Check for updates

J. Appl. Probab. **61,** 1199–1219 (2024)
*doi:*10.1017/jpr.2023.120

INEQUALITIES BETWEEN TIME AND CUSTOMER AVERAGES FOR HNB(W)UE ARRIVAL PROCESSES

SHIGEO SHIODA,* ** AND
KANA NAKANO,* *** *Chiba University*

Abstract

We show that for arrival processes, the 'harmonic new better than used in expectation' (HNBUE) (or 'harmonic new worse than used in expectation', HNWUE) property is a sufficient condition for inequalities between the time and customer averages of the system if the state of the system between arrival epochs is stochastically decreasing and convex and the lack of anticipation assumption is satisfied. HNB(W)UE is a wider class than NB(W)UE, being the largest of all available classes of distributions with positive (negative) aging properties. Thus, this result represents an important step beyond existing result on inequalities between time and customer averages, which states that for arrival processes, the NB(W)UE property is a sufficient condition for inequalities.

Keywords: Time average; customer average; stochastic order; HNBUE; HWBUE; piecewise exponential distribution

2020 Mathematics Subject Classification: Primary 60K25; 68U35
Secondary 60K05; 60G10

1. Introduction

Measuring the bandwidth utilization of a communication link or the number of packets in an output buffer of a router, often called traffic measurement, is an important task for the operation and management of a communication network. Since traffic measurement is usually performed at regular intervals, it is equivalent to examining the time average of the network states (e.g. bandwidth utilization or the number of packets in an output buffer). Note that the time averages of the network states are measured by an observer outside the system. In a communication network, the average of the network states at packet arrival instants is called the customer average. The customer average, which is a state of the communication network experienced by user packets, is directly related to the quality of service (QoS) experienced by users. Packets from users do not always arrive at regular intervals, so the time average and the customer average are generally different. The formal definitions of time and customer averages are given in Section 2.

The relationship between the time and customer averages for a queueing system, especially the condition under which the time and customer averages are identical, has been extensively studied [4, 12, 14, 16, 18, 29]. For example, [29] showed that if customers arrive according

Received 19 August 2022; accepted 4 December 2023.
* Postal address: Graduate School of Engineering, Chiba University, 1-33 Yayoi, Inage, Chiba 263-8522, Japan.
** Email address: shioda@faculty.chiba-u.jp
*** Email address: gkn.bb.e@gmail.com

© The Author(s), 2024. Published by Cambridge University Press on behalf of Applied Probability Trust.

https://doi.org/10.1017/jpr.2023.120 Published online by Cambridge University Press

to a Poisson process and the lack of anticipation assumption holds, then the two averages are identical, which is widely known as the 'Poisson arrivals see time averages' (PASTA) property.

In general, time and customer averages are not identical. Several studies have attempted to identify the conditions under which time averages are larger (or smaller) than customer averages for a queue with a renewal arrival process. For example, [15, 17] showed that if the inter-arrival times for customers are 'new better than used in expectation' (NBUE) (or 'new worse than used in expectation', NWUE) for GI/G/1 queues, the time averages for some states of the system (e.g. workload) are larger (or smaller) than their customer averages. The same conclusion has been proved to hold for GI/G/c/K queues [13]. The relationship between time and customer averages based on the martingale approach in [29] was discussed in [20], which showed that the time averages for queueing systems are larger (or smaller) than their customer averages if the following three conditions hold: (i) the inter-arrival time for customers is NBUE (or NWUE), (ii) the states of the system observed at time t depend only on the past arrival epochs and what happened at those time points, and (iii) the sample path of the state of the system is decreasing between arrival epochs. It was shown in [26] that the second condition can be replaced with a weaker condition, referred to as the *lack of anticipation assumption*, and also that the state of the system between arrival epochs is not necessarily decreasing with respect to the sample path; it is sufficient that it be stochastically decreasing. The same results were derived in [21] using the 'coupling from the past' (CFTP) algorithm. Related results can also be found in [3, 6, 24, 28].

According to these existing studies, the conclusion that, for arrival processes, the NB(W)UE property is a sufficient condition for the above inequalities between the time and customer averages might seem to be the last word on the relationship between the time and customer averages. Here, we give an intuitive explanation for how the NB(W)UE property of the inter-arrival times yields the inequalities between the time and customer averages. Let us consider the change over time of the workload, or the amount of work remaining in the system, of a single-server queue. Assume that a customer arrives at the queue at time 0 and let $W(t)$ denote the workload at time t under the condition that no customer arrives in the interval $(0, t]$. Since each customer brings a certain amount of work, the workload rises at the instant a customer arrives. After that, it continuously decreases until the arrival of the next customer, that is, $W(t)$ is decreasing in the interval $(0, t]$. Now, let τ be a random variable describing the inter-arrival time for customers, and let $\tau^{(\mathrm{e})}$ denote a random variable following an equilibrium distribution of τ; $\tau^{(\mathrm{e})}$ corresponds to the time interval between an arbitrarily chosen time and the arrival time for a customer who arrived just before that arbitrarily chosen time. Note that $\mathbb{E}[W(\tau-)]$ is the so-called customer average and $\mathbb{E}[W(\tau^{(\mathrm{e})})]$ is the so-called time average. Since $W(t)$ is decreasing for $t > 0$, the customer-averaged workload should be smaller than the time-averaged workload if τ is larger than $\tau^{(\mathrm{e})}$ (Figure 1). In fact, if τ is NB(W)UE, then τ is stochastically larger (smaller) than $\tau^{(\mathrm{e})}$ in the usual stochastic-order sense (Definition 2.2), and thus the customer average is smaller (larger) than the time average, which is precisely the conclusion found in previous studies.

In this paper, we focus on the fact that if τ is the 'harmonic new better than used in expectation' (HNBUE) (or 'harmonic new worse than used in expectation', HNWUE), then τ is stochastically larger (smaller) than $\tau^{(\mathrm{e})}$ in the following sense: $\mathbb{E}[g(\tau)] \leq (\geq)\ \mathbb{E}[g(\tau^{(\mathrm{e})})]$ for all decreasing and convex functions $g(t)$. Since $W(t)$ is a decreasing and convex function of t for GI/G/c/K queueing systems, the customer average of the workload for GI/G/c/K queueing systems is smaller (larger) than the time average of the workload if τ is HNB(W)UE. The aim of the present paper is to formally show that if the state (e.g. the workload) of the system between

https://doi.org/10.1017/jpr.2023.120 Published online by Cambridge University Press

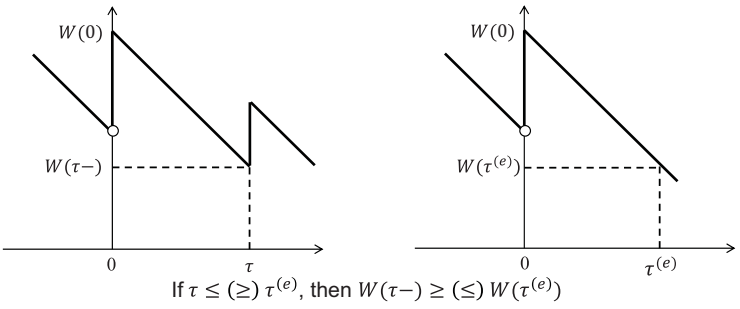


FIGURE 1. Comparison of $W(\tau-)$ and $W(\tau^{(e)})$.

arrival epochs is stochastically decreasing and convex and the lack of anticipation assumption is satisfied, then the arrival process having the HNB(W)UE property is a sufficient condition for the inequalities between the time and customer averages. HNB(W)UE is a wider class than NB(W)UE [19], being the largest of all available classes of distributions with positive (negative) aging properties [8]. In addition to this, as shown in this paper, there are a considerable number of systems in which the state of the system between arrival epochs is stochastically decreasing and convex. Thus, the result of this paper represents a small but important step beyond existing results on inequalities between time and customer averages.

The rest of this paper is organized as follows. Section 2 shows the notation and definitions used in the paper. Section 3 gives proofs of the main results. Section 4 applies the main results to derive the inequalities between the time and customer averages of the workload for GI/G/c/K queues, the number of customers in GI/M/c/K queues, the workload for GI+G/GI/1 queues, and the age process of the superposition of two independent point processes. Section 5 shows some numerical examples, which support the theoretical findings of this paper, using two piecewise exponential distributions; one is not NBUE but HNBUE, and the other is not NWUE but HNWUE.

2. Notation and definitions

2.1. Time and customer averages

Consider a real-valued stochastic process $\{X(t); t \in \mathbb{R}\}$ with left-continuous sample paths and a point process $\{T_n; n \in \mathbb{Z}\}$. $X(t)$ represents the state of a system just prior to time t (because of its left continuity) and $\{T_n\}$ represents arrivals of customers to the system. Customers are labeled following the standard convention such that $T_0 \le 0 < T_1$. We assume no batch arrivals; that is, $T_n < T_{n+1}$ for all $n \in \mathbb{Z}$. The 'state' of the system may represent the number of customers or the total workload for the queueing systems.

The time average of $\{X(t)\}$ from time 0 up to time t is defined by

$$T_t = \frac{1}{t}\int_0^t X(s)\,\mathrm{d}s,$$

and the customer average of $\{X(t)\}$ from time 0 to time t is defined by

$$C_t = \frac{1}{N(t)}\sum_{n=1}^{N(t)} X(T_n), \qquad N(t) \stackrel{\text{def}}{=} \sum_{n=1}^{\infty} \mathbf{1}_{T_n \le t},$$

https://doi.org/10.1017/jpr.2023.120 Published online by Cambridge University Press

where $\mathbf{1}_A$ is an indicator function, which is equal to 1 (0) if A is true (false). The analysis in this paper is conducted in the stationary and ergodic framework like [22, 26]; that is, $\{X(t)\}$ and $\{T_n; n \in \mathbb{Z}\}$ are assumed to be jointly stationary and ergodic under probability measure $\mathbb{P}$. Under this framework,

$$\lim_{t\to\infty} T_t = \lim_{t\to\infty} \frac{1}{t}\int_0^t X(s)\,\mathrm{d}s = \mathbb{E}[X(0)],$$

where $\mathbb{E}$ denotes the expectation with respect to $\mathbb{P}$, and

$$\lim_{t\to\infty} C_t = \lim_{t\to\infty} \frac{1}{N(t)}\sum_{n=1}^{N(t)} X(T_n) = \mathbb{E}^0[X(0)],$$

where $\mathbb{E}^0$ denotes the expectation with respect to $\mathbb{P}^0$, which is the Palm transformation [1] of $\mathbb{P}$ with respect to $\{T_n\}$. Similarly, the time-averaged probability that $\{X(t)\}$ is in any measurable set A is equal to $\mathbb{P}(X(0) \in A)$ because

$$\mathbb{P}(X(0) \in A) = \lim_{t\to\infty} \frac{1}{t}\int_0^t \mathbf{1}_{X(s)\in A}\,\mathrm{d}s,$$

and the customer-averaged probability that $\{X(t)\}$ is in A is equal to $\mathbb{P}^0(X(0) \in A)$ because

$$\mathbb{P}^0(X(0) \in A) = \lim_{t\to\infty} \frac{1}{N(t)}\sum_{n=1}^{N(t)} \mathbf{1}_{X(T_n)\in A}.$$

Thus, the issue of the inequalities between time and customer averages becomes one on the inequalities between the expectations by $\mathbb{P}$ and $\mathbb{P}^0$.

2.2. Lack of anticipation assumption

We define a process $X_0(t)$ ($t \in (0, \infty)$) to be the state of the above system at time t given that it is initiated at time 0 by a customer arrival and according to the Palm measure, so that $\mathbb{P}(X_0(0) \in A) = \mathbb{P}^0(X(0) \in A)$, but it continues without allowing any further arrivals to enter the system after time 0 [22].

The distribution function for the inter-arrival times for customers is denoted by $F_\tau(t) \stackrel{\text{def}}{=} \mathbb{P}^0(T_1 \le t)$, and τ denotes a random variable with distribution function F_τ. The inter-arrival time for customers is assumed to have a finite mean; that is, $\mathbb{E}^0[T_1] = \mathbb{E}[\tau] = 1/\lambda < \infty$. Throughout this paper, the following lack of anticipation assumption is made.

Definition 2.1. The state of the system $X(t)$ is said to satisfy the lack of anticipation assumption if, for any positive bounded real function h, $X(t)$ satisfies

$$\mathbb{E}^0[h(X(t)) \mid T_1 > t] = \mathbb{E}^0[h(X(t)) \mid T_1 = t] = \mathbb{E}[h(X_0(t))] \quad \text{for } t > 0. \tag{2.1}$$

According to [26], (2.1) is equivalent to the following condition:

$$\mathbb{E}^0[h(X(t)) \mid T_1 \ge t] = \mathbb{E}^0[h(X(t)) \mid T_1 = t] = \mathbb{E}[h(X_0(t))] \quad \text{for } t > 0. \tag{2.2}$$

The lack of anticipation assumption was introduced in [26] and used in [22]. This assumption intuitively states that, when customer 0 arrives at time 0 ($= T_0$), the state of the system at time $t > 0$ is conditionally independent of T_1, which is the arrival time for the next customer (customer 1), given that $T_1 > t$. The lack of anticipation assumption holds for a queue driven by a renewal arrival process [22].

https://doi.org/10.1017/jpr.2023.120 Published online by Cambridge University Press

2.3. Stochastic order

Since the main results of this paper are stated using the notions of usual stochastic order, (increasing) convex order, NB(W)UE, and HNB(W)UE, their definitions and related results used here are summarized below. Concerning the details of these definitions and related results, please see a textbook on stochastic orders, such as [19, 24, 25].

Definition 2.2. Let Z_1 and Z_2 be random variables with distribution functions F_{Z_1} and F_{Z_2}, respectively. Assume that Z_1 and Z_2 have finite means. Then, we say that

(i) Z_1 is less than Z_2 with respect to usual stochastic order (written $Z_1 \leq_{\mathrm{st}} Z_2$ or $F_{Z_1} \leq_{\mathrm{st}} F_{Z_2}$) if $\mathbb{E}[f(Z_1)] \leq \mathbb{E}[f(Z_2)]$ for all increasing functions f.

(ii) Z_1 is less than Z_2 with respect to convex order (written $Z_1 \leq_{\mathrm{cx}} Z_2$ or $F_{Z_1} \leq_{\mathrm{cx}} F_{Z_2}$) if $\mathbb{E}[f(Z_1)] \leq \mathbb{E}[f(Z_2)]$ for all convex functions f.

(iii) Z_1 is less than Z_2 with respect to increasing convex order (written $Z_1 \leq_{\mathrm{icx}} Z_2$ or $F_{Z_1} \leq_{\mathrm{icx}} F_{Z_2}$) if $\mathbb{E}[f(Z_1)] \leq \mathbb{E}[f(Z_2)]$ for all increasing convex functions f.

Proposition 2.1. *Let Z_1 and Z_2 be random variables. The following statements are equivalent:*

(i) $Z_1 \leq_{\mathrm{icx}} Z_2$*;*

(ii) $\mathbb{E}[(Z_1 - x)_+] \leq \mathbb{E}[(Z_2 - x)_+]$ *for all* $x \in \mathbb{R}$*,*

where

$$(x)_+ = \begin{cases} x & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

In the following two definitions, Z_t denotes a random variable with distribution function $F_{Z_t}(x) \stackrel{\mathrm{def}}{=} \mathbb{P}(Z \leq x + t \mid Z > t)$, which is the distribution of the residual lifetime of Z after time t.

Definition 2.3. If $\mathbb{E}[Z_t] \leq (\geq) \mathbb{E}[Z]$ for all $t > 0$, then we say that Z is NBUE (NWUE).

Definition 2.4. If

$$\frac{1}{t} \int_0^t \frac{\mathrm{d}s}{\mathbb{E}[Z_s]} \geq (\leq) \frac{1}{\mathbb{E}[Z]}$$

for all $t > 0$, then we say that Z is HNBUE (HNWUE).

The notion of HNBUE (HNWUE) was introduced in [23] and studied in [10]. Although HNBUE (HNWUE) seems to be less familiar than NBUE (NWUE), it will take a main role in this paper. Note that the expression in Definition 2.4 can be written as

$$\left\{ \frac{1}{t} \int_0^t \frac{\mathrm{d}s}{\mathbb{E}[Z_s]} \right\}^{-1} \leq (\geq)\, \mathbb{E}[Z],$$

which means that the integral harmonic mean of $\mathbb{E}[Z_t]$ is less than (greater than) or equal to $\mathbb{E}[Z]$ for all t if Z is HNBUE (HNWUE). It follows from Definitions 2.3 and 2.4 that if Z is NBUE (NWUE), then Z is HNBUE (HNWUE). That is, the HNBUE (HNWUE) class is larger than the NBUE (NWUE) class. The following result will be used in the next section.

https://doi.org/10.1017/jpr.2023.120 Published online by Cambridge University Press

Proposition 2.2. *Let Z be a random variable with mean a. The following statements are equivalent:*

(i) *Z is HNBUE (HNWUE);*

(ii) $Z^{(e)} \leq_{st} (\geq_{st}) \mathrm{Exp}(a)$*;*

(iii) $Z \leq_{cv} (\geq_{cv}) \mathrm{Exp}(a)$*,*

where $Z^{(e)}$ is a random variable with the equilibrium distribution of Z defined by

$$F_Z^{(e)}(x) \stackrel{\text{def}}{=} \frac{1}{\mathbb{E}[Z]} \int_0^x (1 - F_Z(t))\, \mathrm{d}t,$$

and $\mathrm{Exp}(a)$ *denotes an exponential random variable with mean a.*

In what follows, we say $\{T_n\}$ is HNBUE (HNWUE) if its inter-arrival time τ is HNBUE (HNWUE). According to Proposition 2.2, $\{T_n\}$ is HNBUE (HNWUE) if and only if

$$\frac{1}{\mathbb{E}^0[T_1]} \int_0^t \mathbb{P}^0(T_1 > s)\, \mathrm{d}s \leq (\geq)\ \exp\left\{-\frac{t}{\mathbb{E}^0[T_1]}\right\}.$$

2.4. Total time on test transform

For its use in Section 5, we describe the scaled total time on test (TTT) transform of a life distribution (that is, a distribution function with $F(0-) = 0$) [2, 10, 11].

Definition 2.5. Let $F(t)$ be a life distribution with mean λ^{-1}. The scaled TTT-transform, φ_F, of F is then defined by

$$\varphi_F(x) = \lambda \int_0^{F^{-1}(x)} (1 - F(t))\, \mathrm{d}t \quad \text{for } 0 \leq x \leq 1,$$

where $F^{-1}(x) = \inf\{x\colon F(t) \geq x\}$.

The scaled TTT-transform is defined for values of $x \in [0, 1]$, and the transformed values are also in [0,1]. This means that the scaled TTT-transform can be illustrated by a curve within the unit square. It is easy to see that $\varphi_F(x) = x$ if $F(t) = 1 - \mathrm{e}^{-\mu t}$. That is, the diagonal of the unit square corresponds to an exponential distribution.

The shape of the scaled TTT-transform of a life distribution shows the aging properties of the distribution [10, 11]. For example, whether a random variable is NBUE or NWUE can be seen from the scaled TTT-transform of its distribution function, as shown in the next theorem.

Theorem 2.1. ([10].) *A random variable Z with distribution function F is NBUE (NWUE) if and only if $\varphi_F(x) \geq (\leq) x$ for all $x \in [0, 1]$.*

Whether a random variable is HNBUE or HNWUE can also be seen from the scaled TTT-transform of its distribution function, as shown in the next theorem.

Theorem 2.2 ([10].) *A random variable Z with distribution function F is HNBUE (HNWUE) if and only if $\varphi_F(x) \geq (\leq) 1 - \exp\{-(1/\mathbb{E}[Z])F^{-1}(x)\}$ for all $x \in [0, 1]$.*

https://doi.org/10.1017/jpr.2023.120 Published online by Cambridge University Press

3. Main results

Proposition 3.1. *Let $g(t)$ and $h(t)$ be decreasing convex and increasing concave functions, respectively. If Z is HNBUE (HNWUE), then $\mathbb{E}[g(Z)] \le (\ge)\ \mathbb{E}[g(Z^{(\mathrm{e})})]$ and $\mathbb{E}[h(Z)] \ge (\le)\ \mathbb{E}[h(Z^{(\mathrm{e})})]$.*

Proof. Assume that Z is HNBUE. It follows from Proposition 2.2 that $Z \le_{\mathrm{cx}} \mathrm{Exp}(\mathbb{E}[Z])$ and $Z^{(\mathrm{e})} \le_{\mathrm{st}} \mathrm{Exp}(\mathbb{E}[Z])$. Thus, we obtain

$$\mathbb{E}[g(Z)] \le \mathbb{E}[g(\mathrm{Exp}(\mathbb{E}[Z]))] \le \mathbb{E}[g(Z^{(\mathrm{e})})],$$

where the first inequality follows from $Z \le_{\mathrm{cx}} \mathrm{Exp}(\mathbb{E}[Z])$ and the assumption that $g(t)$ is convex, and the second inequality follows from $Z^{(\mathrm{e})} \le_{\mathrm{st}} \mathrm{Exp}(\mathbb{E}[Z])$ and the assumption that $g(t)$ is decreasing. In the same way, we also obtain

$$\mathbb{E}[h(Z)] \ge \mathbb{E}[h(\mathrm{Exp}(\mathbb{E}[Z]))] \ge \mathbb{E}[h(Z^{(\mathrm{e})})].$$

The result when Z is HNWUE can be proved by reversing the inequalities at appropriate places in the above argument. □

Theorem 3.1. *If $\mathbb{E}[h(X_0(t))]$ is a decreasing convex function of t (>0) and $\{T_n\}$ is HNBUE (HNWUE), then $\mathbb{E}^0[h(X(0))] \le (\ge)\ \mathbb{E}[h(X(0))]$.*

Proof. We first assume that $\{T_n\}$ (and thus τ) is HNBUE. The stationarity and the lack of anticipation assumption yield

$$\begin{aligned}\mathbb{E}^0[h(X(0))] = \mathbb{E}^0[h(X(T_0))] = \mathbb{E}^0[h(X(T_1))] &= \int_0^\infty \mathbb{E}^0[h(X(t)) \mid T_1 = t]\, F_\tau(\mathrm{d}t) \\ &= \int_0^\infty \mathbb{E}[h(X_0(t))]\, F_\tau(\mathrm{d}t), \end{aligned} \tag{3.1}$$

where $\mathbb{E}^0[h(X(T_0))] = \mathbb{E}^0[h(X(T_1))]$ follows from the stationarity, and the second line follows from the lack of anticipation assumption (2.1) (or (2.2)). Letting $g(t) \overset{\mathrm{def}}{=} \mathbb{E}[h(X_0(t))]$, we have

$$\int_0^\infty \mathbb{E}[h(X_0(t))]\, F_\tau(\mathrm{d}t) = \int_0^\infty g(t)\, F_\tau(\mathrm{d}t) = \mathbb{E}[g(\tau)] \le \mathbb{E}[g(\tau^{(\mathrm{e})})], \tag{3.2}$$

where the last inequality follows from Proposition 3.1 and the assumption that $g(t)$ is a decreasing convex function of $t(>0)$. Note that $g(0)$ may be larger than $g(0+)$ but this does not matter, because $\tau^{(\mathrm{e})}$ does not have probability mass at $\tau^{(\mathrm{e})} = 0$. The last term of (3.2) can be expressed as

$$\begin{aligned}\mathbb{E}[g(\tau^{(\mathrm{e})})] &= \int_0^\infty \mathbb{E}[h(X_0(t))]\, F_\tau^{(\mathrm{e})}(\mathrm{d}t) \\ &= \int_0^\infty \mathbb{E}^0[h(X(t)) \mid T_1 > t]\, F_\tau^{(\mathrm{e})}(\mathrm{d}t) \\ &= \lambda \int_0^\infty \frac{\mathbb{E}^0[h(X(t))\mathbf{1}_{\{T_1>t\}}]}{1 - F_\tau(t)} (1 - F_\tau(t))\, \mathrm{d}t \\ &= \lambda \int_0^\infty \mathbb{E}^0[h(X(t))\mathbf{1}_{\{T_1>t\}}]\, \mathrm{d}t \\ &= \lambda \mathbb{E}^0\left[\int_0^\infty h(X(t))\mathbf{1}_{\{T_1>t\}}\, \mathrm{d}t \right] \\ &= \lambda \mathbb{E}^0\left[\int_0^{T_1} h(X(t))\, \mathrm{d}t \right] = \mathbb{E}[h(X(0))], \end{aligned} \tag{3.3}$$

https://doi.org/10.1017/jpr.2023.120 Published online by Cambridge University Press

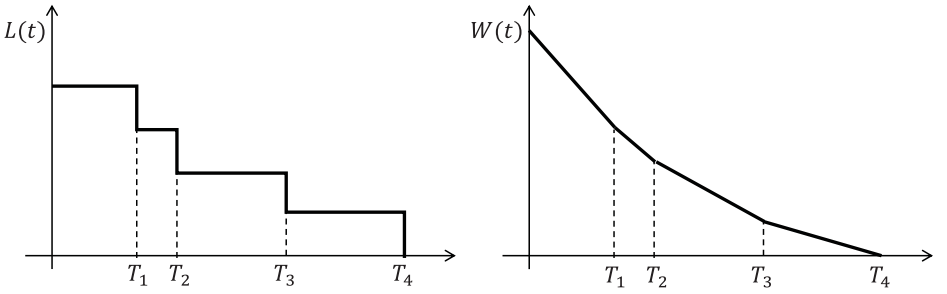


FIGURE 2. Change of workload over time ($c \geq 4$).

where the second line follows from the lack of anticipation assumption (2.1) and the last equality follows from the Palm inversion formula [1]. Combining (3.1), (3.2), and (3.3) completes the proof. The result when τ is HNWUE can be proved by reversing inequalities at appropriate places in the above argument. □

The arguments used in the proof of Theorem 3.1 yield the next result.

Corollary 3.1. *If* $\mathbb{E}[h(X_0(t))]$ *is an increasing concave function of* t (>0) *and* τ *is HNBUE (HNWUE), then* $\mathbb{E}^0[h(X(0))] \geq (\leq)\ \mathbb{E}[h(X(0))]$.

Remark 3.1. Assuming that $\mathbb{E}[h(X_0(t))]$ is decreasing and convex is equivalent to assuming that the state of a system between arrival epochs is stochastically decreasing and convex. In fact, as shown in Section 4, the workload of GI/G/c/K queues between arrival epochs is decreasing and convex with respect to sample path, and the number of customers in GI/M/c/K queues between arrival epochs is stochastically decreasing and convex.

4. Examples

4.1. Workload of GI/G/c/K queue

Consider a GI/G/c/K queue, where an arriving customer is assigned to an empty server or waits in a queue if all servers are busy. Once assigned to a server, a customer is served at the unit rate until completion. Let $W(t)$ and $L(t)$ respectively denote the workload and the number of customers in the queue at time t. The time- and customer-stationary distribution functions of the workload are respectively denoted by $F_W(x) \stackrel{\text{def}}{=} \mathbb{P}(W(0) \leq x)$ and $F_W^0(x) \stackrel{\text{def}}{=} \mathbb{P}^0(W(0-) \leq x)$, where $W(0-) \stackrel{\text{def}}{=} \lim_{t \downarrow 0} W(-t)$ is the left-hand limit of $W(0)$. Let $\{W_0(t); t \in [0, \infty)\}$ denote a process which represents the workload of a virtual queue without allowing any further arrivals to enter the system after time 0 and satisfies $\mathbb{P}(W_0(0) \leq x) = \mathbb{P}^0(W(0) \leq x)$ for all $x \geq 0$. Likewise, let $\{L_0(t); t \in [0, \infty)\}$ denote a process which represents the number of customers in the virtual queue and satisfies $\mathbb{P}(L_0(0) = x) = \mathbb{P}^0(L(0) = x)$ for all $x \geq 0$.

Let $L_s(t) \stackrel{\text{def}}{=} \min\{s, L_0(t)\}$. Since $L_s(t)$ is the number of busy servers for the virtual queue, $W_0(t)$ can be expressed as (Figure 2)

$$W_0(t) = W_0(0) - \int_0^t L_s(t)\, \mathrm{d}t. \tag{4.1}$$

Using this fact, we first show the following result.

https://doi.org/10.1017/jpr.2023.120 Published online by Cambridge University Press

Lemma 4.1. *If $\{T_n\}$ is HNBUE (HNWUE), then $F_W^0 \le_{\mathrm{icx}} (\ge_{\mathrm{icx}})\, F_W$.*

Proof. We define $g(t) \stackrel{\mathrm{def}}{=} \mathbb{E}[(W_0(t-) - x)_+]$. Note that $\mathbb{E}[(W_0(t-) - x)_+]$ is equal to $\mathbb{E}[(W_0(t) - x)_+]$ because there are no arrivals after time 0 in the virtual queue, and thus $W_0(t)$ is continuous for $t > 0$. It follows from (4.1) that, for $t > 0$,

$$\frac{\mathrm{d}}{\mathrm{d}t}(W_0(t) - x)_+ = -L_s(t)\mathbf{1}_{W_0(t)\ge x} \le 0.$$

In addition to this, $\frac{\mathrm{d}}{\mathrm{d}t}(W_0(t) - x)_+$ is increasing because $L_s(t)$ is decreasing. In summary, $W_0(t)$ is decreasing and convex with respect to sample path, and thus $g(t)$ is also decreasing and convex. Hence, it follows from Theorem 3.1 that if τ is HNBUE (HNWUE), then

$$\mathbb{E}^0[(W(0-) - x)_+] \le (\ge)\, \mathbb{E}[(W(0) - x)_+].$$

The above equality holds for all x, and thus the stated result follows from Proposition 2.1. □

We could have a stronger result for GI/G/1/K queues than for GI/G/c/K queues. To show this, let $F_{\mathrm{S}}^0(x) \stackrel{\mathrm{def}}{=} \mathbb{P}^0(W(0) \le x)$. Note that $F_{\mathrm{S}}^0(x)$ is the distribution function of the sojourn time for a customer when the service discipline is first in, first out (FIFO).

Lemma 4.2. *If $\{T_n\}$ is HNBUE (HNWUE) and $F_{\mathrm{S}}^0(x)$ is increasing and concave for $s > 0$ then, for GI/G/1/K queues, $F_W^0 \le_{\mathrm{st}} (\ge_{\mathrm{st}})\, F_W$.*

Proof. First, assume that $\{T_n\}$ is HNBUE. We define, for $t > 0$,

$$g(t; x) \stackrel{\mathrm{def}}{=} \mathbb{E}[\mathbf{1}_{W_0(t-)\le x}] = \mathbb{E}[\mathbf{1}_{W_0(t)\le x}].$$

We can see that

$$g(t; x) = \mathbb{P}(W_0(t) \le x) = \mathbb{P}(W_0(0) \le x + t) = \mathbb{P}^0(W(0) \le x + t) = F_{\mathrm{S}}^0(x + t), \tag{4.2}$$

where the second equality follows from $W_0(t)$ being continuous for $t > 0$ because there are no arrivals after time 0 in the virtual queue. The fact that $F_{\mathrm{S}}^0(x)$ is increasing and concave, together with (4.2), proves that $g(t; x)$ is increasing and concave in t. Hence, it follows from Corollary 3.1 that $\mathbb{E}^0[\mathbf{1}_{W(0-)\le x}] \ge \mathbb{E}[\mathbf{1}_{W(0)\le x}]$, which means that $F_W^0 \le_{\mathrm{st}} F_W$. Reversing the inequalities in the above arguments proves that $F_W^0 \ge_{\mathrm{st}} F_W$ when $\{T_n\}$ is HNWUE. □

Corollary 4.1. *If $\{T_n\}$ is HNBUE (HNWUE) then, for GI/M/1 queues, $F_W^0 \le_{\mathrm{st}} (\ge_{\mathrm{st}})\, F_W$.*

Proof. It is known [27] that

$$F_{\mathrm{S}}^0(x) = 1 - \mathrm{e}^{-\mu(1-\eta)x}, \tag{4.3}$$

where μ is the inverse of the mean service time for a customer, and η is the unique solution to the following equation for $z \in (0, 1)$:

$$z = \mathbb{E}[\mathrm{e}^{\mu(z-1)\tau}]. \tag{4.4}$$

Since the $F_{\mathrm{S}}^0(x)$ in (4.3) is increasing and concave, applying Lemma 4.2 completes the proof. □

https://doi.org/10.1017/jpr.2023.120 Published online by Cambridge University Press

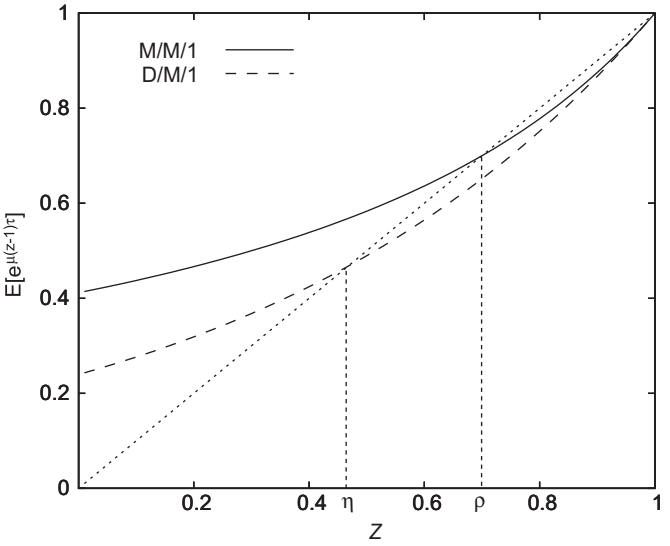


FIGURE 3. Comparison of $\mathbb{E}[e^{\mu(z-1)\tau}]$ for D/M/1 and M/M/1 queues.

Remark 4.1. Corollary 4.1 can be shown by the following elementary consideration. For GI/M/1 queues,

$$F_W^0(x) = 1 - \eta e^{-\mu(1-\eta)x}, \qquad F_W(x) = 1 - \rho e^{-\mu(1-\eta)x}, \tag{4.5}$$

where $\rho \stackrel{\text{def}}{=} \lambda/\mu$ and η is the solution to (4.4) in (0,1). Note that $F_W^0(x)$ is the distribution of the actual waiting time for a GI/M/1 queue for FIFO discipline and its expression was given in [27]; $F_W(x)$ is the distribution of virtual waiting time. Now, suppose that $\{T_n\}$ is HNBUE. It follows from Proposition 2.2 that $\tau \le_{\text{cx}} \text{Exp}(1/\lambda)$, and thus

$$\mathbb{E}[e^{\mu(z-1)\tau}] \le \mathbb{E}[e^{\mu(z-1)\text{Exp}(1/\lambda)}].$$

Since η is the intersection of $f(z) = \mathbb{E}[e^{\mu(z-1)\tau}]$ and $f(z) = z$, we see that $\eta \le \rho$ when $\mathbb{E}[e^{\mu(z-1)\tau}] \le \mathbb{E}[e^{\mu(z-1)\text{Exp}(1/\lambda)}]$. As an example, we compare $\mathbb{E}[e^{\mu(z-1)\tau}]$ when $\tau = 1/\lambda$ (D/M/1 queue: solid curve) and $\mathbb{E}[e^{\mu(z-1)\tau}]$ when $\tau = \text{Exp}(1/\lambda)$ (M/M/1 queue: dashed curve) in Figure 3. It follows from (4.5) that, if $\eta \le \rho$, then $F_W^0(x) \ge F_W(x)$ for all $x \ge 0$. Thus, if $\{T_n\}$ is HNBUE, then $F_W^0 \le_{\text{st}} F_W$. We can also obtain the desired result when $\{T_n\}$ is HNWUE in a similar way.

4.2. Number of customers of GI/M/*c*/*K* queue

Consider a GI/M/*c*/*K* queue where the mean service time for a customer is $1/\mu$. The time- and customer-stationary distribution functions of the number of the customers in the queues are respectively denoted by

$$F_L(x) \stackrel{\text{def}}{=} \mathbb{P}(L(0) \le x), \qquad F_L^0(x) \stackrel{\text{def}}{=} \mathbb{P}^0(L(0-) \le x),$$

where $L(0-) \stackrel{\text{def}}{=} \lim_{t \downarrow 0} L(-t)$ is the left-hand limit of $L(0)$. Let $D_0(t)$ denote the number of customers departed from the virtual queue (see Section 4.1) during $(0, t]$. $L_0(t)$ can be expressed as

https://doi.org/10.1017/jpr.2023.120 Published online by Cambridge University Press

$L_0(t) = L_0(0) - D_0(t)$. Note that $D_0(t)$ admits the $\mathscr{F}_t$-predictable stochastic intensity $\mu L_s(t-)$, where $\mathscr{F}_t$ is the history [1] of the virtual queue up to time t and the arrival process up to time 0. It follows from a property of stochastic intensity [1] that

$$\mathbb{E}[D_0(t) \mid \mathscr{F}_0] = \mathbb{E}\left[\int_0^t \mu L_s(u-)\,\mathrm{d}u\right]. \tag{4.6}$$

Lemma 4.3. *If $\{T_n\}$ is HNBUE (HNWUE), then $F_L \le_{\mathrm{icx}} (\ge_{\mathrm{icx}})\ F_L^0$.*

Proof. We define $g(t) \stackrel{\mathrm{def}}{=} \mathbb{E}[(L_0(t-) - x)_+]$ and $g_A(t) \stackrel{\mathrm{def}}{=} \mathbb{E}[(L_0(t-) - x)_+ \mid \mathscr{F}_0]$. It follows from (4.6) that

$$\frac{\mathrm{d}}{\mathrm{d}t} g_A(t) = -E\big[\mu L_s(t-)\mathbf{1}_{L_0(t)\ge x} \mid \mathscr{F}_0\big] \le 0.$$

Thus, $g_A(t)$ is decreasing. Since $g(t) = \mathbb{E}[g_A(t)]$, $g(t)$ is also decreasing. In addition to this, $\frac{\mathrm{d}}{\mathrm{d}t} g_A(t)$ is increasing because $L_s(t)$ is decreasing. Thus, $g_A(t)$ and therefore also $g(t)$ are convex for $t > 0$. Hence, if $\{T_n\}$ is HNBUE (HNWUE), then

$$\mathbb{E}^0[(L(0-) - x)_+] \le (\ge)\ \mathbb{E}[(L(0) - x)_+].$$

This inequality holds for all x, and thus the stated result follows from Proposition 2.1. □

Next, consider GI/M/1 queues.

Corollary 4.2. *If $\{T_n\}$ is HNBUE (HNWUE) then, for GI/M/1 queues, $F_L^0 \le_{\mathrm{st}} (\ge_{\mathrm{st}})\ F_L$.*

Proof. Letting

$$\mathbb{P}(k, t) \stackrel{\mathrm{def}}{=} \mathbb{E}[\mathbf{1}_{L_0(t-)\le k}] = \mathbb{P}(L_0(t-) \le k),$$
$$\mathbb{P}(k, t) \stackrel{\mathrm{def}}{=} \mathbb{E}[\mathbf{1}_{L_0(t-)=k}] = \mathbb{P}(L_0(t-) = k),$$

it follows that $\frac{\mathrm{d}}{\mathrm{d}t} p(k, t) = \mu(p(k+1, t) - p(k, t))$ and $\frac{\mathrm{d}}{\mathrm{d}t} p(0, t) = \mu p(1, t)$. Hence,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}(k, t) = \sum_{l=0}^{k} \frac{\mathrm{d}}{\mathrm{d}t} p(l, t) = \mu p(1, t) + \sum_{l=1}^{k} \mu(p(k+1, t) - p(k, t)) = \mu p(k+1, t) \ge 0, \tag{4.7}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbb{P}(k, t) = \mu \frac{\mathrm{d}}{\mathrm{d}t} p(k+1, t) = \mu^2(p(k+2, t) - p(k+1, t)). \tag{4.8}$$

It can be seen that

$$\begin{aligned} p(k, t) &= \sum_{n=0}^{\infty} \Pr\{\{L(0) = k+n\} \cap \{n \text{ customers have departed during } [0,t)\}\} \\ &= \sum_{n=0}^{\infty} p(k+n-1, 0)\frac{(\mu t)^n}{n!}\mathrm{e}^{-\mu t} = (1-\eta)\eta^{k-1}\mathrm{e}^{-\mu t(1-\eta)}, \end{aligned}$$

where we use $p(k, 0) = \mathbb{P}(L_0(0-) = k) = (1-\eta)\eta^k$ [27]. Substituting the above equality into (4.8) yields

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbb{P}(k, t) = -\mu^2(1-\eta)^2\eta^k\mathrm{e}^{-\mu t(1-\eta)} \le 0. \tag{4.9}$$

https://doi.org/10.1017/jpr.2023.120 Published online by Cambridge University Press

As (4.7) and (4.9) mean that $p(k, t)$ is increasing and concave with respect to t, it follows from Corollary 3.1 that $\mathbb{E}^0[\mathbf{1}_{L(0-)\le k}] \ge (\le)\ \mathbb{E}[\mathbf{1}_{L(0)\le k}]$, or $\mathbb{P}^0(L(0-) > k) \le (\ge)\ \mathbb{P}(L(0) > k)$, which completes the proof. □

Remark 4.2. Consider a GI/GI/1/K queue and let $\{T_n^{\mathrm{D}} : n \in \mathbb{Z}\}$ denote the departure times for the customers from the queue. The departure times are assumed to be labeled such that $T_0^{\mathrm{D}} \le 0 < T_1^{\mathrm{D}}$. Let $F_\sigma^{(\mathrm{e})}$ denote the equilibrium distribution of the service time for the customers. If the remaining service time for a customer in service at customer arrival instants follows the equilibrium distribution $F_\sigma^{(\mathrm{e})}$, that is,

$$\mathbb{P}^0(T_1^{\mathrm{D}} \le x \mid L_s(0+) > 0) = F_\sigma^{(\mathrm{e})}(x), \tag{4.10}$$

then (4.6) holds because $D_1, D_2, \ldots$ becomes a stationary renewal process and thus $D_0(t) = \mu t$ [5]. This argument suggests that Lemma 4.3 holds for GI/GI/1/K queues if (4.10) holds, and the same argument can also be applied to GI/GI/c/K queues. However, note that (4.10) does not hold in general, and thus this argument does not prove that Lemma 4.3 holds for all GI/GI/1/K queues. Nevertheless, as shown in Section 5, we have numerically found that $\mathbb{E}^0[L] \le (\ge)\ \mathbb{E}[L]$ seems to hold for GI/GI/1 queues if the arrival process is HNBUE (HNWUE) even for non-exponential service time distributions. This result implies that (4.10) approximately holds for most of the conditions of these queues.

Remark 4.3. For GI/M/1 queues,

$$\mathbb{P}^0(L(0-) > k) = \eta^{k+1}, \qquad k = 0, 1, \ldots,$$
$$\mathbb{P}(L(0) > k) = \rho\eta^k, \qquad k = 0, 1, \ldots,$$

where η is the solution of (4.4) in (0,1). If τ is HNBUE (HNWUE), then $\eta \le (\ge)\ \rho$, as shown in Remark 4.1. From this fact, Corollary 4.2 also follows.

4.3. Workload of GI+G/GI/1 queue

Next, we consider an example given in [26] where the superposition of two stationary and ergodic arrival processes, $\{T_n^i; n \in \mathbb{Z}\}$, $i = 0, 1$, is fed into a single-server queue. The two arrival processes are independent. We assume that $\{T_n^0\}$ is a renewal arrival process; that is, its inter-arrival times are independent and identically distributed.

Lemma 4.4. *If $\{T_n^0\}$ is HNBUE (HNWUE) then, for GI+G/GI/1 queues,*

$$\mathbb{E}^0[W(0-)] \le (\ge)\ \mathbb{E}[W(0)].$$

Proof. In the proof, we call customers arriving at $\{T_n^i\}$, $i = 0, 1$, type-i customers, and assume that type-1 customers have preemptive priority over type-0 customers. Note that this assumption is not at all essential, because the concern of this lemma is the total workload. We let $W(t) = W^0(t) + W^1(t)$, where $W^0(t)$ ($W^1(t)$) is the workload due to type-0 (type-1) customers. Note that $W^1(t)$ is independent of the arrival process $\{T_n^0\}$. Because of this, the statistics of $W^1(t)$ under P^0, which is the Palm probability measure with respect to $\{T_n^0\}$, and P are the same, as mentioned in [26]. For $t > 0$, we let $g(t) \stackrel{\mathrm{def}}{=} \mathbb{E}[W_0(t-)]$. Note that $W_0(t-)$ is expressed as the sum of $W_0^0(t-)$ and $W_0^1(t-)$, where $W_0^0(t)$ ($W_0^1(t)$) is the workload in the virtual queue due to type-0 (type-1) customers. Since $\mathbb{E}[W_0^1(t-)] = \mathbb{E}^0[W^1(t-)]$ (from the definition of W_0)

https://doi.org/10.1017/jpr.2023.120 Published online by Cambridge University Press

and $\mathbb{E}^0[W^1(t-)] = \mathbb{E}[W^1(t-)]$ ($W^1(t)$ under P^0 is statistically the same as under P), it follows that

$$g(t) = \mathbb{E}[W_0^0(t-)] + \mathbb{E}[W^1(t-)] = \mathbb{E}[W_0^0(t)] + \mathbb{E}[W^1(0)],$$

where the second equality follows from the stationarity of $W^1(t)$ and the continuity of $W_0^0(t)$ for $t > 0$. Since $\frac{\mathrm{d}}{\mathrm{d}t} W_0^0(t) = -1$ if $W_0^1(t) = 0$ with respect to sample path, it follows that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} g(t) &= -\mathbb{E}[\mathbf{1}_{W_0^0(t)=0} \mathbf{1}_{W_0^1(t)>0}] \\
&= -\mathbb{E}[\mathbb{E}[\mathbf{1}_{W_0^1(t)=0} \mathbf{1}_{W_0^0(t)>0} \mid \mathbf{1}_{W_0^1(t)>0}]] \\
&= -\mathbb{E}[\mathbf{1}_{W_0^1(t)=0} \mathbf{1}_{W_0^0(t)>0} \mid W_0^1(t) > 0] \mathbb{P}(W_0^1(t) > 0) \\
&\quad - \mathbb{E}[\mathbf{1}_{W_0^1(t)=0} \mathbf{1}_{W_0^0(t)>0} \mid W_0^1(t) = 0] \mathbb{P}(W_0^1(t) = 0) \\
&= -\mathbb{E}[\mathbf{1}_{W_0^1(t)=0} \mathbf{1}_{W_0^0(t)>0} \mid W_0^1(t) = 0] \mathbb{P}(W_0^1(t) = 0) \\
&= -\mathbb{E}[\mathbf{1}_{W_0^0(t)>0} \mid W^1(t) = 0] \mathbb{P}(W^1(0) = 0) \le 0, \qquad (4.11)
\end{aligned}$$

where the last equality follows from the fact that $\mathbb{P}(W_0^1(t) = 0) = \mathbb{P}^0(W^1(t) = 0) = \mathbb{P}(W^1(t) = 0)$ and the stationarity of $W^1(t)$ under P. From (4.11), we see that $g(t)$ is decreasing. In addition to this, $\frac{\mathrm{d}}{\mathrm{d}t} g(t)$ is increasing because $W_0^0(t)$ is decreasing. These arguments give us the conclusion that $g(t)$ is decreasing and convex. Hence, it follows from Theorem 3.1 that if $\{T_n^0\}$ is HNBUE (HNWUE), then $\mathbb{E}^0[W(0-)] \le (\ge)\ \mathbb{E}[W(0)]$. □

4.4. Age process for superposition of two independent point processes

Finally, we consider another example given in [26]. Let $\{T_n^i; n \in \mathbb{Z}\}$, $i = 0, 1$, be two point processes that are assumed to be jointly stationary and ergodic under the probability measure $\mathbb{P}$. These two point processes are independent. Let $\mathbb{P}^i$ denote the Palm probability measure with respect to the point process $\{T_n^i\}$ and $\mathbb{E}^i$ be the corresponding expectation. Let $\{R_n; n \in \mathbb{Z}\}$ denote the superposition of the two point processes, and define the following 'age' process:

$$A(t) \stackrel{\mathrm{def}}{=} \sum_{n=-\infty}^{\infty} \mathbf{1}_{R_n < t \le R_{n+1}} (t - R_n).$$

Note that $A(t)$ is left-continuous. The inter-arrival times for point processes $\{T_n^i; n \in \mathbb{Z}\}$ ($i = 0, 1$) are not necessarily independent of each other.

Lemma 4.5. *If $\{T_n^0\}$ is HNBUE (HNWUE) then, for all x,*

$$\mathbb{E}[(A(0) - x)_+] \le (\ge)\ \mathbb{E}^0[(A(0) - x)_+].$$

That is, the age at the arrival instants of the 0th point process is greater than the age at an arbitrary instant with respect to the increasing convex order.

Proof. Define $g(t) \stackrel{\mathrm{def}}{=} \mathbb{E}^0[(A(t) - x)_+ \mid T_1^0 = t]$. Let $F_{\tau_1}(t) \stackrel{\mathrm{def}}{=} P^1(T_1^1 \le x)$ and $F_{\tau_1}^{(\mathrm{e})}(t)$ denote its equilibrium distribution, that is,

$$F_{\tau_1}^{(\mathrm{e})}(t) = \frac{1}{\mathbb{E}^1[T_1^1]} \int_0^t (1 - F_{\tau_1}(s))\, \mathrm{d}s.$$

https://doi.org/10.1017/jpr.2023.120 Published online by Cambridge University Press

It can be seen [26] that

$$\begin{aligned} g(t) &= \int_0^t (s-x)_+ \,\mathrm{d}F_{\tau_1}^{(\mathrm{e})}(s) + (t-x)_+\bigl(1-F_{\tau_1}^{(\mathrm{e})}(t)\bigr) \\ &= -\bigl[(s-x)_+\bigl(1-F_{\tau_1}^{(\mathrm{e})}(s)\bigr)\bigr]_0^t + \int_x^t \bigl(1-F_{\tau_1}^{(\mathrm{e})}(s)\bigr)\,\mathrm{d}s + (t-x)_+\bigl(1-F_{\tau_1}^{(\mathrm{e})}(t)\bigr) \\ &= \int_x^t \bigl(1-F_{\tau_1}^{(\mathrm{e})}(s)\bigr)\,\mathrm{d}s. \end{aligned} \tag{4.12}$$

Note that the first term on the right-hand side of the first line is the expectation of age conditioned on the first arrival of the first point process after time 0 occurring before time t, and the second term is the expectation of age conditioned on the first arrival of the first point process after time 0 being after time t. It follows from (4.12) that

$$\frac{\mathrm{d}}{\mathrm{d}t}g(t) = 1 - F_{\tau_1}^{(\mathrm{e})}(t) \geq 0, \qquad \frac{\mathrm{d}^2}{\mathrm{d}t^2}g(t) = -\frac{\mathrm{d}}{\mathrm{d}t}F_{\tau_1}^{(\mathrm{e})}(t) = -\frac{1-F_{\tau_1}(t)}{\mathbb{E}^1[T_1^1]} \leq 0.$$

This means that $g(t)$ is increasing and concave, and thus the stated result follows from Proposition 2.1. □

5. Numerical examples

In this section, we show two numerical examples concerning the inequalities between customer and time averages.

5.1. Utilization of a GI/M/1 queue

In this subsection, we numerically investigate the relationship between customer-averaged and time-averaged utilization of a GI/M/1 queue. We also show how this relationship is related to the coefficient of variation of inter-arrival times for customers. The utilization is the probability that a non-zero workload remains in the queue. The time-averaged utilization is thus equal to $\mathbb{P}(W(0) > 0)$ and the customer-averaged utilization (utilization at a customer arrival instant) is equal to $\mathbb{P}^0(W(0-) > 0)$. As shown in Remark 4.1, for a GI/M/1 queue, $\mathbb{P}(W(0) > 0)$ and $\mathbb{P}^0(W(0-) > 0)$ are respectively given as $\mathbb{P}(W(0) > 0) = \rho = \lambda/\mu$ and $\mathbb{P}^0(W(0-) > 0) = \eta$, where μ is the inverse of the mean service time, λ is the inverse of the mean inter-arrival time for customers, and η is the solution of $z = \mathbb{E}[\mathrm{e}^{\mu(z-1)\tau}]$ in (0,1). Now assume that the inter-arrival time for customers is distributed according to the nth Erlang distribution, where $\mathbb{E}[\mathrm{e}^{\mu(z-1)\tau}]$ is given as

$$\mathbb{E}[\mathrm{e}^{\mu(z-1)\tau}] = \left(\frac{n\rho}{n\rho+1-z}\right)^n.$$

In Figure 4(a), we show the customer-averaged utilization of the GI/M/1 queue, in which the inter-arrival time for customers is distributed according to the nth Erlang distribution, by changing n from 1 to 25. The time-averaged utilization is set to 0.5, 0.7, or 0.9. The horizontal axis of the figure shows the coefficient of variation of inter-arrival times for customers, instead of showing the values of n. Note that the coefficient of variation of the nth Erlang distribution is equal to $n^{-1/2}$. Because the Erlang distribution is HNBUE, it follows from Corollary 4.1 that

$$\mathbb{P}^0(W(0-) > 0)\bigl[= 1 - F_W^0(0)\bigr] \leq \mathbb{P}(W(0) > 0)\bigl[= 1 - F_W(0)\bigr].$$

https://doi.org/10.1017/jpr.2023.120 Published online by Cambridge University Press

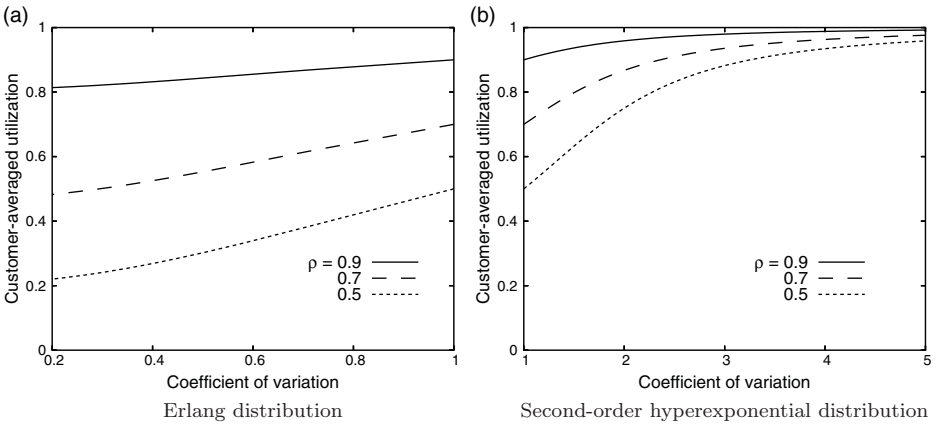


FIGURE 4. Utilization of a GI/M/1 at queue the instant of packet arrival.

Thus, the customer-averaged utilization is smaller than (or equal to) the time-averaged utilization, which is consistent with the results in Figure 4(a). Figure 4(a) also shows that the customer-averaged utilization becomes smaller as the coefficient of variation of inter-arrival times for customers becomes smaller. This result can be explained using theory as follows. The proof of Corollary 4.1 (and the proof of Lemma 4.2) shows that, for a GI/M/1 queue, $g(t;x) \stackrel{\text{def}}{=} \mathbb{E}^0[\mathbf{1}_{W_0(t)\le x}]$ is a concave function of t. Thus, $\mathbb{E}^0[\mathbf{1}_{W_0(t)>0}] = 1 - g(t, 0)$ is a convex function of t. Now, let $\tau^{\mathrm{Er}(n)}$ denote a random variable following the nth Erlang distribution. If $\tau^{\mathrm{Er}(n_1)}$ and $\tau^{\mathrm{Er}(n_2)}$ have the same mean and $n_2 \le n_1$, $\tau^{\mathrm{Er}(n_1)} \le_{\mathrm{cv}} \tau^{\mathrm{Er}(n_2)}$. Thus,

$$\mathbb{E}^0\big[\mathbf{1}_{W_0(\tau^{\mathrm{Er}(n_1)})>0}\big] \le \mathbb{E}^0\big[\mathbf{1}_{W_0(\tau^{\mathrm{Er}(n_2)})>0}\big],$$

showing that when the inter-arrival time for customers follows the nth Erlang distribution, the customer-averaged utilization decreases as n increases. Since the coefficient of variation of the nth Erlang distribution becomes smaller as n increases, the customer-averaged utilization decreases as the coefficient of variation of inter-arrival times decreases.

Next, assume that the inter-arrival time for customers is distributed according to a second-order hyperexponential distribution with mean $1/\lambda$, whose distribution function is given as

$$P(\tau \le t) = \begin{cases} \dfrac{1}{n+1}(1 - \mathrm{e}^{-\lambda t/n}) + \dfrac{n}{n+1}(1 - \mathrm{e}^{-n\lambda t}), & t \ge 0, \\ 0, & t < 0. \end{cases} \tag{5.1}$$

Note that the distribution in (5.1) is parametrized with n, where n is not necessarily an integer, and its coefficient of variation is equal to $\sqrt{(2n^2 - 3n + 2)/n}$. Under the distribution in (5.1), $\mathbb{E}[\mathrm{e}^{\mu(z-1)\tau}]$ is given as

$$\mathbb{E}[\mathrm{e}^{\mu(z-1)\tau}] = \frac{\rho}{n+1}\left(\frac{1}{\rho + n(1-z)} + \frac{n^2}{n\rho + 1 - z}\right).$$

In Figure 4(b), we show the customer-averaged utilization of the GI/M/1 queue, in which the distribution of the inter-arrival time for packets is given by (5.1), by increasing n from 1. The time-averaged utilization is set to 0.5, 0.7, or 0.9. As in Figure 4(a), the horizontal axis of the figure shows the coefficient of variation of inter-arrival times for customers.

https://doi.org/10.1017/jpr.2023.120 Published online by Cambridge University Press

Because the hyperexponential distribution is HNWUE, it follows from Corollary 4.1 that $\mathbb{P}^0(W(0-) > 0) \geq \mathbb{P}(W(0) > 0)$. Thus, the customer-averaged utilization is larger than the time-averaged utilization, which is consistent with the results in Figure 4(b). Figure 4(b) also shows that the customer-averaged utilization becomes larger as the coefficient of variation of inter-arrival times for customers becomes larger, which can also be confirmed via theory by an argument similar to that for the Erlang distribution.

5.2. Piecewise exponential distributions

If a random variable Z has a piecewise-constant hazard function, then Z is called a piecewise exponential random variable [7]. The distribution function of an n-piece exponential random variable Z with cut points $t_0 = 0 < t_1 < \cdots < t_n = \infty$ is

$$F_Z(t) = 1 - \sum_{k=1}^{n} c_k \mathrm{e}^{-\lambda_k t} \mathbf{1}_{t_{k-1} < t \leq t_k}, \tag{5.2}$$

where $c_k = \prod_{l=1}^{k} \mathrm{e}^{(\lambda_l - \lambda_{l-1}) t_{l-1}}$. Note that its hazard function $h_Z(t)$ and its expectation $\mathbb{E}[Z]$ are given by

$$h_Z(t) = \sum_{k=1}^{n} \lambda_i \mathbf{1}_{t_{k-1} < t \leq t_k}, \qquad \mathbb{E}[Z] = \sum_{k=1}^{n} \frac{c_k}{\lambda_k} \left(\mathrm{e}^{-\lambda_k t_{k-1}} - \mathrm{e}^{-\lambda_k t_k} \right).$$

The scaled TTT-transform of a piecewise exponential distribution is a piecewise linear function. In fact, the scaled TTT-transform of the n-piece exponential distribution (5.2) is given as

$$\varphi_{F_Z}(x) = \sum_{i=1}^{n} \left(\varphi_Z(x_{i-1}) + \frac{\lambda}{\lambda_i}(x - x_{i-1}) \right) \mathbf{1}_{x_{i-1} < x \leq x_i}, \qquad \lambda \stackrel{\text{def}}{=} \frac{1}{\mathbb{E}[Z]},$$

where $x_i \stackrel{\text{def}}{=} F_Z(t_i)$. Figure 5(a) shows the distribution function of the four-piece exponential random variable with parameters (5.3), and Figure 5(b) shows its scaled TTT-transform. According to Theorems 2.1 and 2.2, Figure 5(b) proves that this four-piece exponential random variable is not NBUE but HNBUE.

The set of piecewise exponential random variables includes those that are not NBUE (NWUE) but HNBUE (HNWUE). Klefsjö showed that a four-piece exponential random variable with the following parameters is not NBUE but HNBUE [9, 10]:

$$\begin{aligned} &t_1 = 0.359, \quad t_2 = 0.592, \quad t_3 = 1.662, \\ &\lambda_1 = 0.143, \quad \lambda_2 = 3.600, \quad \lambda_3 = 0.175, \quad \lambda_4 = 3.400. \end{aligned} \tag{5.3}$$

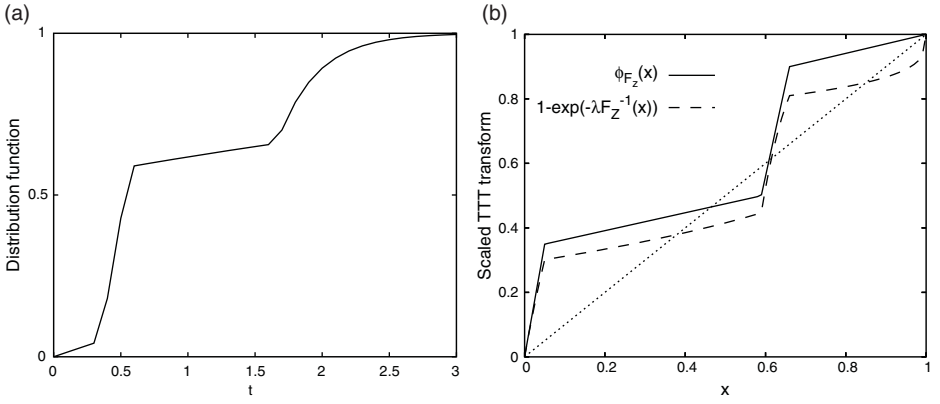
We found that a four-piece exponential random variable with the following parameters is not NWUE but HNWUE:

$$\begin{aligned} &t_1 = 0.2, \quad t_2 = 1.0, \quad t_3 = 2.0, \\ &\lambda_1 = 2.0, \quad \lambda_2 = 0.1, \quad \lambda_3 = 2.0, \quad \lambda_4 = 0.2. \end{aligned} \tag{5.4}$$

Figure 6(a) shows the distribution function of the four-piece exponential random variable with parameters (5.4), and Figure 6(b) shows its scaled TTT-transform, which proves that the four-piece exponential random variable with parameters (5, 4) is not NWUE but HNWUE.

Figure 7 shows the customer and time averages of the workload of a GI/GI/1 queue under the condition that the inter-arrival time for customers follows the four-piece exponential

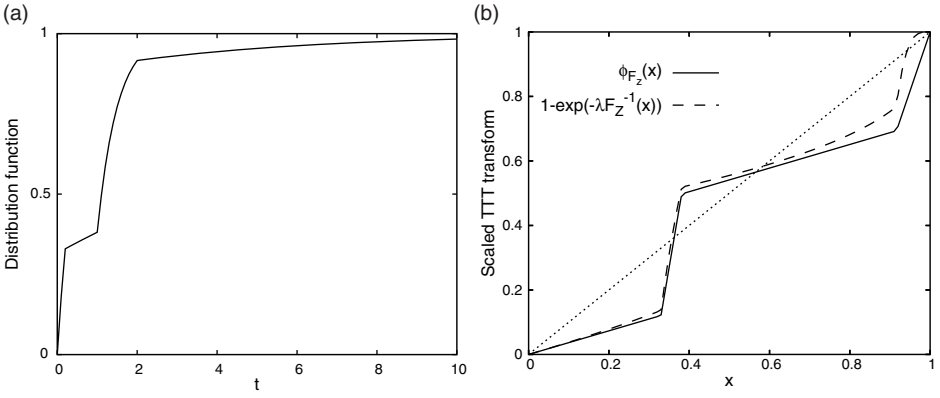
https://doi.org/10.1017/jpr.2023.120 Published online by Cambridge University Press



Distribution function

Scaled TTT-transform (solid curve $\varphi_{F_Z}(x)$, dotted curve $1-\exp\left(-\lambda F_Z^{-1}(x)\right)$)

FIGURE 5. Piecewise exponential distribution that is not NBUE but HNBUE.



Distribution function

Scaled TTT-transform (solid curve $\varphi_{F_Z}(x)$, dotted curve $1-\exp\left(-\lambda F_Z^{-1}(x)\right)$)

FIGURE 6. Piecewise exponential distribution that is not NWUE but HNWUE.

distribution with parameters (5.3). We considered three different service-time distributions: constant (GI/D/1), exponential distribution (GI/M/1), and the second-order hyperexponential distribution (GI/H_2/1) whose distribution function is

$$F(t)=\begin{cases}\frac{1}{4}(1-\mathrm{e}^{-\mu t/3})+\frac{3}{4}(1-\mathrm{e}^{-3\mu t}), & t\geq 0,\\ 0, & t<0.\end{cases}$$

The results in Figure 7(a) (GI/D/1) and Figure 7(c) (GI/H_2/1) were obtained from simulation, and the results in Figure 7(b) were obtained by theory (for the GI/M/1 queue). Figure 7 confirms the conclusion of Lemma 4.1 that the customer average of the workload is smaller than the time average when the arrival process is HNBUE. (The customer and time averages are very close when $\rho \geq 0.8$, and we tabulate these customer and time averages in tables in the

https://doi.org/10.1017/jpr.2023.120 Published online by Cambridge University Press

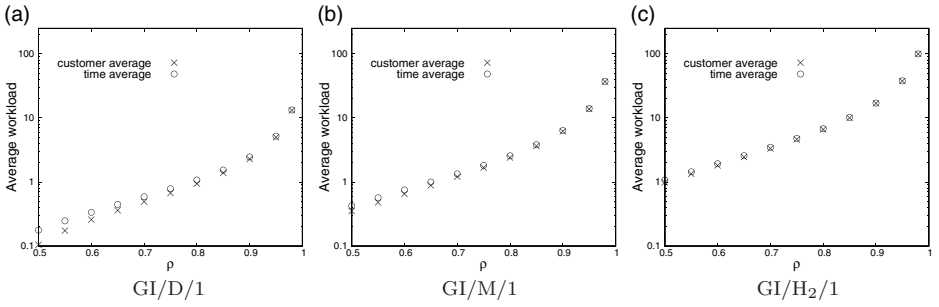


FIGURE 7. Average workload (arrival process is not NBUE but HNBUE).

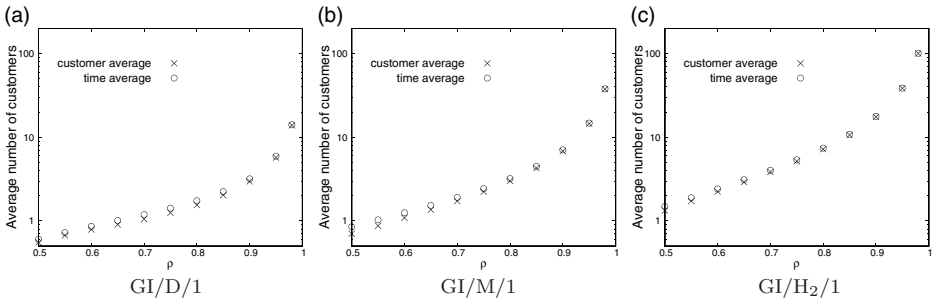


FIGURE 8. Average number of customers (arrival process is not NBUE but HNBUE).

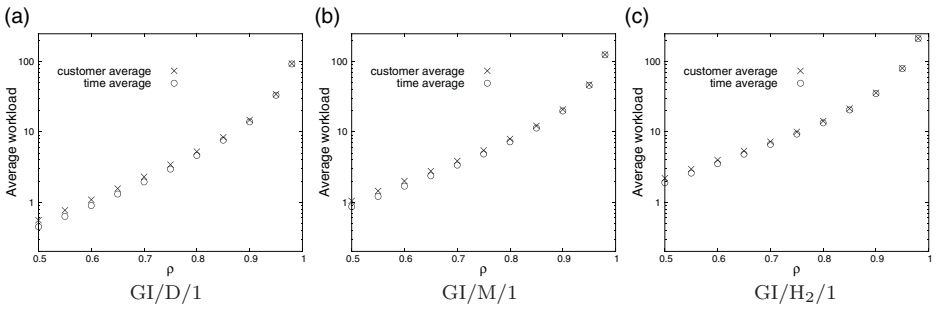


FIGURE 9. Average workload (arrival process is not NWUE but HNWUE).

Appendix.) Figure 8 compares the customer and time averages of the number of customers of a GI/GI/1 queue under the same conditions with the workload. Figure 8 shows that the customer average of the number of customers in the queue is smaller than the time average when the arrival process is HNBUE. Note that the inequalities between the customer and time averages of the number of customers are proved only for GI/M/1 queues (Lemma 4.3). Thus, these numerical examples imply that inequalities between the customer and time averages of the number of customers may hold for GI/GI/1 queues.

Figure 9. (workload) and Figure 10 (number of customers) show the customer and time averages of the GI/GI/1 queue under the condition that the inter-arrival time for customers

https://doi.org/10.1017/jpr.2023.120 Published online by Cambridge University Press

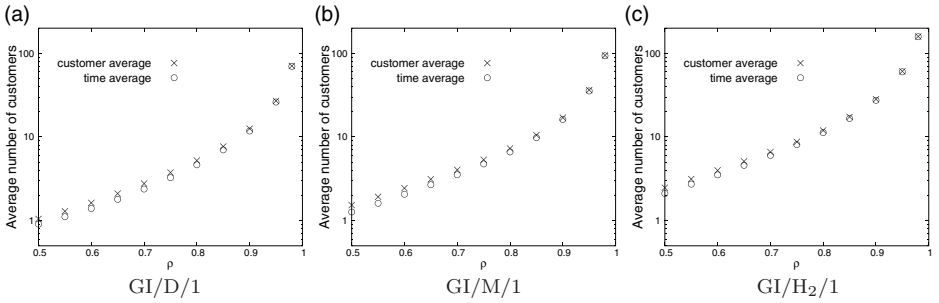


FIGURE 10. Average number of customers (arrival process is not NWUE but HNWUE).

follows a four-piece exponential distribution with parameters (5.4), which is not NWUE, but HNWUE. These figures numerically confirm that the customer average is larger than the time average when the arrival process is HNWUE.

Appendix A. Average workload and number of customers in numerical examples of Section 5

For reference, in Tables 1 and 2 we respectively list the values of the average workload and the number of customers under the same conditions as Figures 7 and 8 for the cases where $\rho = 0.8$, 0.85, 0.9, 0.95, and 0.98. In Tables 3 and 4, we also respectively list the values of the

TABLE 1. Workload when arrival process is not NBUE but HNBUE.

	GI/D/1		GI/M/1		GI/H_2/1	
λ/μ	customer average	time average	customer average	time average	customer average	time average
0.98	13.135	13.353	36.959	37.181	101.196	101.419
0.95	4.991	5.193	13.844	14.055	37.932	38.145
0.90	2.285	2.462	6.175	6.368	16.939	17.137
0.85	1.388	1.541	3.648	3.824	10.022	10.206
0.8	0.941	1.073	2.407	2.566	6.624	6.794

TABLE 2. Number of customers in system when arrival process is not NBUE but HNBUE.

	GI/D/1		GI/M/1		GI/H_2/1	
λ/μ	customer average	time average	customer average	time average	customer average	time average
0.98	13.882	14.107	37.692	37.919	101.890	102.117
0.95	5.718	5.939	14.564	14.786	38.639	38.860
0.90	2.970	3.184	6.857	7.071	17.615	17.829
0.85	2.031	2.237	4.290	4.499	10.660	10.867
0.8	1.550	1.740	3.007	3.206	7.220	7.420

https://doi.org/10.1017/jpr.2023.120 Published online by Cambridge University Press

TABLE 3. Workload when arrival process is not NWUE but HNWUE.

	GI/D/1		GI/M/1		GI/H_2/1	
λ/μ	customer average	time average	customer average	time average	customer average	time average
0.98	93.2819	92.0702	126.4702	125.2531	214.5624	213.3298
0.95	34.3390	33.2379	47.2338	46.1053	80.7055	79.5496
0.90	14.8240	13.8947	20.9210	19.9357	36.1857	35.1504
0.85	8.3940	7.6283	12.2393	11.3907	21.5155	20.5924
0.8	5.2469	4.6347	7.9717	7.2519	14.2849	13.4683

TABLE 4. Number of customers in system when arrival process is not NWUE but HNWUE.

	GI/D/1		GI/M/1		GI/H_2/1	
λ/μ	customer average	time average	customer average	time average	customer average	time average
0.98	70.1557	69.2446	94.4432	93.5343	158.9153	158.0006
0.95	26.9426	26.0796	36.3863	35.5170	60.8870	60.0118
0.90	12.5292	11.7484	17.0117	16.2106	28.1937	27.3814
0.85	7.6825	6.9928	10.5377	9.8071	17.3471	16.5954
0.8	5.2304	4.6397	7.2924	6.6339	11.9462	11.2540

average workload and the number of customers under the same conditions as Figures 9 and 10.

Acknowledgements

We wish to thank the referees for their insightful comments and suggestions which had led to a substantial improvement on an earlier version of the manuscript.

Funding information

The present study was supported by the Japan Society for the Promotion of Science (JSPS) through KAKENHI Grant Number JP20K21783.

Competing interests

There are no competing interests to declare that arose during the preparation or publication process of this article.

References

[1] BACCELLI, F. AND BREMAUD, P. (2002). *Elements of Queueing Theory*, 2nd edn. Springer, Berlin.
[2] BARLOW, R. E. AND DOKSUM, K. A. (1972). Isotonic tests for convex orderings. In *Proc. Sixth Berkeley Symp. Mathematical Statistics and Probability, Volume 1: Theory of Statistics*, eds. L. M. LE CARN, J. NEYMAN AND E. L. SCOTT, University of California Press, Berkeley, CA, pp. 293–324.
[3] BRANDT, A., FRANKEN, P. AND LISEK, B. (1990). *Stationary Stochastic Models*. Wiley, Chichester.

https://doi.org/10.1017/jpr.2023.120 Published online by Cambridge University Press

[4] BRÉMAUD, P., KANNURPATTI, R. AND MAZUMDAR, R. (1992). Event and time averages: A review. *Adv. Appl. Prob* **24**, 377–411.
[5] DURRETT, R. (2019). *Probability: Theory and Examples*, 5th edn. Cambridge University Press.
[6] FRANKEN, P., KÖNIG, D., ARNDT, U. AND SCHMIDT, V. (1982). *Queues and Point Processes*. Wiley, Chichester.
[7] FRIEDMAN, M. (1982). Piecewise exponential models for survival data with covariates. *Ann. Statist.* **10**, 101–113.
[8] GHOSH, S. AND MITRA, M. (2020). A new test for exponentiality against HNBUE alternatives. *Commun. Statist. Theory Meth.* **49**, 27–43.
[9] KLEFSJÖ, B. (1980). *Some Properties of the HNBUE and HNWUE Classes of Life Distributions*. Research report 1980-8, Department of Mathematical Statistics, University of Umeå.
[10] KLEFSJÖ, B. (1982). On aging properties and total time on test transforms. *Scand. J. Statist.* **9**, 37–41.
[11] KLEFSJÖ, B. (1991). TTT-plotting—a tool for both theoretical and practical problems. *J. Statist. Planning Infer.* **29**, 99–110.
[12] KÖNIG, D. AND SCHIMIDT, V. (1980). Imbedded and non-imbedded stationary characteristics of queueing systems with varying service rate and point processes. *J. Appl. Prob.* **17**, 753–767.
[13] KÖNIG, D. AND SCHIMIDT, V. (1980). Stochastic inequalities between customer-stationary and time-stationary characteristics of queueing systems with point processes. *J. Appl. Prob.* **17**, 768–777.
[14] KÖNIG, D. AND SCHIMIDT, V. (1989). EPSTA: The coincidence of time-stationary and customer-stationary distributions. *QUESTA* **5**, 247–263.
[15] KÖNIG, D., SCHMIDT, V. AND STOYAN, D. (1976). On some relations between stationary distributions of queue lengths and imbedded queue lengths in G/G/*s* queueing systems. *Statistics* **7**, 577–586.
[16] MELAMED, B. AND WHITT, W. (1990). On arrivals that see time averages. *Operat. Res.* **38**, 156–172.
[17] MIYAZAWA, M. (1976). Stochastic order relations among GI/G/1 queues with a common traffic intensity. *J. Operat. Res. Soc. Japan* **19**, 193–208.
[18] MIYAZAWA, M. AND WOLFF, R. (1990). Further results on ASTA for general stationary processes and related problems. *J. Appl. Prob.* **27**, 792–804.
[19] MÜLLER, A. AND STOYAN, D. (2002). *Comparison Methods for Stochastic Models and Risks*. Wiley, Chichester.
[20] NIU, S.-C. (1984). Inequalities between arrival averages and time averages in stochastic processes arising from queueing theory. *Operat. Res.* **32**, 785–795.
[21] PEKÖZ, E. AND ROSS, S. (2008). Relating time and customer averages for queues using 'forward' coupling from the past. *J. Appl. Prob.* **45**, 568–574.
[22] PEKÖZ, E., ROSS, S. AND SESHADRI, S. (2008). How nearly do arriving customers see time-average behavior? *J. Appl. Prob.* 45, 963–971.
[23] ROLSKI, T. (1975). Mean residual life. *Bull. Inst. Internat. Statist.* **46**, 266–270.
[24] ROLSKI, T. (1981). *Stationary Random Proceses Associated with Point Proceses*. Springer, Berlin.
[25] SHAKED, M. AND SHANTHIKUMAR, J. G. (1993). *Stochastic Orders and Their Applications*. Academic Press, New York.
[26] SHANTHIKUMAR, J. G. AND ZAZANIS, M. (1999). Inequalities between event and time averages. *Prob. Eng. Inf. Sci.* **13**, 293–308.
[27] SHORTLE, J. F., THOMPSON, J. M., GROSS, D. AND HARRIS, C. M. (2018). *Fundamentals of Queueing Theory*. Wiley, Chichester.
[28] SZEKLI, R. (1995). *Stochastic Ordering and Dependence in Applied Probability*. Springer, Berlin.
[29] WOLFF, R. (1982). Poisson arrivals see time averages. *Operat. Res.* **30**, 223–231.

https://doi.org/10.1017/jpr.2023.120 Published online by Cambridge University Press