J. Appl. Prob. **53,** 715–732 (2016)
*doi:*10.1017/jpr.2016.36
© *Applied Probability Trust* 2016

QUANTILE SENSITIVITY ESTIMATION FOR DEPENDENT SEQUENCES

GUANGXIN JIANG,* *City University of Hong Kong*
MICHAEL C. FU,** *University of Maryland*

Abstract

In this paper we estimate quantile sensitivities for dependent sequences via infinitesimal perturbation analysis, 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 process; 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 measures such as mean and variance when characterizing the extreme properties of a stochastic system. For complex stochastic systems, simulation is commonly used to estimate performance. Serfling (1980) provides an overview of quantile estimation for independent and identically distributed (i.i.d.) data. For dependent sequences, Sen (1972) provides a representation of the quantile by order statistics and a confidence interval for the estimator. Other methods include interpolation (Iglehart (1976)), batching (Seila (1982)), and maximum transformation (Heidelberger and Lewis (1984)).

Our work addresses quantile sensitivity estimation. Previous work goes back to Hong (2009), who introduced infinitesimal perturbation analysis (IPA) to estimate the quantile sensitivity based on analysis of a probability sensitivity. Jiang and Fu (2015) presented an alternative more direct derivation of the IPA estimators for both batched and unbatched estimators. Fu *et al.* (2009) applied smoothed perturbation analysis 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 (2009)) and weak derivative estimator (cf. Volk-Makarewicz and Heidergott (2012)) 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

Received 28 April 2015; revision received 22 September 2015.
* Postal address: Department of Economics and Finance, City University of Hong Kong, Kowloon, Hong Kong. Email address: guajiang@cityu.edu.hk
** Postal address: Robert H. Smith School of Business and Institute for Systems Research, University of Maryland, College Park, MD 20742, USA.

https://doi.org/10.1017/jpr.2016.36 Published online by Cambridge University Press

and queueing theory, e.g. analyzing the steady-state waiting time. More generally, a strong mixing condition is introduced to capture asymptotic properties of dependent random variables (Bradley (1975)), and a quantile sensitivity estimator is derived for this setting.

The most closely related works are Hong (2009), who 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 (2009), who 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. We conclude with Section 5.

2. IPA quantile sensitivity estimation

IPA has been used for estimating quantile sensitivities of i.i.d. sequences; see, e.g. Hong (2009) and Jiang and Fu (2015). In the setting of these two papers, the quantile sensitivity estimator is obtained from the IPA derivatives of the corresponding order statistics. Here, 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, so 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 nonstationary processes. We provide a batched estimator that 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 (CDF) $F(x;\theta) = \mathbb{P}\{X_t \le x\}$, 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 $\mathbb{P}\{X \le q_\alpha\} \ge \alpha$ and $\mathbb{P}\{X \ge q_\alpha\} \ge 1-\alpha$. If X_t is a continuous random variable, then $\mathbb{P}\{X_t \le q_\alpha\} = F(q_\alpha;\theta) = \alpha$, where $F(\cdot,\theta)$ has a continuous support. We make the following assumption.

Assumption 1. (i) *In a neighborhood of $x = q_\alpha$, $F(x;\theta)$ is continuously differentiable with respect to both arguments.*

(ii) *The density $\partial_1 F(x;\theta)$ is strictly positive for all $\theta \in \Theta$.*

Let $q'_\alpha := \mathrm{d}q_\alpha/\mathrm{d}\theta$ denote the α-quantile sensitivity of X_t with respect to 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 CDFs. 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

https://doi.org/10.1017/jpr.2016.36 Published online by Cambridge University Press

CDFs assumed known. Since $\alpha = F(q_\alpha; \theta)$, differentiating both sides with respect to 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 with respect to the ith argument of F, and Assumption 1 guarantees the right-hand side of (1) exists.

If $F(x; \theta)$ is known, we can first estimate the quantile q_α, then substitute into (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. We need the following additional assumption.

Assumption 2. *We have* $\hat{q}^n_\alpha \xrightarrow{\mathrm{D}} q_\alpha$ *as* $n \to \infty$, *where* '$\xrightarrow{\mathrm{D}}$' *denotes convergence in distribution.*

Remark 1. Assumption 1 is a typical assumption used in quantile sensitivity estimation (cf. Hong (2009) and Jiang and Fu (2015)). For i.i.d. sequences, $\hat{q}^n_\alpha$ converges to q_α with probability 1, which is the key step for proving the asymptotic unbiasedness of the quantile sensitivity estimator (Jiang and Fu (2015)). Assumption 2 holds for many classes of dependent sequences. For example, in regenerative processes, if the derivative of the density of the steady-state CDF $\partial_1^2 F(x; \theta)$ exists and is bounded by M in a neighbourhood of q_α, then Assumption 2 holds (Iglehart (1976)). We will discuss this case in detail in the next section.

Let $\mathrm{d}X/\mathrm{d}\theta$ denote the ordinary IPA estimator of X with respect to θ. For example, $X = A + \theta S$, where A is a random variable with parameter θ and S does not contain θ. If $\mathrm{d}A/\mathrm{d}\theta$ is well defined then $\mathrm{d}X/\mathrm{d}\theta = \mathrm{d}A/\mathrm{d}\theta + S$.

For the sequence (2), the corresponding derivatives with respect to θ are given by

$$\frac{\mathrm{d}X_{(1)}}{\mathrm{d}\theta}, \frac{\mathrm{d}X_{(2)}}{\mathrm{d}\theta}, \ldots, \frac{\mathrm{d}X_{(\lceil \alpha n \rceil)}}{\mathrm{d}\theta}, \ldots, \frac{\mathrm{d}X_{(n)}}{\mathrm{d}\theta},$$

and we define the IPA quantile sensitivity estimator as

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

Jiang and Fu (2015, Theorem 1) introduced their theorem in order to establish asymptotic unbiasedness of the IPA estimator for dependent sequences.

Lemma 1. (Jiang and Fu (2015).) *Assume that* X *is differentiable with respect to* $\theta \in \Theta$ *with probability* 1, *and Assumption 1 is satisfied at the point* y. *Then,*

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

where $F(\cdot; \theta)$ *is the CDF of* X.

https://doi.org/10.1017/jpr.2016.36 Published online by Cambridge University Press

Equation (5) connects the IPA estimator $\mathrm{d}X/\mathrm{d}\theta$ with the CDF of X through conditional expectation. By Fu (2006), if this parameter is a *location*, *scale,* or *generalized scale* parameter, the conditional expectation can be removed. For example, consider the generalized scale parameter $\mathrm{d}X/\mathrm{d}\theta = (X - \bar{\theta})/\theta$, then $\mathbb{E}[\mathrm{d}X/\mathrm{d}\theta \mid X] = \mathbb{E}[(X - \bar{\theta})/\theta \mid X] = (X - \bar{\theta})/\theta$. Lemma 1 can be regarded as an extension of the classical results of the IPA estimator in Suri and Zazanis (1988), where the random variable is simple and the derivative can be represented by the quotient of partial derivatives of CDFs. 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 $\mathrm{d}X_t/\mathrm{d}\theta = X_1$, 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. The proof closely follows the i.i.d. case in Jiang and Fu (2015), since the process is stationary, and is given in Appendix A.

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

2.2. The IPA estimator for quantile sensitivity in nonstationary 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 $X_t \xrightarrow{\mathrm{D}} X$ as $t \to \infty$. Stationary processes are the special case where $X_t \overset{\mathrm{D}}{=} X$, with equality in distribution denoted by '$\overset{\mathrm{D}}{=}$'. Let $F(x; \theta)$ denote the CDF 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. $\mathrm{d}q_\alpha/\mathrm{d}\theta$. Let $\{F_t(x; \theta),\ t = 0, 1, \ldots\}$ denote the CDF 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 CDFs, Assumption 1 is replaced by the following assumption.

Assumption 3. *In a neighbourhood of* $x = q_\alpha$, $\{F_t(x; \theta),\ t = 0, 1, 2, \ldots\}$ *and* $F(x; \theta)$ *are continuously differentiable with respect to 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 CDF, the conditional expectation of* $\mathrm{d}X_{(t)}/\mathrm{d}\theta$ *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 Assumption 3 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 *and* b *can take value* $\pm\infty$.

Then $\partial_1 F(x, \theta) = \lim_{t\to\infty} \partial_1 F_t(x, \theta)$ *and* $\partial_2 F(x, \theta) = \lim_{i\to\infty} \partial_2 F_t(x, \theta)$.

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 $[a, b]$ with respect to x for fixed θ. Then, it remains to show that $\varphi(x, \theta) = \partial_1 F(x, \theta)$. Because

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

https://doi.org/10.1017/jpr.2016.36 Published online by Cambridge University Press

then

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

and the second equality holds because of the uniform convergence of the original sequence $F_t(x,\theta)$. Differentiating both sides with respect to 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 $X_t \xrightarrow{\mathrm{D}} X$, by Assumption 3, $F(x;\theta)$ is continuous with respect to x,

$$\sup_{x\in[a,b]} |F_t(x;\theta) - F(x;\theta)| \to 0,$$

i.e. $F_t(x;\theta)$ converges uniformly to $F(x;\theta)$ with respect to x. Moreover, when Θ is a compact set, pointwise convergence is equivalent to uniform convergence.

For ease of notation, let $F_{(t)}$ denote the CDF 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 CDFs $\{F_t,\ t = 0, 1, \ldots, 5\}$, which satisfies $x_0 < x_5 < x_3 < x_2 < x_1 < x_4$. So for this realization, the fourth-order statistic observation $x_{(4)}$ corresponds to x_2, which is sampled from X_2 with CDF 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 that* $F_{(\lceil \alpha n\rceil)}(x;\theta)$ *satisfies Lemmas 2(i) and 2(ii). If Assumptions 2 and 3 hold, and* $\mathrm{d}X/\mathrm{d}\theta$ *exists, then* $\mathbb{E}[I_n] \to q'_\alpha$ *as* $n \to \infty$.

Proof. As in Theorem 1, the quantile sensitivity of X can be obtained by

$$q'_\alpha = \mathbb{E}\left[\frac{\mathrm{d}X}{\mathrm{d}\theta} \,\middle|\, X = q_\alpha\right] = -\frac{\partial_2 F(q_\alpha;\theta)}{\partial_1 F(q_\alpha;\theta)}. \tag{6}$$

Equation (6) represents a conditional expectation of the derivative of the random variable X as a quotient of partial derivatives of its CDF, i.e. $-\partial_2 F(X;\theta)/\partial_1 F(X;\theta)$. Analogously, for the random variables $\{X_t,\ t = 0, 1, 2, \ldots\}$,

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

Now consider the expectation of $\mathrm{d}X_{(\lceil \alpha n\rceil)}/\mathrm{d}\theta$. 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 Assumption 1 replaced by Assumption 3, so

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

Then, similar to Jiang and Fu (2015, Equation (11)) (or (A.1) in Appendix A),

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

https://doi.org/10.1017/jpr.2016.36 Published online by Cambridge University Press

Now we show that $|\partial_i F_{(\lceil \alpha n \rceil)}(\hat{q}_\alpha^n; \theta) - \partial_i F(q_\alpha, \theta)| \xrightarrow{\text{D}} 0$ for $i = 1, 2$ as $n \to \infty$. Thus,

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

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 $\varepsilon > 0$, there exists N' such that for all $n \ge N'$, $|\partial_i F_{(\lceil \alpha n \rceil)}(\hat{q}_\alpha^n; \theta) - \partial_i F(\hat{q}_\alpha^n, \theta)| < \varepsilon/2$ almost surely for all $\theta \in \Theta$. By Assumption 3, $F(x, \theta)$ is continuously differentiable with respect to both arguments, i.e. $\partial_i F(x, \theta)$ is continuous with respect to both x and θ for $i = 1, 2$. Then by Assumption 2, for any $\varepsilon > 0$, there exists N'' such that, for all $n \ge N''$, $|\partial_i F(\hat{q}_\alpha^n, \theta) - \partial_i F(q_\alpha, \theta)| < \varepsilon/2$ in distribution. So it is easy to prove that, for $i = 1, 2$,

$$|\partial_i F_{(\lceil \alpha n \rceil)}(\hat{q}_\alpha^n; \theta) - \partial_i F(q_\alpha, \theta)| \xrightarrow{\text{D}} 0 \quad \text{as } n \to \infty,$$

i.e. for $i = 1, 2$, $\partial_i F_{(\lceil \alpha n \rceil)}(\hat{q}_\alpha^n; \theta) \xrightarrow{\text{D}} \partial_i F(q_\alpha, \theta)$ as $n \to \infty$. By Assumption 3,

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

Similar to the proof of Jiang and Fu (2015, Lemma 2) (or see also (A.2) in Appendix A),

$$\frac{\partial_2 F_{(\lceil \alpha n \rceil)}(x; \theta)}{\partial_1 F_{(\lceil \alpha n \rceil)}(x; \theta)}$$

is uniformly integrable, so $\mathbb{E}[I_n] \to q'_\alpha$ as $n \to \infty$. □

Remark 3. (i) If all $\{F_t(x; \theta),\ t = 0, 1, \ldots\}$ satisfy Lemmas 2(i) and 2(ii), then $F_{(\lceil \alpha n \rceil)}(x; \theta)$ satisfies the conditions automatically.

(ii) Although $X_t \xrightarrow{\text{D}} X$, i.e. $F_t(x; \theta) \to F(x; \theta)$ as $t \to \infty$, it does not directly imply that $F_{(\lceil \alpha n \rceil)}(x; \theta) \to F(x; \theta)$ as $n \to \infty$. However, after an appropriate 'warm-up' period, all the CDFs can be regarded as the same as the steady-state CDF $F(x; \theta)$. 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 CDFs of X_t in the ith batch ($i \ge 2$) converge to the steady-state CDF $F(x; \theta)$. 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 (2009) and Jiang and Fu (2015)). 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, a widely used simulation method for analyzing the steady-state mean of a stochastic process (Steiger and Wilson (2001)). By Theorems 1 and 2, we see that 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

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

https://doi.org/10.1017/jpr.2016.36 Published online by Cambridge University Press

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

$$\hat{q}'^{m}_{\alpha,k} = \frac{1}{k}\sum_{i=1}^{k} I_{i,m_i}. \tag{7}$$

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 $\mathrm{var}(\hat{q}'^{m}_{\alpha,k}) = 1/k^2 \sum_{i=1}^{k} \mathrm{var}(I_{i,m_i})$. Suppose that $\max_i \mathrm{var}(I_{i,m_i}) < \infty$, when $k \to \infty$, $\mathrm{var}(\hat{q}'^{m}_{\alpha,k}) < 1/k \max_i \mathrm{var}(I_{i,m_i}) \to 0$, 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, $\sup_{m_i} \mathbb{E}[I^2_{i,m_i}] < \infty$ and $\mathrm{var}(\hat{q}'^{m}_{\alpha,k}) \to 0$ as $k \to \infty$. If Assumptions 1 and 2 hold, and $\mathrm{d}X/\mathrm{d}\theta$ exists, then $\hat{q}'^{m}_{\alpha,k} \xrightarrow{\mathbb{P}} q'_\alpha$ as $m \to \infty$ and $k \to \infty$, where '$\xrightarrow{\mathbb{P}}$' denotes convergence in probability.*

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

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

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

$$\mathbb{P}\{|\hat{q'}^{m}_{\alpha,k} - \mathbb{E}[\hat{q}'^{m}_{\alpha,k}]| \geq \varepsilon\} \leq \frac{\mathrm{var}[\hat{q'}^{m}_{\alpha,k}]}{\varepsilon^2}.$$

Since $\sup_{m_i} \mathbb{E}[I^2_{i,m_i}] < \infty$ for each i, when $k \to \infty$, $\mathrm{var}(\hat{q}'^{m}_{\alpha,k}) \to 0$, 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, $\hat{q}'^{m}_{\alpha,k} \xrightarrow{\mathbb{P}} \mathbb{E}[\hat{q}'^{m}_{\alpha,k}]$. Then, let $k \to \infty$, and it is easy to prove that $\lim_{k\to\infty} \lim_{m\to\infty} \hat{q}'^{m}_{\alpha,k} = q'_\alpha$ in probability. □

For nonstationary processes in Section 2.2, we can still prove that the batched estimator (7) is weakly consistent. According to Remark 3, if $m_1 \to \infty$, the CDFs from the second batch onwards follow the steady-state CDF, 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, $\sup_{m_i} \mathbb{E}[I^2_{i,m_i}] < \infty$ and $\mathrm{var}(\hat{q}'^{m}_{\alpha,k}) \to 0$ as $k \to \infty$. Additionally, $\{F_t(x;\theta),\ t = 0, 1, 2, \ldots\}$ satisfies Lemma 2(ii). If Assumptions 2 and 3 hold, and $\mathrm{d}X/\mathrm{d}\theta$ exists, then $\hat{q}'^{m}_{\alpha,k} \xrightarrow{\mathbb{P}} q'_\alpha$ as $m \to \infty$, $k \to \infty$.*

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

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

https://doi.org/10.1017/jpr.2016.36 Published online by Cambridge University Press

then

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

When $m \to \infty$, by Remark 3, we know that $\mathbb{E}[I_{i,m_i}] \to q'_\alpha$ for $i = 2, \ldots, k$, then

$$\mathbb{E}[\hat{q}'^{m}_{\alpha,k}] \to \left(\frac{1}{k}\right)\gamma + \left(1 - \frac{1}{k}\right)q'_\alpha,$$

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 $\varepsilon > 0$,

$$\mathbb{P}\{|\hat{q}'^{m}_{\alpha,k} - \mathbb{E}[\hat{q}'^{m}_{\alpha,k}]| \geq \varepsilon\} \leq \frac{\mathrm{var}[\hat{q}'^{m}_{\alpha,k}]}{\varepsilon^2}.$$

By assumption when $k \to \infty$, $\mathrm{var}(\hat{q}'^{m}_{\alpha,k}) \to 0$, 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, $\hat{q}'^{m}_{\alpha,k} \xrightarrow{\mathbb{P}} \mathbb{E}[\hat{q}'^{m}_{\alpha,k}]$. Then, let $k \to \infty$, and it is easy to prove that $\mathbb{E}[\hat{q}'^{m}_{\alpha,k}] \to q'_\alpha$. □

Remark 4. (i) Note that in Theorem 4, we do *not* need $F_{(\lceil\alpha n\rceil)}(x;\theta)$ to satisfy Lemma 2(i), 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.

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 (1995)), 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 such that there exists a sequence $0 = T_0 \leq T_1 \leq T_2 \leq \cdots$ 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 nondelayed regenerative processes, i.e. the first cycle $\{X_t,\ T_0 \leq t \leq T_1\}$ has the same CDF 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 Asmussen (2003, Chapter 4, Theorem 1.2), 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

https://doi.org/10.1017/jpr.2016.36 Published online by Cambridge University Press

random variable W as $i \to \infty$. As in the last section, $F_t(x;\theta)$ and $F(x;\theta)$ denote the CDFs. 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 for $\tilde{n}$ cycles, generating $X_0, X_1, \ldots, X_{T_{\tilde{n}}-1}$ with a total number of samples $T_{\tilde{n}}$. Then the order statistics of this sequence and their corresponding IPA estimator are given by $X_{(1)} \le X_{(2)} \le \cdots \le X_{(\lceil \alpha T_{\tilde{n}} \rceil)} \le \cdots \le X_{(T_{\tilde{n}})}$ and $\mathrm{d}X_{(1)}/\mathrm{d}\theta$, $\mathrm{d}X_{(2)}/\mathrm{d}\theta, \ldots, \mathrm{d}X_{(\lceil \alpha T_{\tilde{n}} \rceil)}/\mathrm{d}\theta, \ldots, \mathrm{d}X_{(T_{\tilde{n}})}/\mathrm{d}\theta$, respectively. Denote $\hat{q}_\alpha^{\tilde{n}} = X_{(\lceil \alpha T_{\tilde{n}} \rceil)}$ and $I_{\tilde{n}} = \mathrm{d}X_{(\lceil \alpha T_{\tilde{n}} \rceil)}/\mathrm{d}\theta$. By Iglehart (1976), we know that in a neighbourhood of q_α, if the derivative of density of X exists and are bounded, then

$$\frac{\hat{q}_\alpha^{\tilde{n}} - q_\alpha}{\sigma_\alpha/(\mu \partial_1 F(q_\alpha;\theta)\tilde{n}^{1/2})} \xrightarrow{\mathrm{D}} \mathcal{N}(0,1) \quad \text{as } \tilde{n} \to \infty, \tag{8}$$

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 (8) also implies that $\hat{q}_\alpha^{\tilde{n}} \xrightarrow{\mathrm{D}} q_\alpha$ as $\tilde{n} \to \infty$. Then Theorem 1 holds with Assumption 2 replaced by the following assumption.

Assumption 4. *In a neighborhood 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 Lemmas 2(i) and 2(ii). If Assumptions 3 and 4 hold, and $\mathrm{d}X/\mathrm{d}\theta$ 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 (1982). 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 $M_i(\tilde{m}) = T_{i\tilde{m}} - T_{(i-1)\tilde{m}} = N_{(i-1)\tilde{m}+1} + \cdots + N_{i\tilde{m}}$, 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 $M_i(\tilde{m})$, we have $X_{(1,i)} \le X_{(2,i)} \le \cdots \le X_{(\lceil \alpha M_i(\tilde{m}) \rceil, i)} \le \cdots \le X_{(M_i(\tilde{m}),i)}$. The corresponding derivatives with respect to θ are given by

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

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

$$\hat{q}'^{\tilde{m}}_{\alpha,\tilde{k}} = \frac{1}{\tilde{k}} \sum_{i=1}^{\tilde{k}} I_{i,\tilde{m}}. \tag{9}$$

Although $\{X_t,\ t = 0, 1, \ldots\}$ is regenerative, establishing that their derivatives $\{\mathrm{d}X_t/\mathrm{d}\theta,\ t = 0, 1, \ldots\}$ are regenerative requires an additional special structure. One condition is to assume that $\mathrm{d}X_t/\mathrm{d}\theta = \phi(X_t;\theta)$, i.e. the derivative of X_t is only a function of X_t and does not contain other random variables. By Assumption 1, we know that $\phi(X_t;\theta) = -\partial_2 F(X_t;\theta)/\partial_1 F(X_t;\theta)$ is continuous with respect to the first argument. Since X_t is regenerative, then $\mathrm{d}X_t/\mathrm{d}\theta = \phi(X_t;\theta)$ is still regenerative. Glasserman (1993) considered more general conditions for regenerative processes whose derivatives are also regenerative. We summarize the results in the following lemma.

Lemma 3. (Glasserman (1993).) *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:*

https://doi.org/10.1017/jpr.2016.36 Published online by Cambridge University Press

(i) *each X_t and U_t, $t \geq 0$, is continuous in a neighbourhood of some point θ and differentiable at θ almost surely;*

(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: $A_1 \times \cdots \times A_r \to B$;*

(iii) *it holds that $\mathbb{P}\{h$ is continuously differentiable at $(U_1(\theta), \ldots, U_r(\theta)) \mid U_i(\theta) \in A_i,\ i = 1, \ldots, r\} = 1$. Then $\{\mathrm{d}X_t/\mathrm{d}\theta,\ t = 0, 1, \ldots\}$ 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 $\sup_{\tilde{m}} \mathbb{E}[I_{\tilde{m}}^2] < \infty$, then $\mathrm{var}(I_{\tilde{m}}) < \infty$ and $\mathrm{var}(\hat{q}'^{\tilde{m}}_{\alpha,\tilde{k}}) = 1/\tilde{k}\,\mathrm{var}(I_{\tilde{m}}) \to 0$ as $k \to \infty$, which satisfies the conditions in Theorem 3. The following corollary states the weak consistency of the batched estimator for regenerative processes.

Corollary 2. *Under the assumptions in Lemma 3, suppose that $\sup_{\tilde{m}} \mathbb{E}[I_{\tilde{m}}^2] < \infty$. If Assumptions 3 and 4 are satisfied, and $\mathrm{d}X/\mathrm{d}\theta$ exists, then $\hat{q}'^{\tilde{m}}_{\alpha,\tilde{k}} \xrightarrow{\mathbb{P}} q'_\alpha$ as $\tilde{m} \to \infty$ and $\tilde{k} \to \infty$.*

In the i.i.d. setting, $\mathbb{E}[\hat{q}_\alpha^m - q_\alpha] = O(m^{-1})$ and $\mathbb{E}[(\hat{q}_\alpha^m - q_\alpha)^2] = O(m^{-1})$, 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 (9) similar to the i.i.d. case. Although characterizing 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 that $\varphi(x) = -\partial_2 F(x;\theta)/\partial_1 F(x;\theta)$ is twice differentiable, $|\varphi''(x)| < K$, $\mathbb{E}[\hat{q}_\alpha^{\tilde{m}} - q_\alpha] = O(\tilde{m}^{-\gamma_1})$, and $\mathbb{E}[(\hat{q}_\alpha^{\tilde{m}} - q_\alpha)^2] = O(\tilde{m}^{-\gamma_2})$. Under the assumptions in Corollary 2, if $\inf_{\tilde{m}} \mathrm{var}(I_{\tilde{m}}) > 0$ then*

$$\frac{\hat{q}'^{\tilde{m}}_{\alpha,\tilde{k}} - q'_\alpha}{\sigma(\hat{q}'^{\tilde{m}}_{\alpha,\tilde{k}})} \xrightarrow{\mathrm{D}} \mathcal{N}(0,1) \quad \text{as } \tilde{m} \to \infty,\ \tilde{k} \to \infty \quad \text{and} \quad \frac{\tilde{k}^{1/2}}{\tilde{m}^{\gamma}} \to 0, \tag{10}$$

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 obtain 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 (10) can be written as

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

https://doi.org/10.1017/jpr.2016.36 Published online by Cambridge University Press

By definition, $\hat{q}'^{\tilde{m}}_{\alpha,\tilde{k}}$ 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 (10) satisfies

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

By the 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 $\text{var}(\hat{q}'^{\tilde{m}}_{\alpha,\tilde{k}}) = 1/\tilde{k}\,\text{var}(I_{\tilde{m}})$ and $\inf_{\tilde{m}} \text{var}(I_{\tilde{m}}) > 0$, we obtain

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

The theorem is proved. □

The batched estimator for regenerative processes is different from that in the previous section. To distinguish these two batched estimators, we call the batched estimator (7) 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 two different 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 (1982), the last cycle contains significantly more information than a typical cycle. Hong (2009) 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

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 (1968), (1972), and Heidelberger and Lewis (1984)), short-range dependent (SRD) linear processes (Hesse (1990)), and long-range dependent (LRD) linear processes (Ho and Hsing (1996)). In this subsection we consider dependent processes satisfying a ϕ-mixing condition, which was considered in the stationary setting by Liu and Hong (2009) for a kernel estimator.

All the notation is 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 $n \ge 1$, $|\mathbb{P}(B \mid A) - \mathbb{P}(B)| \le \phi(n)$, $\phi(n) \ge 0$, where $\phi(n)$ is nonincreasing with respect to n and $\lim_{n\to\infty} \phi(n) = 0$. Sen (1972) studied the Bahadur representation of sample quantile for ϕ-mixing sequences. Denote

$$A_0(\phi) = \sum_{n=1}^{\infty} (\phi(n))^{1/2} \quad \text{and} \quad v_i = \mathbb{E}[\mathbf{1}_{\{q_\alpha \ge X_0\}} \mathbf{1}_{\{q_\alpha \ge X_i\}}], \qquad i = 0, 1, 2, \ldots.$$

Let $v^2 = v_0 + 2\sum_{i=1}^\infty v_i$. By Sen (1972), $A_0(\phi) < \infty$ implies that $v < \infty$. From the main theorem in Sen (1972), if $A_0(\phi) < \infty$ and $0 < \partial_1 F(x;\theta) < \infty$, and $\partial_1^2 F(x;\theta)$

https://doi.org/10.1017/jpr.2016.36 Published online by Cambridge University Press

exists and is bounded, then $(\hat{q}_\alpha^n - q_\alpha)/(v/(\partial_1 F(q_\alpha;\theta)n^{1/2})) \xrightarrow{\mathrm{D}} \mathcal{N}(0,1)$ as $n \to \infty$. Therefore, Assumption 2 can be replaced by the following assumption.

Assumption 5. *It holds that* $A_0(\phi) < \infty$.

Corollary 3. *Suppose that* $\sup_n \mathbb{E}[I_n^2] < \infty$ *and* $F_{(\lceil \alpha n \rceil)}(x;\theta)$ *satisfy Lemmas 2(i) and 2(ii). If Assumptions 1 and 5 are satisfied, and* $\mathrm{d}X/\mathrm{d}\theta$ *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 $\{(X_t, \mathrm{d}X_t/\mathrm{d}\theta),\ t \geq 0\}$ satisfies the ϕ-mixing condition, the same condition required in Liu and Hong (2009). Under certain conditions, ϕ-mixing of $\{X_t,\ t \geq 0\}$ implies that $\{\mathrm{d}X_t/\mathrm{d}\theta,\ t \geq 0\}$ is ϕ-mixing. For example, if there is a function f satisfying some mild conditions, e.g. it is a linear function, such that $\{\mathrm{d}X_t/\mathrm{d}\theta = f(X_t),\ t \geq 0\}$, then $\mathrm{d}X_t/\mathrm{d}\theta$ satisfies a ϕ-mixing condition. When $m \to \infty$ for all $i \neq j$, $\mathrm{cov}(I_{i,m}, I_{j,m}) \to 0$. Then

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

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* $\{(X_t, \mathrm{d}X_t/\mathrm{d}\theta),\ t \geq 0\}$ *satisfies the* ϕ*-mixing condition. If Assumptions 1 and 5 are satisfied, then* $\hat{q}'^m_{\alpha,k} \xrightarrow{\mathbb{P}} q'_\alpha$ *as* $m \to \infty$ *and* $k \to \infty$.

The sequence $I_{i,m}, i = 1, 2, \ldots$, is derived from $\mathrm{d}X_t/\mathrm{d}\theta$, so it satisfies the ϕ-mixing condition. By Deo (1975), if Assumption 5 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

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

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 that* $\varphi(x) = -\partial_2 F(x;\theta)/\partial_1 F(x;\theta)$ *is twice differentiable,* $|\varphi''(x)| < K$, $\mathbb{E}[\hat{q}_\alpha^m - q_\alpha] = O(m^{-\gamma_1})$, *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 3, 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*

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

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 OU process, i.e. AR(1) process. For the G/G/1 queue, we consider the waiting (queueing) time, and compare the bias, mean-squared error (MSE), and coverage probability of the REG estimator with the DET estimator considered in Hong (2009). Although the central limit theorem results should be applicable to these examples (Theorem 5 for the queueing example and Corollary 5 for the discrete OU process), verifying each of the

https://doi.org/10.1017/jpr.2016.36 Published online by Cambridge University Press

conditions directly is difficult. For the first example, Appendix B provides details for verifying Assumptions 3 and 4 and the sup second-moment conditions required for Corollaries 1, 3, and 4, and a discussion of justifying the rate conditions required for Theorem 5.

4.1. The G/G/1 queue

Let W_i denote the waiting time of the ith customer in a first-come–first-served 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 $(i + 1)$th customer. Under appropriate stability conditions, W_i is a regenerative process with respect to indices of customers who initiate busy periods. The IPA estimator of W_i is given by direct differentiation, i.e.

$$\frac{\mathrm{d}W_{i+1}}{\mathrm{d}\theta} = \left(\frac{\mathrm{d}W_i}{\mathrm{d}\theta} + \frac{\mathrm{d}X_i}{\mathrm{d}\theta}\right)\mathbf{1}_{\{W_i+X_i\geq 0\}}. \tag{11}$$

Similar to Hong (2009, Section 8), 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 mean $\mu_A - \mu_S$, so the quantile sensitivities are analytically available for comparison.

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 $\tilde{n} = 10\,000$ 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 10 000 cycles generates 50 112 customers served in a replication, the simulation replication is continued until 50 200 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 Figure 1, the MSE in Figure 2, and the coverage probabilities in Figure 3. In both Figure 1 and Figure 2, the performance of the REG is generally better than the DET for $\alpha < 0.8$. When $\alpha > 0.8$, both absolute bias and MSE increase significantly, more so for the REG than for the DET. In Figure 3 we show the estimated coverage probabilities of the 90% confidence interval. When α is 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. In Figure 3 it is indicated that the REG coverage probabilities are more stable.

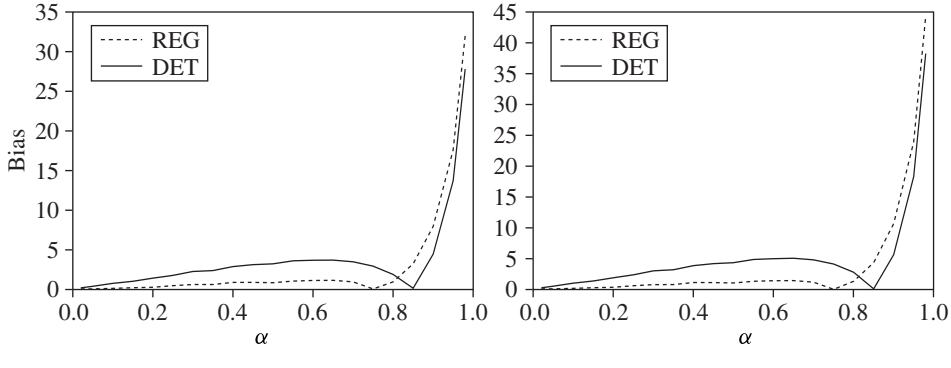


FIGURE 1: Absolute bias of quantile sensitivities with respect to μ_A (*left*) and μ_S (*right*).

https://doi.org/10.1017/jpr.2016.36 Published online by Cambridge University Press

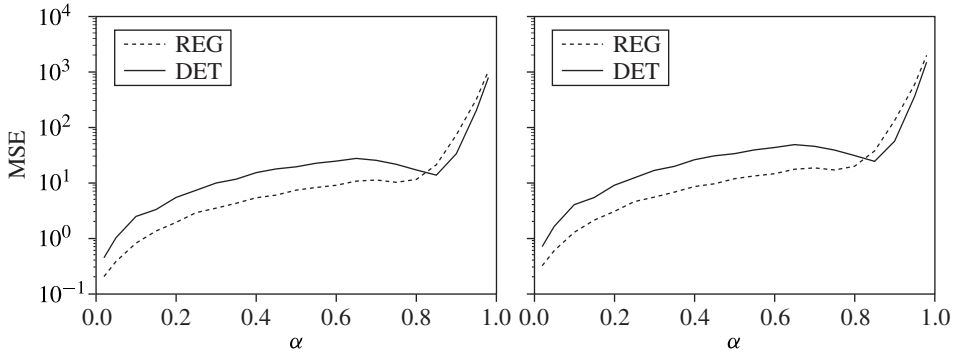


FIGURE 2: MSE of quantile sensitivities with respect to μ_A (*left*) and μ_S (*right*).

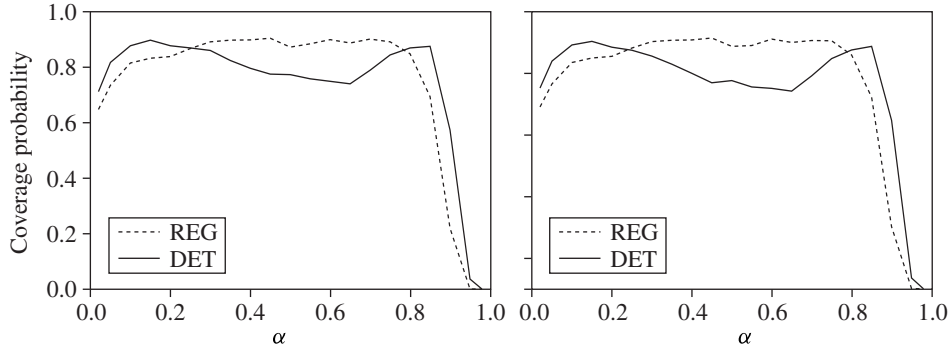


FIGURE 3: Coverage probabilities of quantile sensitivities with respect to μ_A (*left*) and μ_S (*right*).

4.2. The OU process

The OU process is given by the stochastic differential equation

$$\mathrm{d}X_t = \theta(\mu - X_t)\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, \qquad t \geq 0,$$

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

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

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 $\theta = 0.2$, $\mu = 0.1$, $\sigma = 0.1$, $\Delta t = 0.05$, 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. In Figure 4 it is indicated 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. In Figure 5 we show the bias and MSE as the total sequence length N increases for fixed $k = 100$ and $\alpha = 0.8$.

https://doi.org/10.1017/jpr.2016.36 Published online by Cambridge University Press

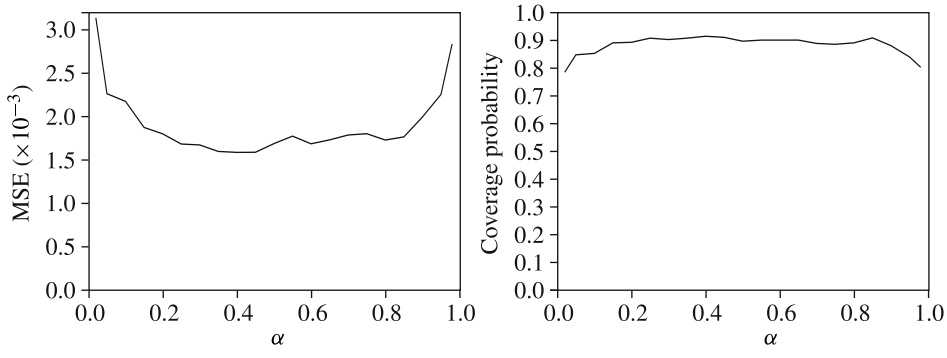


FIGURE 4: MSE (*left*) and 90% coverage probability (*right*) of quantile sensitivity for the OU example ($N = 10^6$, $m = 10^4$, $k = 100$).

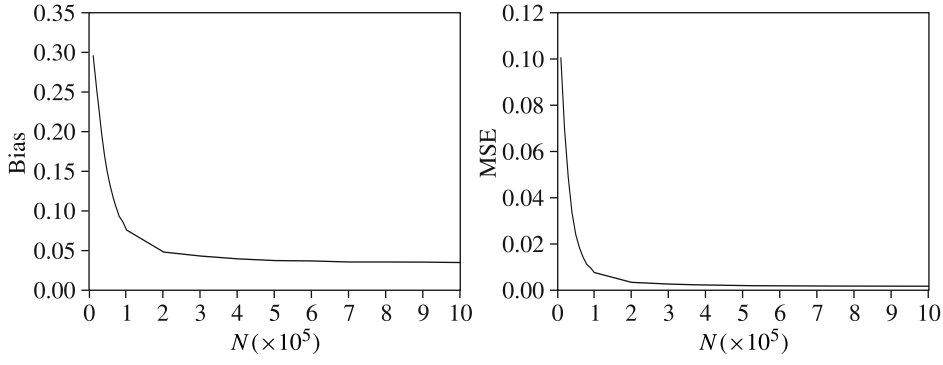


FIGURE 5: Bias (*left*) and MSE (*right*) of quantile sensitivity for the OU example($K = 100$, $\alpha = 0.8$).

5. Conclusion

In this paper we estimated the quantile sensitivities of dependent sequences via IPA, and constructed a batched estimator which was asymptotically unbiased and weakly consistent. Two widely used processes, regenerative processes and ϕ-mixing processes, were studied. For regenerative processes, we improved the batched estimator based on regenerative cycles, and a central limit theorem was given. For ϕ-mixing processes, we also provided 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, as long as the appropriate quantile estimator converges to the true quantile. Two numerical examples, the G/G/1 queue and discrete OU process, were 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 *et al.* (2014)). A limitation of our work is that the assumptions in the theoretical results are typically difficult to verify in practice, especially some of the conditions in Theorem 5 and Corollary 5, so an important topic for future research is to be able to find sufficient conditions that are more readily verifiable.

https://doi.org/10.1017/jpr.2016.36 Published online by Cambridge University Press

Appendix A.

Proof of Theorem 1. By Lemma 1, $\mathbb{E}[\mathrm{d}X_t/\mathrm{d}\theta \mid X_t = x] = -\partial_2 F(x;\theta)/\partial_1 F(x;\theta)$. Substituting $x = q_\alpha$ into the equation above, and by (1),

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

By Assumption 2, $\hat{q}^n_\alpha \xrightarrow{\mathrm{D}} q_\alpha$ as $n \to \infty$, and by Assumption 1, for $i = 1, 2$, $\partial_i F(\hat{q}^n_\alpha, \theta) \xrightarrow{\mathrm{D}} \partial_i F(q_\alpha, \theta)$, so with Assumption 1(ii), $-\partial_2 F(\hat{q}^n_\alpha, \theta)/\partial_1 F(\hat{q}^n_\alpha; \theta) \xrightarrow{\mathrm{D}} -\partial_2 F(q_\alpha, \theta)/\partial_1 F(q_\alpha; \theta)$ as $n \to \infty$. Note that

$$\begin{aligned}
\mathbb{E}[I_n] &= \mathbb{E}\left[\frac{\mathrm{d}X_{(\lceil \alpha n \rceil)}}{\mathrm{d}\theta}\right] \\
&= \int_{-\infty}^{\infty} \mathbb{E}\left[\frac{\mathrm{d}X_{(\lceil \alpha n \rceil)}}{\mathrm{d}\theta} \;\middle|\; X_{(\lceil \alpha n \rceil)} = x\right] \mathrm{d}F_{X_{(\lceil \alpha n \rceil)}}(x) \\
&= \int_{-\infty}^{\infty} -\frac{\partial_2 F(x;\theta)}{\partial_1 F(x;\theta)} \,\mathrm{d}F_{X_{(\lceil \alpha n \rceil)}}(x) \\
&= \mathbb{E}\left[-\frac{\partial_2 F(\hat{q}^n_\alpha;\theta)}{\partial_1 F(\hat{q}^n_\alpha;\theta)}\right], \qquad \text{(A.1)}
\end{aligned}$$

where $F_{X_{(\lceil \alpha n \rceil)}}(x)$ is the CDF 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 that $\partial_2 F(\hat{q}^n_\alpha;\theta)/\partial_1 F(\hat{q}^n_\alpha;\theta)$ is uniformly integrable.

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

where the inequality follows by Jensen's inequality for the 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,

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

Appendix B. Verifying Assumptions 3, 4, and $\sup_n \mathbb{E}[I_n^2]$

For Assumption 3, since the queue in the first example is stable, the steady-state distribution is known to be an exponential distribution with parameter $1/\mu_S - 1/\mu_A$, so it is easy to verify that Assumptions 3 and 4 are satisfied. The main difficulty is to verify the conditions for the sequential distributions $\{F_t(x;\theta),\ t = 0, 1, 2, \ldots\}$. However, we can verify a sufficient condition, similar to Jiang and Fu (2015).

In the first example, $W_{i+1} = (W_i + X_i)^+$, where $X_i = S_i - A_i$. The density functions of S_i and A_i are $f_S(x) = 1/\mu_S \exp(-x/\mu_S)$, $x \ge 0$, and $f_A(x) = 1/\mu_A \exp(-x/\mu_A)$, $x \ge 0$, respectively. Suppose that we are interested in estimating the sensitivity with respect to μ_S.

https://doi.org/10.1017/jpr.2016.36 Published online by Cambridge University Press

First we calculate the density function of X_i, which is given by

$$f_{X_i}(x;\mu_S) = \begin{cases} \dfrac{1}{\mu_S+\mu_A}\exp\left(-\dfrac{1}{\mu_S}x\right), & x \geq 0, \\ \dfrac{1}{\mu_S+\mu_A}\exp\left(\dfrac{1}{\mu_A}x\right), & x < 0. \end{cases} \tag{B.1}$$

Denote $F_{X_i}(x;\mu_S) = \int_{-\infty}^{x} f_{X_i}(y;\mu_S)\,\mathrm{d}y$. By the Lindley integral equation,

$$F_{W_i}(x;\mu_S) = \int_{0^-}^{\infty} F_{X_i}(x-y;\mu_S)F_{W_{i-1}}(\mathrm{d}y;\mu_S). \tag{B.2}$$

We suppose that $W_1 = 0$ and $W_2 = X_1^+$. So when $y > 0$, $F_{W_2}(\mathrm{d}y;\mu_S) = f_{X_1}(y;\mu_S) = 1/(\mu_S+\mu_A)\exp(-y/\mu_S)$. Then by (B.2),

$$F_{W_3}(x;\mu_S) = \int_{0^+}^{\infty} F_{X_3}(x-y;\mu_S)F_{W_2}(\mathrm{d}y;\mu_S) + F_{X_3}(x;\mu_S)F_{W_2}(0;\mu_S), \tag{B.3}$$

where $F_{W_2}(0;\mu_S) = \int_{-\infty}^{0^-} 1/(\mu_S+\mu_A)\exp(x/\mu_A)\,\mathrm{d}x > 0$. Then we check the partial derivative with respect to the first argument. By $F_{X_3}(x-y;\mu_S) \leq 1$, $\int_{0^+}^{\infty} F_{W_2}(\mathrm{d}y;\mu_S) < \infty$, and the dominated convergence theorem, the interchangeability of the differentiation and integration holds, and

$$\partial_x F_{W_3}(x;\mu_S) = \int_{0^+}^{\infty} \partial_x F_{X_3}(x-y;\mu_S)F_{W_2}(\mathrm{d}y;\mu_S) + \partial_x F_{X_3}(x;\mu_S)F_{W_2}(0;\mu_S).$$

Then it suffices to verify that $\partial_x F_{X_3}(x;\mu_S)$ exists and is continuous with respect to x. Note that for a given x, $F_{X_3}(x;\mu_S) = \int_{-\infty}^{x} f_{X_3}(y;\mu_S)\,\mathrm{d}y$, so by (B.1), $\partial_x F_{X_3}(x;\mu_S) = f_{X_3}(x;\mu_S)$ exists and is continuous. Moreover, for given x, $\partial_x F_{X_3}(x;\mu_S) > 0$, then

$$\partial_x F_{W_3}(x;\mu_S) \geq \partial_x F_{X_3}(x;\mu_S)F_{W_2}(0;\mu_S) > 0,$$

which means $\partial_x F_{W_3}(x;\mu_S)$ is strictly positive.

Take the partial derivative with respect to the second argument on both sides of (B.3). Note that by (B.1), $\partial_{\mu_S} F_{X_3}(x;\mu_S)$ and $\partial_{\mu_S} F_{W_2}(\mathrm{d}y;\mu_S)$ exist and are continuous with respect to μ_S. Then by the dominated convergence theorem, we can exchange differentiation and integration, and use the product rule of differentiation to check that $F_{W_3}(x;\mu_S)$ is continuously differentiable with respect to the second argument. By (B.2), we can check that $F_{W_i}(x;\mu_S)$ satisfies Assumption 3 step by step.

Next, consider $\sup_n \mathbb{E}[I_n^2]$. By the definition of I_n in (4), I_n is one of $\mathrm{d}W_i/\mathrm{d}\theta$, $i = 1,\ldots,n$, so if we can verify that $\mathbb{E}[(\mathrm{d}W_i/\mathrm{d}\theta)^2] < \infty$ for all $i = 1,\ldots,n$, then $\sup_n \mathbb{E}[I_n^2] < \infty$. By (11),

$$\begin{aligned} \mathbb{E}\left[\left(\frac{\mathrm{d}W_{i+1}}{\mathrm{d}\theta}\right)^2\right] &= \mathbb{E}\left[\left(\frac{\mathrm{d}W_i}{\mathrm{d}\theta}+\frac{\mathrm{d}X_i}{\mathrm{d}\theta}\right)^2 \mathbf{1}_{\{W_i+X_i\geq 0\}}\right] \\ &\leq \mathbb{E}\left[\left(\frac{\mathrm{d}W_i}{\mathrm{d}\theta}+\frac{\mathrm{d}X_i}{\mathrm{d}\theta}\right)^2\right] \\ &\leq 2\mathbb{E}\left[\left(\frac{\mathrm{d}W_i}{\mathrm{d}\theta}\right)^2\right] + 2\mathbb{E}\left[\left(\frac{\mathrm{d}X_i}{\mathrm{d}\theta}\right)^2\right]. \end{aligned}$$

https://doi.org/10.1017/jpr.2016.36 Published online by Cambridge University Press

By (B.1), we have $\mathbb{E}[(\mathrm{d}X_i/\mathrm{d}\theta)^2] < \infty$. Then induction yields $\mathbb{E}[(\mathrm{d}W_i/\mathrm{d}\theta)^2] < \infty$ for all i, so $\sup_n \mathbb{E}[I_n^2] < \infty$. For the REG estimator, the subscript n is changed by the number of regenerative cycles $\tilde{m}$, and the result holds.

For the second example, Assumption 3 can be verified by checking the convolution of each random variable as discussed above. Assumption 5 is a common assumption in most papers about quantile and quantile sensitivity (Heidelberger and Lewis (1984), Liu and Hong (2009)), and is typically difficult to verify for practical problems. Similar to checking the convergence rate in Theorem 5, we can verify it roughly in the numerical example.

Acknowledgements

We thank the anonymous referee for useful comments and suggestions that have helped improve the paper. 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

Asmussen, S. (2003). *Applied Probability and Queues*, 2nd edn. Springer, New York.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Prob. Surv.* **2**, 107–144.

Deo, C. M. (1975). A functional central limit theorem for stationary random fields. *Ann. Prob.* **3,** 708–715.

Doukhan, P. (1994). *Mixing: Properties and Examples*. Springer, New York.

Fu, M. C. (2006). Gradient estimation. In *Handbooks in Operations Research and Management Science, Simulation*, eds S. G. Henderson and B. L. Nelson, Elsevier, Amsterdam, pp. 575–616.

Fu, M. C., Hong, L. J. and Hu, J. Q. (2009). Conditional Monte Carlo estimation of quantile sensitivities. *Manag. Sci.* **55,** 2019–2027.

Glasserman, P. (1993). Regenerative derivatives of regenerative sequences. *Adv. Appl. Prob.* **25,** 116–139.

Heidelberger, P. and Lewis, P. A. W. (1984). Quantile estimation in dependent sequences. *Operat. Res.* **32,** 185–209.

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.

Ho, H-C. and Hsing, T. (1996). On the asymptotic expansion of the empirical process of long-memory moving averages. *Ann. Statist.* **24,** 992–1024.

Hong, L. J. (2009). Estimating quantile sensitivities. *Operat. Res.* **57,** 118–130.

Iglehart, D. L. (1976). Simulating stable stochastic systems. VI. Quantile estimation. *J. Assoc. Comput. Mach.* **23,** 347–360.

Jiang, G. and Fu, M. C. (2015). Technical note—on estimating quantile sensitivities via infinitesimal perturbation analysis. *Operat. Res.* **63,** 435–441.

Jiang, G., Fu, M. C. and Xu, C. (2014). Bias reduction in estimating quantile sensitivities. In *Proceedings of the 19th IFAC World Congress*, eds E. Boje and X. Xia, IFAC, pp. 10463–10468.

Liu, G. and Hong, L. J. (2009). Kernel estimation of quantile sensitivities. *Naval Res. Logistics* **56,** 511–525.

Meketon, M. S. and Heidelberger, P. (1982). A renewal theoretic approach to bias reduction in regenerative simulations. *Manag. Sci.* **28,** 173–181.

Seila, A. F. (1982). A batching approach to quantile estimation in regenerative simulations. *Manag. Sci.* **28,** 573–581.

Sen, P. K. (1968). Asymptotic normality of sample quantiles for m-dependent processes. *Ann. Math. Statist.* **39,** 1724–1730.

Sen, P. K. (1972). On the Bahadur representation of sample quantiles for sequences of ϕ-mixing random variable. *J. Multivariate Anal.* **2,** 77–95.

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

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

Suri, R. and Zazanis, M. A. (1988). Perturbation analysis gives strongly consistent sensitivity estimates for the $M/G/1$ queue. *Manag. Sci.* **34,** 39–64.

Volk-Makarewicz, W. and Heidergott, B. (2012). Sensitivity analysis of quantiles. *AENORM* **20,** 26–31.

https://doi.org/10.1017/jpr.2016.36 Published online by Cambridge University Press