

$$\frac{\partial q_\alpha}{\partial \mu_A} = -\left(\frac{\mu_A}{\mu_A - \mu_S}\right)^2 \log(1-\alpha).$$

First, we compare the two batched estimators, i.e., DET and REG, for $\mu_A = 10$ and $\mu_S = 8$. The simulation experiments fixed the total number of regenerative cycles at $\tilde{n} = 10000$ per replication, divided into 100 batches ($\tilde{k} = 100$) of 100 cycles ($\tilde{m} = 100$). Since DET uses fixed batch sizes rather than cycles, the total number of customers was rounded up to allow equal batch sizes. For example, if 10000 cycles generates 50112 customers served in a replication, the simulation replication is continued until 50200 customers are served, so that DET uses 100 batches of 502 customers. All the results are based on 1000 independent replications. The absolute bias of the two batched estimators are plotted against different α in Fig.1, the mean square error (MSE) in Fig.2, and the coverage probabilities in Fig.3. In both Fig.1 and Fig.2, the

FIGURE 1: Absolute bias of quantile sensitivities w.r.t. μ_A (left panel) and μ_S (right panel)

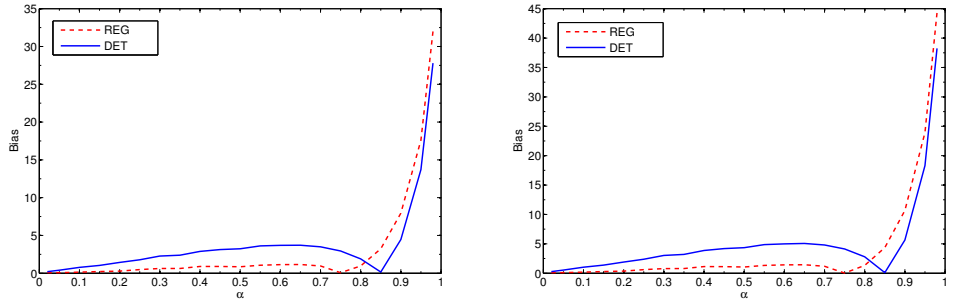
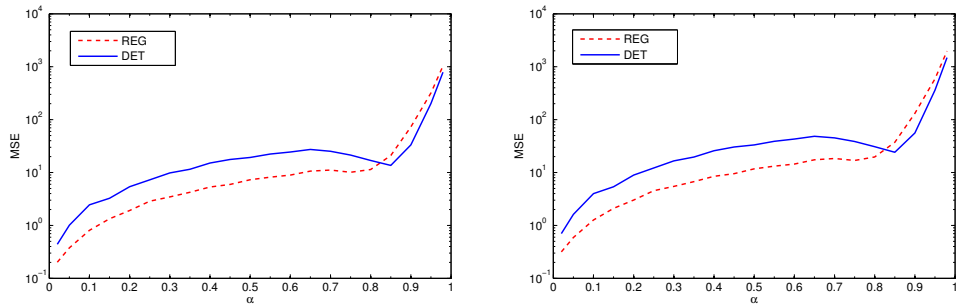
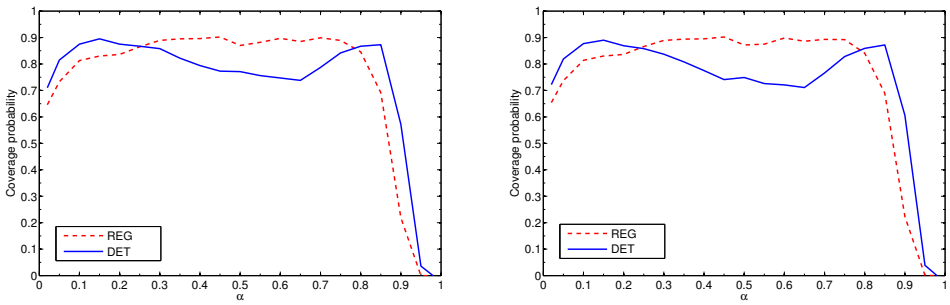


FIGURE 2: MSE of quantile sensitivities w.r.t. μ_A (left panel) and μ_S (right panel)



performance of REG is generally better than DET for $\alpha < 0.8$. When $\alpha > 0.8$, both absolute bias and MSE increase significantly, more so for REG than for DET. Fig.3 shows the estimated coverage probabilities of the 90% confidence interval. When α is

FIGURE 3: Coverage probabilities of quantile sensitivities w.r.t. μ_A (left panel) and μ_S (right panel)



far from 0 or 1, the coverage probabilities are close to 0.9. When α is near 0 or 1, the coverage probabilities decrease significantly. Fig.3 indicates that the REG coverage probabilities are more stable.

Next, we consider the tradeoff between the number of cycles in each batch and the number of batches for the REG. Denote the number of cycles in each batch by $\tilde{m} = \tilde{n}^{\delta}$ and the number of batches by $\tilde{k} = \tilde{n}^{1-\delta}$. Three values for the total number of cycles were tested: $\tilde{n} = 1000, 10000, 100000$, as a function of δ for different values of the traffic intensity $\rho = \mu_S/\mu_A$ and the quantile level α. From Fig.4, Fig.5 and Fig.6, the MSE

FIGURE 4: MSE of quantile sensitivities w.r.t. μ_A; traffic intensity $\rho = 0.8$; $\alpha = 0.2$, 0.4 and 0.8 in the left, middle, and right panels, respectively.

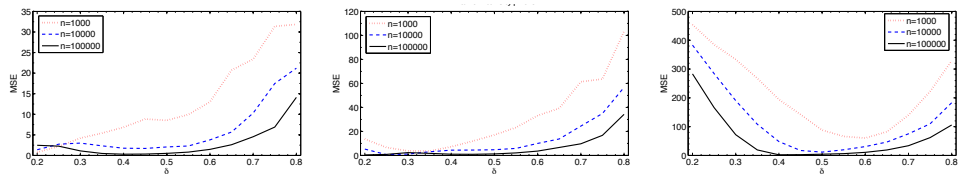
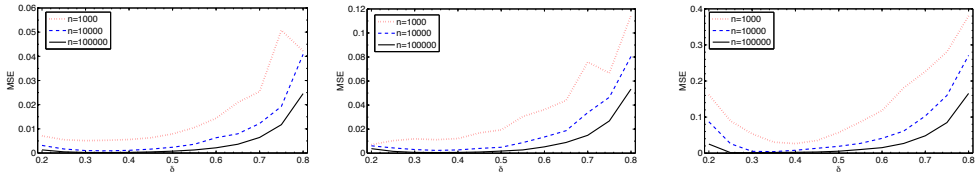
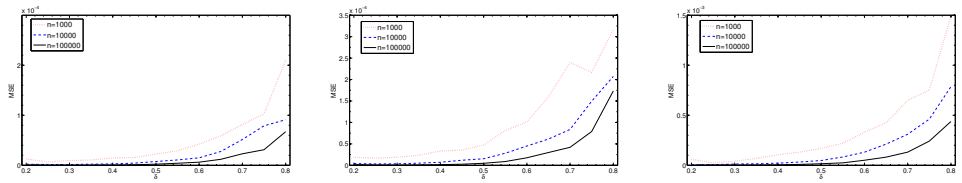


FIGURE 5: MSE of quantile sensitivities w.r.t. μ_A; traffic intensity $\rho = 0.4$; $\alpha = 0.2$, 0.4 and 0.8 in the left, middle, and right panels, respectively.



increases with δ in most situations. However, if δ is small, i.e, the number of cycles in each batch is small, the bias is large, and the coverage probabilities are not close to the

FIGURE 6: MSE of quantile sensitivities w.r.t. μ_A; traffic intensity $\rho = 0.1$; $\alpha = 0.2$, 0.4 and 0.8 in the left, middle, and right panels, respectively.



theoretical probability. For $\tilde{n} = 10000$, Table 1 gives the 90% coverage probabilities and MSE for different combinations of $\tilde{m}$ and $\tilde{k}$.

TABLE 1: Coverage probabilities (CP) and the MSE with different traffic intensities and batch sizes ($\tilde{n} = 10000$)

		$\rho = 0.8$		$\rho = 0.4$		$\rho = 0.2$	
$\tilde{m}$	$\tilde{k}$	CP	MSE	CP	MSE	CP	MSE
20	500	0.001	38.9	0.900	0.0109	0.843	0.0001
50	200	0.791	5.08	0.897	0.0316	0.886	0.0003
100	100	0.895	10.4	0.869	0.0557	0.881	0.0005
200	50	0.863	19.2	0.860	0.102	0.825	0.0010
500	20	0.825	39.7	0.830	0.248	0.775	0.0023

4.2. Ornstein-Uhlenbeck Process

The Ornstein-Uhlenbeck (OU) process is given by the stochastic differential equation

$$dX_t = \theta(\mu - X_t)dt + \sigma dW_t, \ t \geq 0,$$

where $\theta > 0$, $\mu > 0$ and $\sigma > 0$ are parameters, and W_t denotes the standard Wiener process. We consider the discrete-time OU process, i.e., AR(1) process:

$$X_{t+1} = X_t + \theta(\mu - X_t)\Delta t + \sigma\sqrt{\Delta t}Z_t, \ t = 0, 1, \ldots,$$

where Δt is the discretization interval and $\{Z_t, t = 0, 1, \ldots\}$ are i.i.d. standard normal random variables. The limiting distribution of X_t is normally distributed with mean μ and variance $\sigma^2/(2\theta - \theta^2\Delta t)$, so we can obtain an analytical expression for the quantile.

Let $\theta = 0.2$, $\mu = 0.1$, $\sigma = 0.1$, $\Delta t = 0.05$, and total sequence length $N = 10^6$ with batch size $m = 10^4$ and total number of batches $k = 100$. Again, all results are based on 1000 replications. Fig.7 indicates that the batched estimator MSE is quite small and the estimated probability coverage is close to the theoretical value of 90%. Similar to the G/G/1 queue, when the quantile level α tends to 0 or 1, the MSE increases substantially and the coverage probability deteriorates. Fig. 8 shows the bias and MSE as the total sequence length N increases for fixed $k = 100$ and $\alpha = 0.8$.

FIGURE 7: MSE and 90% coverage probability of quantile sensitivity for the OU example ($N = 10^6$, $m = 10^4$, $k = 100$)

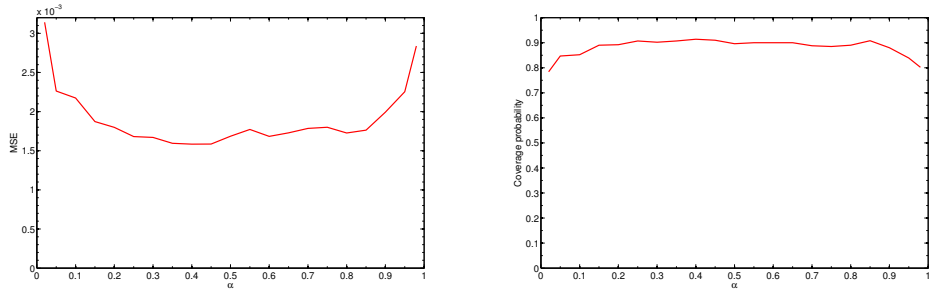
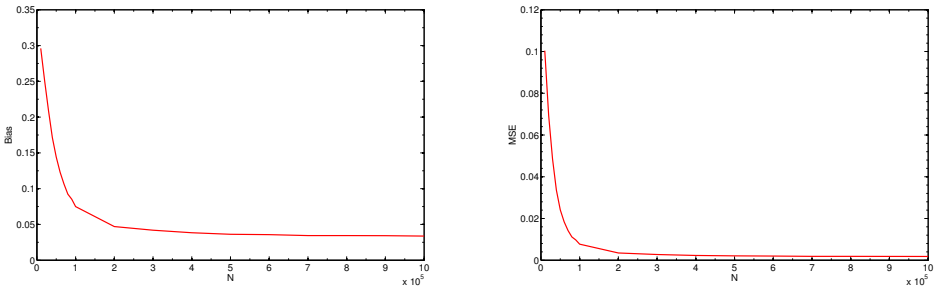


FIGURE 8: Bias and MSE of quantile sensitivity for the OU example ($K = 100$, $\alpha = 0.8$)



5. Conclusion

In this paper, we estimate the quantile sensitivities of dependent sequences via IPA, and construct a batched estimator which is asymptotically unbiased and weakly consistent. Two widely used processes, regenerative processes and ϕ-mixing processes, are studied. For regenerative processes, we improve the batched estimator based on regenerative cycles, and a central limit theorem is given. For ϕ-mixing processes, we also provide a central limit theorem, and the framework can be used for other classes

of dependent processes with quantiles that can be estimated by the corresponding sequence of order statistics. Two numerical examples, the G/G/1 queue and discrete OU process, i.e., AR(1) process, are given to illustrate the effectiveness of the estimators. For both numerical examples, when α is close to 0 or 1, the performance of the estimators deteriorates, as the bias becomes significantly large. Bias reduction methods such as jackknifing can improve the performance of the estimators (Jiang, Fu and Xu [22]).

Acknowledgements

This work was supported in part by the National Science Foundation (NSF) under Grants CMMI-0856256, CMMI-1362303, CMMI-1434419, and EECS-0901543, by the Air Force of Scientific Research (AFOSR) under Grant FA9550-15-10050, and by the National Natural Science Foundation of China (Project 11171256).

References

[1] Alexopoulos, C., Wilson, J. R. A new perspective on batched quantile estimation. In Proceedings of the 2012 Winter Simulation Conference, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher. 190–200. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

[2] Asmussen, S. 2003. *Applied Probability and Queues.* Wiley, New York.

[3] Babu, G. J., Singh, K. 1978. On deviations between empirical and quantile processes for mixing random variables. *J. Multivariate Anal.* **8**, 532-549.

[4] Bekki, J. M., Fowler, J. W., Mackulak, G. T., Nelson, B.L. 2010. Indirect cycle time quantile estimation using the Cornish-Fisher expansion. *IIE Trans.* **42**, 31-44.

[5] Bradley, R. C. 2005. Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.* **2**, 107-144.

[6] Deo, C. M. 1975. A functional central limit theorem for stationary random fields. *Ann. Probab.* **3**(4), 708-715.

[7] Doukhan, P. 1995. *Mixing: Properties and Examples.* Springer-Verlag, New York.

[8] Fu, M. C. 2006. Gradient estimation. S. G. Henderson, B. L. Nelson, eds. *Handbooks in Operations Research and Management Science: Simulation*, Chap. 19. Elsevier, Amsterdam, 575-616.

[9] Fu, M. C., Hong, L. J., Hu, J. Q. 2010. Conditional Monte Carlo estimation of quantile sensitivities. *Management Sci.* **55**(12), 2019-2027.

[10] Gelinas, G., Martel, A., Lefrancois, P. 1995. SOS: A quantile estimation procedure for dynamic lot-sizing problems. *J. Oper. Res. Soc.* **46**, 1337-1351.

[11] Ghosh, S., Pasupathy, R. Low-storage online estimator for quantile and densities. In proceedings of the 2013 Winter Simulation Conference, edited by R. Pasupathy, S. H. Kim, A. Tolk, R. Hill, and M. E. Kuhl. 778–789. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

[12] Glasserman, P. 1993. Regenerative derivatives of regenerative sequences. *Adv. Appl. Probab.* **25**(1), 116-139.

[13] Glasserman, P., Heidelberger P., Shahabuddin, P. 2000. Variance reduction techniques for estimating Value-at-Risk. *Management Sci.* **46**(1), 1349-1364.

[14] Heidelberger, P., Lewis, P. A. W. 1984. Quantile estimation in dependent sequences. *Oper. Res.* **32**(1), 185-209.

[15] Heidergott, B., Volk-Makarewicz, W. 2010. Sensitivity analysis of quantiles. *AENORM*, 2012, 26-31. http://www.aenorm.nl/editions/?edt=3&art=18

[16] Hesse, C. H. 1990. A Bahadur-type representation for empirical quantiles of a large class of stationary, possibly infinite-variance, linear processes. *Ann. Statist.* **18**, 1188-1202.

[17] Ho, H. C., Hsing, T. 1996. On the asymptotic expansion of the empirical process of long-memory moving averages. *Ann. Statist.* **24**, 992-1024.

[18] Hong, L. J. 2009. Estimating quantile sensitivities. *Oper. Res.* **57**(1), 118-130.

[19] Iglehart, D. L. 1972. Simulating stable stochastic systems, VI: Quantile estimation. *J. ACM.* **23**, 347-360.

[20] Jain, R., Chlamtac, I. 1985. The P^2 algorithm for dynamic calculation of quantiles and histograms without storing observations. *Commun. ACM.* **28**(10), 1076-1085.

[21] Jiang, G., Fu, M. C. 2015. Technical note: On estimating quantile sensitivities via infinitesimal perturbation analysis. *Oper. Res.* **63**(2), 435-441.

[22] Jiang, G., Fu, M. C., Xu, C. 2014. Bias reduction in estimating quantile sensitivities. In Proceedings of the 19th IFAC World Congress, edited by E. Boje, X. Xia. 10463-10468. Austria: International Federation of Automatic Control.

[23] Liu, G. W., Hong, L. J. 2009. Kernel estimation of quantile sensitivities. *Naval Res. Logist.* **56**(6), 511-525.

[24] Meketon, M.S., Heidelberger, P. 1982. A renewal theoretic approach to bias reduction in regenerative simulations. *Management Sci.* **28**, 173-181.

[25] Nair, N. U., Sankaran, P. G. 2008. Quantile based reliability analysis. *Commun. Statist. Theory Meth.* **38**, 222–232.

[26] Seila, A. F. 1982. A batching approach to quantile estimation in regenerative simulations. *Management Sci.* **28**, 573-581.

[27] Sen, P. K. 1968. Asymptotic normality of sample quantiles for m-dependent processes. *Ann. Math. Statist.* **39**, 1724-1730.

[28] Sen, P. K. 1972. On the Bahadur's representation of sample quantiles for sequences of ϕ-mixing random variable. *J. Multivariate Anal.* **2**, 77-95.

[29] Serfling, R. J. 1980. *Approximation Theorems of Mathematical Statistics.* Wiley, New York.

[30] Serfozo, R. 2009. *Basics of Applied Stochastic Process.* Springer-Verlag, Berlin.

[31] Steiger, N. M., Wilson J. R. 2001. Convergence properties of the batch means method for simulation output analysis. *INFORMS J. Comput.* **13**(4), 277-293.

[32] Suri, R., Zazanis, M. A. 1988. Perturbation analysis gives strongly consistent sensitivity estimates for the $M/G/1$ queue. *Management Sci.* **34**, 39-64.

[33] Wu, W. 2005. On the Bahadur representation of sample quantiles for dependent sequences. *Ann. Statist.* **33**(4), 1934-1963.