Gradient Estimation for Quantiles of Stationary Waiting Times

Bernd Heidergott ∗ Warren Volk-Makarewicz ∗ Felisa Vázquez-Abad **

∗ Vrije Universiteit Amsterdam, dep. Econometrics, Amsterdam, (e-mail: {bheidergott,wmakarewicz}@feweb.vu.nl).
** The City University of New York, dep. Computer Science, New York, (felisa.vazquez-abad@hunter.cuny.edu).

Abstract:
Quantiles of customer based performance characteristics have been adopted in many areas for measuring the quality of service. Recently, sensitivity analysis of quantiles has attracted quite some attention. Sensitivity analysis of quantiles is particularly challenging as quantiles cannot be expressed as the expected value of some sample performance function, and it is therefore not evident how standard gradient estimation methods can be applied. While sensitivity analysis of quantiles of waiting times for static or fixed time horizon problems is well understood, quantile estimation for stationary waiting times remains an open question. This paper will close this gap and will provide a framework for gradient estimation for quantiles of stationary waiting times.

Keywords: quantile, Markov chains, gradient estimation, control, measure-valued differentiation.

1. INTRODUCTION

The α-quantile of a random variable Y equals u if and u is the largest feasible value such that \( P(Y \geq u) \leq \alpha \). Quantiles are also known as value at risk (VaR) and are widely used as measures of quality in the service industry. Unfortunately, the α-quantile can only be obtained in closed form in special cases and one usually has to resort to simulation (resp. statistical estimation) for evaluating a quantile. In this paper we assume that Y can be influenced through some control parameter \( \theta \) and we provide estimators for the sensitivity of the α-quantile with respect to \( \theta \). Specifically, we consider in this paper the case \( Y \) is a stationary version of a Markov chain.

This research is motivated by the following. Consider the sojourn time in a G/G/1 queue modeling a communication link. Typically, a service contract with a provider usually guarantees that the “typical” α-quantile of the sojourn time of a typical job through the systems does not exceed a specified value \( \eta \). Suppose that \( \theta \) is the capacity of a server in a communication link. Then the sensitivities of the α-quantile of the stationary sojourn time of a job yields the effect of \( \theta \) on the service quality. This information is of particular interest to the provider as the sensitivity information allows pricing of the service. To see this note that from the provider’s point of view the standard performance should be optimal subject to the service requirement expressed in the quantile. Under suitable conditions, the performance optimization problem can be solved in an iterative way by the Lagrange multiplier method, Kushner (2003), yielding the shadow price of service quality restriction and thus providing information on the economic impact of \( \theta \) on the overall profit. Quantile sensitivity estimation has been studied intensively in the literature for the case of a static problem (i.e., a performance measure that is defined over a finite deterministic time horizon), see Fu et al. (2009); Liu and Hong (2009); Hong and Liu (2009); Hong (2009). These papers apply sample-path differentiation to the quantile sensitivity estimation problem.

A main challenge with this approach is that quantile sensitivity estimation involves differentiating probabilities (details will be given later in the text), which is related to differentiating indicator mappings in the sample path sense. As this is impossible in the pathwise sense, the literature has developed methods to circumvent this problem. It is worth noting that pathwise differentiation of indicator functions can, in principle, be dealt with through proper conditioning, this is called smoothed perturbation analysis (SPA), see Fu and Hu (1997) for details. In this paper we choose a different approach based on measure-valued differentiation (MVD). In particular, differentiating indicator mappings constitutes no obstacle for the MVD method and estimators for stationary characteristics of Markov chains are well known, see, for example, Heidergott et al. (2006). Moreover, for many problems that are of importance in practice, MVD is straightforward to apply, see Heidergott et al. (2009) for details. It is worth noting that while MVD, in general, requires to simulate two alternative versions of a system, single run implementations do exist and have been shown to be computationally more efficient than IPA, Heidergott et al. (2009). MVD is related to the Perturbation Realization (PR) approach for discrete state Markov Chains, Cao (2007). For MVD and PR alike finite perturbations are generated and propagated via two “phantom” sequences and the value of such a propagation is cumulated until both “phantom” sequences reach the
same state. For a more detailed discussion on the relation between MVD and PR we refer to Heidergott and Cao (2002). In this paper we will apply the MVD approach to the quantile sensitivity estimation problem. In case the quantile is known but its derivative is not, we will provide sufficient condition for unbiasedness of our estimator. In case that the quantile itself is unknown, we will replace the known quantile by an appropriate order statistic estimator and we will show that the resulting estimator has good numerical performance in the sense that the bias introduced by replacing the quantile by the order statistics does result in a negligible bias. A particular neat version of our estimator can be obtained if one is able to directly sample the stationary Markov chain. While this in general impossible, we will provide the for case of the stationary workload in the M/G/1 queue a perfect sampling scheme for simulating perfect samples of the workload process.

The paper is organized as follows. Section 2 provides an analytical expression for the derivative of a quantile. The general Markov chain analysis is presented in Section 3. Section 4 is devoted to the analysis of the G/G/1 examples. Numerical examples are provided in Section 5. We conclude the paper by identifying topics of further research.

2. QUANTILE DERIVATIVE

Let \( F_\theta \) denote the cumulative distribution function of a continuous random variable \( X_\theta \). Denote the \( \alpha \)-quantile of \( X_\theta \), resp. of \( F_\theta \), by \( q_\alpha(\theta) \), then

\[
F_\theta(q_\alpha(\theta)) = \alpha. \tag{1}
\]

Provided that \( F_\theta \) and \( q_\alpha(\theta) \) are both differentiable, then differentiating the expression in (1) yields

\[
\left( \frac{\partial}{\partial \theta} \right)_{x=q_\alpha(\theta)} F_\theta(q_\alpha(\theta)) \frac{\partial}{\partial \theta} q_\alpha(\theta) + \frac{\partial}{\partial x} \bigg|_{x=q_\alpha(\theta)} F_\theta(q_\alpha(\theta)) = 0.
\]

Let

\[
\frac{\partial}{\partial \theta} F_\theta(x) = F'_\theta(x), \quad x \geq 0, \quad \text{and} \quad \frac{\partial}{\partial \theta} q_\alpha(\theta) = q'_\alpha(\theta),
\]

then,

\[
q'_\alpha(\theta) = -\frac{F'_\theta(q_\alpha(\theta))}{F_\theta(q_\alpha(\theta))}.
\]

3. MARKOV CHAIN ANALYSIS

Let \( \Theta = (a, b) \subset \mathbb{R} \), with \( a < b \) and \( a, b \in \mathbb{R} \). Consider a Markov process \( X_\theta(n) \) with transition probability \( P_\theta \) with an unique stationary distribution \( \pi_\theta \) and state space \([0, \infty)\). More specifically, assume that a transition from \( X_\theta(n) \) to \( X_\theta(n+1) \) can be constructed by through a stochastically recursive sequence in the following way

\[
X_\theta(n+1) = h(X_\theta(n), Y_\theta(n)),
\]

for a measurable mapping \( h \) and a finite vector of random variables \( Y_\theta(n) \) independent of everything else driving the transition. In a queueing system model, \( Y_\theta(n) \), typically contains samples of service and inter-arrival times and routing variables, for example the waiting time example below.

3.1 Estimating the Value of the Density of a Stationary Markov Chain at a Certain Point

Assume that \( P_\theta(\cdot, x) \) has Lebesgue density \( f_\theta(\cdot, x) \) and denote the density of the stationary distribution \( \pi_\theta \) by \( f_\theta^\infty \). The following expression for the density of \( \pi_\theta \) holds.

Lemma 1. For any \( x \geq 0 \) it holds that

\[
f_\theta^\infty(x) = E[f_\theta(x, X_\theta)],
\]

with \( X_\theta \) being a sample from \( \pi_\theta \).

Proof: By stationarity,

\[
\pi_\theta([0, x]) = \int_0^x f_\theta(r, u)f_\theta^\infty(u) du dr.
\]

Differentiating with respect to \( x \) yields for the density \( f_\theta^\infty(x) \)

\[
f_\theta^\infty(x) = \int_0^\infty f_\theta(x, u)f_\theta^\infty(u) du = E[f_\theta(x, X_\theta)],
\]

with \( X_\theta \) being a sample from \( \pi_\theta \). \( \square \)

3.2 Derivative Estimation for Stationary Markov Chains

Let \( B_\theta \) denote the set of bounded measurable mappings.

We assume the following

(A1) We assume that for all \( u \), the transition kernel \( P_\theta(u, \cdot) \)

is \( B_\theta \)-differentiable with \( B_\theta \)-derivative \( (c_\theta, P_\theta^+, P_\theta^-) \), i.e., we assume that for all measurable and bounded mappings \( g \) it holds that

\[
\int g(u)P_\theta^+(s, du) - \int g(u)P_\theta^-(s, du) = c_\theta \int \left( \int g(u)P_\theta^+(s, du) - \int g(u)P_\theta^-(s, du) \right) ds
\]

for all \( s \).

(A2) \( \{X_\theta(n)\} \) is Harris recurrent with atom \( \alpha \) and

\[
\tau_\theta = \inf\{n \geq 0 : X_\theta(n) \in \alpha\}
\]

denotes the first entrance time into the atom \( \alpha \).

Specifically, we assume that \( E[\tau_\theta] < \infty \) for \( \theta \in \Theta \).

Note that condition (A1) assumes that the normalizing constant of \( P_\theta^+(u, \cdot) \) is independent of \( u \). This typically holds in applications and simplifies the following analysis.

We now construct two variants of the sequence \( X_\theta(n) \) which are called phantoms. For \( k \in \mathbb{N} \), we construct the sequence \( \{X_\theta^+(n; s) : n \geq 1\} \) as follows with the initial state set to \( s \), i.e., \( s = X_\theta(1, s) \). The transition from \( X_\theta(1, s) \) to \( X_\theta(2, s) \) is governed by \( P_\theta^+ \) (compared to \( P_\theta \) for the nominal sequence). After time 2, the transitions of \( X_\theta^+(n, s) \) are governed by \( P_\theta^+ \) again. In the same way, we define the sequence \( \{X_\theta^-(n, s) : n \geq 1\} \), where instead of \( P_\theta^+ \) the kernel \( P_\theta^- \) is used for transition from 1 to 2. Furthermore, denote by \( \tau_\theta^+(s) \) the first time both variants simultaneously hit \( \alpha \), i.e.,

\[
\tau_\theta^+(s) = \inf\{n \geq k : X_\theta^+(n, s) \in \alpha, X_\theta^-(n, s) \in \alpha\}.
\]

In order to establish sufficient conditions for unbiasedness of the MVD estimator, we will require that \( X_\theta(n) \)
is geometrically ergodic. More specifically, for a general transition kernel \( Q \) on \([0,\infty)\), let
\[
||Q|| = \sup_{s,A} \sup_{\theta} |Q(s,A)|,
\]
where the second supremum is carried out over all measurable subsets of \([0,\infty)\). We also introduce the ergodic projector of \( \pi \) by \( \Pi_\theta \), i.e., for any distribution on \([0,\infty)\) it holds that \( \mu_\theta = \pi_\theta \). Note that in case of uni-chain Markov Kernels with discrete state space, \( \Pi_\theta \) is matrix with all rows identical to \( \pi \).

Provided that the value of \( q \) is known, we replace \( \pi_\theta \) by \( \Pi_\theta \) if for all \( n \)
\[
||\Pi_\theta^n - \Pi_\theta|| \leq \gamma \rho^n,
\]
for some finite constant \( \gamma \) and \( \rho \in [0,1) \).

We now can state the main result on the MVD estimator for the stationary distribution function.

**Theorem 2.** Let the inter-arrival times be \( B_s \)-differentiable and suppose that \( (A1) \) and \( (A2) \) hold. If \( X_\theta(n) \) is geometrically ergodic, then
\[
\frac{\partial}{\partial \theta} F_\theta^\infty(x) = c \times \frac{\partial}{\partial \theta} E \left( \sum_{k=n}^{\tau^+_x(x)} \left( 1 \{ X^+_\theta(n,X) \leq x \} - 1 \{ X^-_\theta(n,X) \leq x \} \right) \right),
\]
where \( X_\theta \) is distributed according to \( \pi_\theta \).

**Proof:** Let \( \pi_\theta \) denote the stationary distribution of the waiting time and note that \( F_\theta^\infty(x) = \pi_\theta([0,x]) \), for \( x \geq 0 \). Hence,
\[
\frac{\partial}{\partial \theta} F_\theta^\infty(x) = \int 1_{([0,x])}(u) \pi_\theta(du).
\]
The proof then follows from Theorem 4.2 in Heidergott et al. (2006) (see Section 5.1 of this paper for details on the estimation scheme). \( \square \)

### 3.3 The Quantile Estimator

Provided that the value of \( q_\alpha \) is known, combining the results of the previous sections, we arrive at the MVD sensitivity estimator for the quantile
\[
q'_\alpha(\theta) = \frac{c \theta}{E[f_\alpha(q_{\alpha},X_\theta)]} \times \frac{\partial}{\partial \theta} E \left( \sum_{k=n}^{\tau^+_x(X_\theta)} \left( 1 \{ X^+_\theta(n,X) \leq q_{\alpha} \} - 1 \{ X^-_\theta(n,X) \leq q_{\alpha} \} \right) \right).
\]
If \( q_{\alpha} \) is unknown, we replace \( q_{\alpha} \) by an appropriate order statistic, which is detailed in the following section. Let \( \{X_\theta^n(n) : 1 \leq n \leq m\} \) be a sample of i.i.d. copies of samples from \( \pi_\theta \), and let \( X_\theta^n(m) := X_\theta(\{\alpha m\}) \). It is well known that \( X_\theta^n(m) \) converges with probability one to \( q_{\alpha} \) and it is a widely used test statistic for estimating \( q_{\alpha} \). Bahadur (1966) provides a quantile estimator for i.i.d. data, and shows that the estimator is strongly consistent and has an asymptotic normal distribution. For a thorough review of quantile estimation for i.i.d. data we refer to Serfling (1980). Sen (1972) extends the result of Bahadur (1966), and shows that the strong consistency and the asymptotic normality of the estimator also holds for dependent data under some mild regularity conditions. In the simulation literature, a number of variance reduction techniques have been proposed for quantile estimation; for example, Hsu and Nelson (1990). Hestenberg and Nelson (1998) use control variates and Glynn (1996) applies importance sampling, while Avramidis and Wilson (1998) employ correlation-induction techniques.

If \( q_{\alpha} \) is not known, we replace it by \( X_\theta^n(m) \). As we will illustrate with the numerical examples, the bias introduced by replacing \( q_{\alpha} \) in (3) by the order statistic is negligible for \( m \) sufficiently large.

### 4. EXAMPLES

#### 4.1 Stationary Waiting Times in the G/G/1 Queue

Let \( \Theta = (a,b) \subset \mathbb{R} \), with \( a < b \) and \( a,b \in \mathbb{R} \). We define \( W_\theta(n) \) be the waiting time of the \( n \)th customer arriving to a single server queue with infinite waiting capacity. We also define \( \{A_\theta(n)\} \) be the i.i.d. sequence of inter-arrival times and \( \{S(n)\} \) the i.i.d. sequence of service times. Lindley’s recursion yields:
\[
W_\theta(n+1) = \max(W_\theta(n) + S(n) - A_\theta(n+1), 0) = h(W_\theta(n), S(n), A_\theta(n+1)),
\]
for \( n \geq 1 \), and \( W_\theta(1) = 0 \). For \( w > 0 \), the conditional distribution of waiting time is given by
\[
F_\theta(w|u) = P_\theta(u; (0,w]) = \int_0^w \left( \int_{\max(u+s-w,0)}^{u+s} f_\theta^A(a) \, da \right) f_\theta^S(s) \, ds
\]
and
\[
F_\theta(0|u) = P_\theta(u; \{0\}) = \int_0^{u+s} \left( \int_0^{u+s} f_\theta^A(a) \, da \right) f_\theta^S(s) \, ds = 1 - E[F_\theta^A(u + S) - F_\theta^A(max(u + S - w, 0))],
\]
with \( f_\theta^A \) is the density of the inter-arrival times and \( f_\theta^S \) that of the service times.

For the ergodicity of the Markov chain \( \{W_\theta(n)\} \) we need that the supremum of \( E[S(1)]/E[A_\theta(1)] \) with respect to \( \theta \) over \( \Theta \) is less than one, see, for example, Cohen (1969).

Applying Lemma 1, yields
\[
f_\theta^\infty(w) = E[f_\theta^A(max(W + S - w, 0))], w > 0,
\]
where \( W \) is distributed as the stationary waiting time.

The phantoms \( W_\theta^+(n,s) \) are constructed easily from a weak derivative \((c \theta, A_\theta^+, A_\theta^-) \) of \( A_\theta \). Indeed, a realization of \( W_\theta^+(1,s) \) is obtained from
\[
W_\theta^+(1,s) = \max(S + s - A_\theta^+, 0)
\]
and, in the same way,
\[
W_\theta^-(1,s) = \max(S(1) + s - A_\theta^-, 0).
\]
From then on, the construction of \( W_{\theta}^{\pm}(n, s) \) is just like that of the nominal sequence, i.e.,
\[
W_{\theta}^{\pm}(n+1, s) = \max\{W_{\theta}^{\pm}(n, s) + S(n) - A_{\theta}(n+1), 0\}
\]
for \( n \geq 2 \). See Heidergott et al. (2006) for details.

**Theorem 3.** Let the inter-arrival times be \( B_{\theta} \)-differentiable with \( B_{\theta} \) derivative \((c_{\theta}, A_{\theta}^{0}, A_{\theta}^{0})\). If \( \inf_{\theta \in \Theta} E[S(1) - A_{\theta}(1)] < 0 \), then
\[
q_{\alpha}'(\theta) = -\frac{c_{\theta}}{E[F_{A}(\max(W_{\theta} + S - q_{\alpha}, 0))] \times}
\]
\[
\mathbb{E}\left[\sum_{k=0}^{\infty} \left(1_{[W_{\theta}(n,W_{\theta}) \leq q_{\alpha}]} - 1_{[W_{\theta}(n,W_{\theta}) > q_{\alpha}]}\right)\right],
\]
with \( W_{\theta} \) being a sample of the stationary waiting time.

**Proof:** It has been shown in Heidergott et al. (2006) that \( \inf_{\theta \in \Theta} E[S(1) - A_{\theta}(1)] < 0 \) implies geometric ergodicity of \( X_{\theta}(n) \). Hence, the proof of the theorem follows from combining Theorem 2 and Lemma 1. \( \square \)

### 4.2 Stationary Workload Process in the M/G/1/T Queue

Let \( X_{\theta}(n) \) denote the amount of work seen by the \( n \)-th arriving customer to an M/G/1 queue with workload restriction \( T \). More specifically, service times are assigned to customers upon arrival and the server accepts a maximal amount of work given by the constraint \( T \), i.e., if a customer is only accepted if the amount of work at the server (where the amount of work in the queue is given by the remaining service time of the customer in service and the service times of all customers in queue) after admittance of the customer does not exceed \( T \), the customer is otherwise lost. By the PASTA property, the stationary version of \( X_{\theta}(n) \) reflects the stationary amount of work in the queue in the continuous time model. In this section, we address estimating the sensitivity of the quantile of stationary amount of work in the queue.

The sequence \( X_{\theta}(n) \) follows a Lindley type of recursion
\[
X_{\theta}(n+1) = \max\{X_{\theta}(n) + S(n) - A_{\theta}(n+1), 0\},
\]
for \( n \geq 1 \), and \( X_{\theta}(1) = 0 \). Note that \( X_{\theta}(n) \) regenerates whenever \( A_{\theta}(n+1) \geq T \). Let \( \eta_{\theta} \) denote the probability of the above event, and let \( \eta_{\theta} \) be geometrically distributed with probability of success \( \rho_{\theta} \).

**Theorem 4.** With the definitions above, if \( E[S(1)]/E[A_{\theta}(1)] < 1 \), then \( X_{\theta}(\eta_{\theta} + n) \) is distributed according to \( \pi_{\theta} \) for all \( n \geq 0 \).

**Proof:** Under the condition put forward in the theorem, the queue is stable. We extend the definition of the process on \( Z \), which can be done without any harm as we have assumed that the service and inter-arrival times are i.i.d.

The proof follows from the following Lyapunov scheme.

Let \( X_{\theta}(n,0) \) denote the value of \( X_{\theta}(k) \) at \( n = 0 \) when the recursion is started at time \( -n \) at zero. Suppose that going backward in time from \( n = 0 \), the event \( \{A_{\theta}(n) \geq T\} \) occurs at \( \eta_{\theta} \) for the first time. Note that any sample path of \( \{X_{\theta}(n) : n \in \mathbb{Z}\} \) will regenerate at \( \eta_{\theta} \); i.e.,
\[
\lim_{n \to -\infty} X_{\theta}(n,0) = X_{\theta}(\eta_{\theta},0),
\]
and by stationarity, the limit on the above left hand side is equal to the stationary version of \( X_{\theta}(n) \). Hence, \( X_{\theta}(\eta_{\theta},0) \) is a perfect draw from the stationary distribution and switching back to a forward construction, which can be done without any harm in the i.i.d. case, proves the claim. \( \square \)

### 5. NUMERICAL EXAMPLES

#### 5.1 The M/M/1 Queue

For the M/M/1 queue, the inter-arrival distribution has mean \( 1/\theta \) and the service distribution has mean \( 1/\mu \), with \( \theta < \mu \). The stationary waiting time distribution, \( F_{\theta} \), is known for this case combining a mass at 0 with probability \( 1 - \rho \) and with the exponential(\( \mu - \theta \)) distribution when the stationary waiting times are positive
\[
F_{\theta}(t) = 1 - e^{-(\mu - \theta)t}.
\]
for \( \alpha > 1 - \rho \), and for \( \alpha < 1 - \rho \) it holds \( q_{\alpha}'(\theta) = 0 \).

To begin, we generate \( k \) samples of length \( m \) of the stationary waiting time, \( (W_{\theta,i}(n))_{n=1}^{m} \), with distribution function \( F_{\theta} \) and we keep the service times, \( (S(n))_{n=1}^{m} \) that generated the waiting times. For each \( 1 \leq i \leq k \), the estimator for the expectation of the derivative of \( q_{\alpha}'(\theta) \) is given by
\[
D_{i,m} = \frac{1}{m} \sum_{n=1}^{m} f_{A}(\max(W_{\theta,i}(n) + S(n) - X_{i}(m),0)),
\]
where \( f_{A}(t) = \theta e^{-\theta t} \) for \( t > 0 \) and the definition of \( X_{i}(m) \) will be provided later in the text.

For the estimation of the expectation in the numerator, we use the kept matrix of service times as well as the \( k \times m \) matrix of inter-arrival times, \( (A_{\theta,i}(n))_{n=1}^{m} \), which were also used in the generation of stationary waiting times. However, these matrices were permuted in opposite direction by a number of rows that is uniformly distributed between 1 and the integer value of \( k \times 2 \). For the MVD derivative of the distribution of the inter-arrival times, we let \( A_{\theta,i}(n) \) follow an exponential distribution with rate \( \theta \) and we let \( A_{\theta}(n) \) be gamma distributed such that \( A_{\theta}(n) \) is equal in distribution to the sum of two independent exponential \( \theta \) random variables.

Beginning at state \( W_{\theta,i}(0) \) for each row, we take \( A_{\theta,i}(1) = A_{\theta,i}(1) \) and \( A_{\theta,i}(1) = A_{\theta,i}(1) + A_{\theta,i}(1) \) where \( A_{\theta,i}(1) \) is a i.i.d. copy of \( A_{\theta,i}(1) \). Then for each row we use the same inter-arrival and service time realizations for both positive and negative phantom sequences to generate subsequent perturbed waiting times, \( (W_{\theta,i}(n), W_{\theta,i}(0)))_{n=1}^{\infty} \),
\[
(W_{\theta,i}(n), W_{\theta,i}(0)))_{n=1}^{\infty}, \quad (W_{\theta,i}(n), W_{\theta,i}(0)))_{n=1}^{\infty},
\]
until
\[
\tau_{\theta}(W_{\theta,i}) = \inf\{n \geq 1 : W_{\theta,i}^{+} = 0\},
\]
i.e., until for the first time the positive phantom sequence reaches zero. Note that at \( \tau_{\theta}(W_{\theta,i}) \) also the negative phantom hits zero, since \( W_{\theta,i}^{-}(n, W_{\theta,i}(0))) > W_{\theta,i}^{-}(n, W_{\theta,i}(0))) \)
with probability one, which stems from the fact that $A^+_\theta(1) > A^+_\theta(1)$ with probability one.

Once both phantom sequences hit zero for a given row, we begin a new perturbation estimate using the realizations $S_i(\tau_\theta(W_{\theta,i}))$, $A^+_\theta(\tau_\theta(W_{\theta,i}) + 1)$ as given above where $\tau_\theta(W_{\theta,i})$ is also an i.i.d. copy of an inter-arrival time where the new initial stationary waiting time is obtained sequentially from the sample $W_0(0) = [W_{\theta,i}(0)]_{i=1}^k$ to determine $W_{\theta,i}(\tau_\theta(W_{\theta,i}) + 1)$. For $n > m$, if both phantom sequences did not stop at $n = m$, we then draw i.i.d. realizations for $S_i(n), A_i(n)$, which again are common to both phantom sequences until both phantom sequences reach zero. Including $n = m$, for each row the event in which the positive and negative phantom sequence reach zero is defined as $\tau_{\theta,i}^+(m)$. For below, $\nu_{\theta,i}$ is the number of such cycles up to $n = \tau_{\theta,i}^+(m)$.

$$\tau_{\theta,i}^+(m) = \inf\{n \geq m : W_{\theta,i}^+ = 0\}.$$  

For a given row, $1 \leq i \leq k$, the estimator for the expectation of the numerator of the quantile sensitivity is given as

$$N_{i,m} = \frac{1}{\nu_{\theta,i}} \sum_{n=1}^{\nu_{\theta,i}} \left(1_{W_{\theta,i}^+(n,W_0(0) \leq X_i(m))} - 1_{W_{\theta,i}^+(n,W_0(0) \leq X_i(m))}\right).$$

Hence, for the M/M/1 queue, $q_{\theta}^*(\theta)$ can be expressed as

$$q^*_\theta(\theta) = \frac{1}{2} \sum_{i=1}^{\nu_{\theta,i}} \sum_{m=1}^{\nu_{\theta,i}} N_{i,m}.$$  

In this subsection, we use both $X_i(m) = q_{\theta}(\theta)$ and $X_i(m) = W_{\theta,i}(m)_{i=1}^m$, a common choice of order statistic to approximate the quantile. Since $f_\theta(t) = \theta e^{-\theta t}$, $t > 0$ is not continuous at $t = 0$, we choose $f_\theta(t) = 0$ to ensure the behavior of the estimator is increasing for increasing values of $\alpha$, the same behavior as the analytical value of $q^*_\theta(\theta)$. The parameters for this example are presented in Table 1 where $N_{o,e}$ is the number of estimates computed. In our results, we provide the mean and standard deviation for $\alpha = 0.90, 0.95$ and compare the results to the analytical (true) value of $q_{\theta}^*$ of the stationary waiting time distribution. Results are presented in Tables 2 where $X_i(m)$ is the analytical value of $q_{\theta}(\theta)$ and Table 3 where $X_i(m) = W_{\theta,i}(m)_{i=1}^m$. With other results not presented due to conciseness we observe moderate variance and moderate skewness to the right of the distribution of the estimator, especially for $k = m = 2^k, 2^k$ (though this tendency increases for increasing $\rho$) before becoming symmetric. The sources for the variance is due to discontinuity of the function $g(x) = f_\theta(\max(x,0))$ around zero. Specifically, for $x = W_{\theta,i}(n) + S_i(m) - X_i(m)$ slightly greater than zero $g(\cdot)$ has value $g(0^+) = \theta^*$ compared to $g(0^-) = 0$. As such, a precise estimate of the probability of an idle queue is needed to obtain an accurate result. Also, though there are variance reduction techniques present for the Measure Value Derivative component of the quantile sensitivity, some variance can be attributed to the difference in the number of times each perturbed sequence is above $q_\theta$ due to variations in $A_\theta(1)$, especially for large $\rho$. Thirdly, the final estimation is the quotient of two estimators, itself introducing variance. For increasing $m$, the rate of convergence to the true value is greater than $m^{-1/2}$ since for larger $m$ it allows for any given row, $1 \leq i \leq k$, more estimates to be conducted for the value of the introduced perturbations. This further reduces the error as opposed to including $m$ and $k$ for $D_{i,m}$, or adding more sequences for the numerator, $N_{i,m}$.

### Table 1. The values of parameters $\rho, \theta, \mu, k, m, N_{e,o}$ to determine $q^*_\theta(\theta)$ for the M/M/1 queue.

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$\theta$</th>
<th>$\mu$</th>
<th>$(k, m)$</th>
<th>$N_{e,o}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.45</td>
<td>4.50</td>
<td>(2^2, 2^2), (2^3, 2^3)</td>
<td>200</td>
</tr>
<tr>
<td>0.50</td>
<td>0.90</td>
<td>0.50</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.2 The Workload Process in the M/G/1/T Queue

For this example the procedure of estimation is similar to the previous subsection. The service time of the nth customer, $S(n)$, is equal in distribution to the sum of one plus a Poisson number distributed (with mean $\mu_1$) number of independent exponential distributions with rate $\mu_2$. From Wald’s theorem, the mean service time is given by $E[S(1)] = (1 + \mu_1)/\mu_2$. Again, the inter-arrival
distribution for each customer is $A_q(1) \sim \exp(\theta)$. The generation of the stationary waiting time realizations for each row is conducted via the Loynes scheme and the Lindley type recursion (see the proof of Theorem 5), generating $k$ sequences of length $m$ of the stationary waiting time sequence. In the splitting of the waiting time sequence, we choose the same decomposition for the Measure Valued Derivative for the exponential distribution as for the M/M/1 queue as well as the same generation of the positive and negative inter-arrival random variate to begin the perturbed sequences. We use the same i.i.d. matrix of inter-arrival and service time sequences as for the calculation of the stationary waiting times, permuting both matrices in opposite direction a uniformly distributed number of rows. However, we choose at random, uniformly, the new initial waiting time, $W_{\theta,i}(0)$, from $W_{\theta}(0)$ together with the corresponding queue at $n = 0$, $L_{\theta,i}(0)$ to re-initialize both phantom sequences. We do not continue to evaluate estimates for the introduced perturbation for any row once $n = m$. The parameters for this example are given in Table 4. Recall that $T$ is the maximum amount of work a server is willing to accept, and $N_{st}$ is the number of estimates conducted.

The values of the parameters in determining estimates $q'_\theta(\phi)$ for the M/G/1/T queue

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$\theta$</th>
<th>$\mu_1$</th>
<th>$\mu_2$</th>
<th>$T = 5$</th>
<th>$T = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.45</td>
<td>6</td>
<td>1/3</td>
<td>(5, 10)</td>
<td>(2', 2')</td>
</tr>
<tr>
<td>0.50</td>
<td>3</td>
<td>7/3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. The values of parameters $\rho, \theta, \mu_1, \mu_2, T, k, m, N_{st}$ to determine $q'_\theta(\phi)$ for the M/G/1/T queue

deviation for all combination of parameters are presented in Table 5. With additional quantile statistics for the sensitivity estimator, we observe the distribution of the estimator is weakly right-symmetric, less than the M/M/1 queue due to the sojourn time being bounded above. This distribution becomes symmetric for larger $k$ and $m$. The technique introduced should be asymptotically unbiased with bias from above due to the truncation introduced for the Measure Valued Derivative. Sources of variance are analogous to the previous example.

Estimation of $q'_\theta(\phi)$ for the stationary waiting time of an M/G/1/T queue: $k = m = 2^7$.

<table>
<thead>
<tr>
<th>$\rho = 0.10$</th>
<th>$\rho = 0.50$</th>
<th>$\rho = 0.90$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha = 0.90$</td>
<td>$\alpha = 0.90$</td>
<td>$\alpha = 0.90$</td>
</tr>
<tr>
<td>Mean</td>
<td>4.4505</td>
<td>6.3853</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>0.3983</td>
<td>0.5969</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\rho = 0.90$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha = 0.90$</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Std. Dev.</td>
</tr>
</tbody>
</table>

Table 5. $q'_\theta(\phi)$ for sensitivity of the quantile w.r.t to the inter-arrival parameter $\theta$ for the stationary waiting time of a M/G/1/T queue.

6. CONCLUSION

We presented an unbiased estimator for the sensitivity of quantiles of stationary waiting times. A topic of further research is a thorough numerical study of the estimator.

REFERENCES


