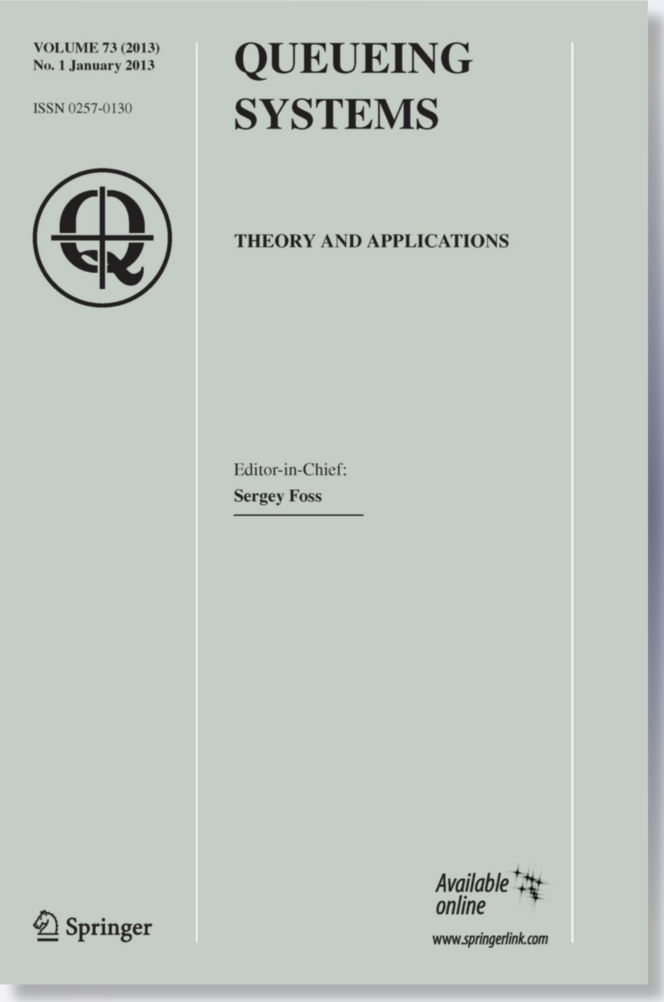
Analysis of queueing systems with customer interjections

Qi-Ming He & Alireza A. Chavoushi

Queueing Systems
Theory and Applications

ISSN 0257-0130
Volume 73
Number 1

Queueing Syst (2013) 73:79-104
DOI 10.1007/s11134-012-9302-1



VOLUME 73 (2013)
No. 1 January 2013

ISSN 0257-0130

QUEUEING SYSTEMS

THEORY AND APPLICATIONS

Editor-in-Chief:
Sergey Foss

Springer

Available online
www.springerlink.com

Springer

Queueing Syst (2013) 73:79–104
DOI 10.1007/s11134-012-9302-1

Analysis of queueing systems with customer interjections

Qi-Ming He · Alireza A. Chavoushi

Received: 25 July 2010 / Published online: 3 May 2012

Abstract In this paper we study queueing systems with customer interjections. Customers are distinguished into normal customers and interjecting customers. All customers join a single queue waiting for service. A normal customer joins the queue at the end and an interjecting customer tries to cut in the queue. The waiting times of normal customers and interjecting customers are studied. Two parameters are introduced to describe the interjection behavior: the percentage of customers interjecting and the tolerance level of interjection by individual customers. The relationship between the two parameters and the mean and variance of waiting times is characterized analytically and numerically.

Keywords Queueing system · Waiting time · Stochastically larger order · Priority queue · Matrix-analytic methods

Mathematics Subject Classification 60K25 · 60J20

1 Introduction

Wherever queues exist, customer interjections may occur. For instance, the first-come-first-served (FCFS) service discipline is usually assumed in public places such as airports, supermarkets, and restaurants. However, customer interjections can still be seen there. Some customers simply try to cut in queue, while others find excuses or find friends in the queue to cut in. For traffic at an intersection with several lanes,

Q.-M. He (✉)
Department of Management Sciences, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada
e-mail: q7he@uwaterloo.ca

A.A. Chavoushi
Department of Industrial Engineering, Dalhousie University, Halifax, Nova Scotia B3J 2X4, Canada
e-mail: AL351118@DAL.CA

some drivers may use a left (right) turning lane to cut into a straight lane or vice versa. Such interjections can cause the traffic to slow down in order to avoid possible accidents. In telecommunications networks, to test the efficiency of data transmission, artificial packages are inserted into the normal traffic in a random manner. Customer interjections may save time for some customers, while they may increase the waiting times of others. In some cases (e.g., traffic at an intersection and data packages in telecommunication networks), interjections can reduce the efficiency of the system. Thus, customer interjection is a factor that should not be ignored in the design and operation control of service, transportation, and telecommunication systems.

Larson [11] discussed social justice and the psychology of queueing. Larson used "slips" and "skips" to describe customer interjections: "he who experiences a slip is victimized; he who skips gets a certain sense of satisfaction from his good fortune." A number of examples where slips and skips occur are presented and analyzed in [11]. For some cases, non-technical but effective solutions are discussed. In [5] some probability laws for slips and skips in a number of different classical queueing systems were derived. The models considered in [5] are different from the model studied in this paper. In particular, they considered the situation where first-in customers leave the queueing system later than others. The difference between the order of arrivals and the order of departures is caused mainly by service, not interjection. In [24] a similar issue was considered for queueing networks, where slips and skips are called customers overtaken.

The queueing model investigated in this paper is closely related to queueing models in which queue positions can be purchased (e.g., [1, 7, 9, 13], and references therein). In such queueing models, the queue positions of customers are determined by how much they pay to the system. Thus, the relative positions of all customers in queue are determined by their payments (or bribe). For a number of cases, when a cost function is introduced, the optimal policy for payment has been obtained. Queueing models in which customers pay for queue position have found applications in auction, supply chain management, social systems, and computer systems. In the model studied in this paper, the queue position of a newly arrived interjecting customer is determined by the kindness of customers already in queue, not by the payment of the customer.

The queueing model investigated in this paper is also closely related to queueing models with customer service priorities ([4, 7, 10, 21, 22], and references therein). In priority queues, customers with higher priority skip over customers with lower priority, which leads to the difference between the waiting times of different types of customers. The queueing process of customers with higher priority is, in general, not affected by that of lower priority customers. The reason is that the service priorities are maintained during the waiting and service periods of customers. Furthermore, customers already in queue have little influence on the positions of new customers. For a queue with customer interjections, an interjecting customer's queue position depends on all customers in queue at the arrival epoch. Consequently, the queueing processes of all types of customers interact with others. Thus, while customer interjections can be viewed as partial priority, its impact on the queueing processes of customers of different types of customers in queue is different from that of customer priority. In Sect. 2.5 a comparison between the waiting times of a queue with customer interjections, a standard FCFS queue, and a priority queue is carried out. The results indicate that the behaviors of the queues are significantly different.

Queues with customer jockeying (or transferring) ([23, 25], and references therein) were studied extensively. For such queueing models, customer "slips" and "skips" can occur since there are multiple queues and jockeying between queues. For some cases, customers can choose which queue to join in order to minimize their waiting times. In those systems the decision for jockeying resides within individual customers and there is no interaction between customers. In our model the customer interjection process is between interjecting customers and non-interjecting customers. The behavior of either type of customers has great impact on the queueing process of the other.

In this paper we study a queueing model for which all customers join a single queue. Some of the customers will try to cut in the queue expecting to be served earlier. We call such an action *customer interjection* and customers with such an action *interjecting customers*. Customers with no intention of interjecting are called *normal customers*. We investigate the effects of customer interjections on the waiting times of interjecting, normal, and *arbitrary* customers. We focus on the relationship between the waiting times and two system parameters related to societal behavior: the percentage of interjecting customers and the tolerance level of interjection by individual customers. Intuitively, it is expected that:

(i) The impact of the societal tolerance level of interjection and the percentage of interjecting customers on the mean and variance of the waiting time of a normal customer is negative; i.e., when the two parameters increase, the mean and variance increase.

(ii) The impact of the societal tolerance level of interjection on the mean of the waiting time of an interjecting customer is, in general, positive; i.e., the higher the tolerance level, the smaller the mean waiting time.

(iii) The societal tolerance level of interjection and the percentage of interjecting customers have no impact on the mean waiting time of an arbitrary customer. On the other hand, their impact on the variance of the waiting time of an arbitrary customer is significant.

Using both theoretical and numerical methods, Sect. 2 shows that the above intuitive results are indeed true for an $M/M/1$ queue. In Sect. 3 the results are confirmed numerically for a more general queueing model. Numerical examples in Sect. 3 also demonstrate that the mean and variance of the waiting time of a normal/interjecting customer increase if the arrival processes of normal and interjecting customers are bursty or if the service time is more variable.

The theory of Markov processes [4, 18, 19] provides us with the basic tool to study the queueing system of interest. In addition, stochastic comparison [3, 20] and matrix-analytic methods [16, 17] are utilized in the study as well. Particularly, the monotonicity of the waiting times in some system parameters is shown by using stochastic comparison. Algorithms for computing the means and variances of waiting times are developed based on matrix-analytic methods, which take the advantage of the tri-diagonal structure in the transition matrix.

The remainder of the paper is organized as follows. In Sect. 2 a basic queueing model with customer interjections, i.e., an $M[2]/M/1$ queue with customer interjections, is introduced and analyzed. Section 2.1 studies the absorption time of a Markov process that is used in finding the waiting times of customers. In Sects. 2.2 and 2.3 the waiting time of a normal customer and the waiting time of an interjecting customer are investigated, respectively. In Sect. 2.4 the waiting time of an arbitrary customer is investigated. Section 2.5 compares the $M/M/1$/FCFS queue, a priority $M/M/1$ queue, and the $M/M/1$ queue with customer interjections. In Sect. 3 the basic queueing model is extended to an $MMAP[2]/PH/1$ queue by incorporating correlations in the customer arrival process and by using more general service time distributions. An algorithm is developed for computing the means and variances of the waiting times of normal, interjecting, and arbitrary customers. Section 4 concludes the paper.

2 An $M[2]/M/1$ queue with customer interjections

We consider a single server queueing system with a single queue and two types of customers. The two types of customers are called *normal customers* and *interjecting customers*, respectively. When a normal customer arrives, it joins the queue at the end. When an interjecting customer arrives, it may join the queue at any queue position depending on the queue length at the arrival epoch and a given probability distribution.

We assume that customers arrive at the system according to a Poisson process with arrival rate λ. The service times of all customers are independent and identically distributed random variables with an exponential distribution with service rate μ. The service process and the arrival process are independent. We assume that an arriving customer is interjecting with probability η_I $(0 \le \eta_I \le 1)$ and is normal with probability $1 - \eta_I$. Note that the subscript "I" is for the word "interjection." According to a classical result for Poisson processes [18], such an arrival process can also be viewed as the superposition of two independent Poisson processes with arrival rates $\eta_I \lambda$ and $(1 - \eta_I)\lambda$, respectively. When an interjecting customer arrives, it tries to join the queue as close to the head of the queue as possible. We assume that the interjecting customers do not interrupt the service in progress. Thus, the arriving customer contacts the first customer waiting in queue for possible interjection. That customer (regardless of its own type) may let the new customer cut in with probability η_C $(0 \le \eta_C \le 1)$ (i.e., taking the first position in queue). Note that the subscript "C" is for "cutting in". If the first customer refuses the interjection request, the new customer contacts the second customer in queue. The process repeats until either the customer interjects successfully or it joins the queue at the end if all waiting customers refuse its interjection request. We assume that the time for finding a position in queue is negligible for an interjecting customer. Given that there are n customers in queue at the arrival epoch, there are positions $\{1, 2, \dots, n + 1\}$ available to the arriving interjecting customer. Thus, the position taken by the interjecting customer has a truncated geometric distribution $\{\eta_C, (1-\eta_C)\eta_C, (1-\eta_C)^2\eta_C, \dots, (1-\eta_C)^{n-1}\eta_C, (1-\eta_C)^n\}$ on positions $\{1, 2, \dots, n + 1\}$.

It is easy to see that, if $\eta_I = 0$ or $\eta_C = 0$, the queueing model is reduced to the classical $M/M/1$ queue with a first-come-first-served (FCFS) service discipline. If $\eta_I = \eta_C = 1$, the queueing model is reduced to the classical $M/M/1$ queue with a last-come-first-served (LCFS) service discipline.

The parameter η_I represents the percentage of customers with interjection intention, which reflects the societal behavior on interjection. The parameter η_C represents the level of tolerance of individuals on interjection. We are mainly interested in the impact of the pair (η_I, η_C) on the waiting times of normal customers and interjecting customers. From a social justice point of view, it is always expected that the values of η_I and η_C should be low, i.e., close to zero. Thus, we shall pay special attention to cases in which η_I or η_C is close to zero.

It is readily seen that the queue length in the system of interest is the same as that in the classical $M/M/1$ queue with an FCFS service discipline, but the waiting time can be different. Define $q(t)$ as the total number of customers in the system at time t, which is usually called the queue length at time t. Then the steady state distribution of $q(t)$ exists if and only if $\rho = \lambda/\mu < 1$, and is given by [4]:

$$\lim_{t\to\infty} P\{q(t)=n\} = (1-\rho)\rho^n, \quad n \geq 0. \tag{2.1}$$

In the rest of this section we study the waiting times of normal customers, interjecting customers, and arbitrary customers. To that end, we first analyze the waiting time of a customer in the nth position in the queue.

2.1 Waiting time of a customer in position n

Let $W_n(\eta_I, \eta_C)$ be the waiting time of a customer currently in position n in the queue; i.e., the length of the time starting from the epoch that a customer is currently in position n in the queue and ending at the epoch that the customer enters the server, $n \geq 1$. The waiting times of normal customers and interjecting customers can be expressed in terms of $\{W_n(\eta_I, \eta_C), n \geq 1\}$ (see Eqs. (2.11) and (2.16)).

To find the distribution of $W_n(\eta_I, \eta_C)$, we introduce an absorbing Markov process to describe the change of position of a customer in the queue. Suppose that a customer is in position n in the queue. The customer moves to position $n-1$ if the current service completes before the next arrival. If the next arrival occurs first, the customer remains in position n if the arrival does not interject. If the new customer interjects into one of the first n positions, the customer in position n moves to position $n+1$. The new customer interjects into one of the first n positions with probability $\eta_I(1-(1-\eta_C)^n)$. Therefore, the change of the position of a customer in queue can be described by a Markov process with a state space $\{0, 1, 2, \ldots\}$:

$$Q_a = \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \\ \vdots \end{array} \begin{pmatrix} 0 & 0 & & & & \\ \mu & Q_{1,1} & Q_{1,2} & & & \\ & Q_{2,1} & Q_{2,2} & Q_{2,3} & & \\ & & Q_{3,2} & Q_{3,3} & Q_{3,4} & \\ & & & \ddots & \ddots & \ddots \end{pmatrix}$$

$$\equiv \begin{pmatrix} 0 & 0 \\ \mathbf{b} & Q \end{pmatrix}, \tag{2.2}$$

where $\mathbf{b}$ is a column vector with all elements being zero except the first one which is μ, and Q is a tri-diagonal matrix with $Q_{i,i} = -\mu - \lambda\eta_I(1-(1-\eta_c)^i)$, $Q_{i+1,i} = \mu$, and $Q_{i,i+1} = \lambda\eta_I(1-(1-\eta_c)^i)$, for $i \geq 1$. By definition, $W_n(\eta_I, \eta_C)$ is the absorption time of state zero of the Markov process Q_a given that the initial state was

$n, n \geq 1$. Thus, the distribution of $W_n(\eta_I, \eta_C)$ has a phase type representation [16] with an infinite state space and

$$P\bigl\{W_n(\eta_I, \eta_C) < t\bigr\} = 1 - \boldsymbol{\alpha}_n \exp\{Qt\}\mathbf{e}, \tag{2.3}$$

where $\boldsymbol{\alpha}_n$ is a vector with the nth component being one and all others zero, and $\mathbf{e}$ is the column vector with all components being one. By a well-known ergodicity condition for birth-and-death processes [4, 18], the following lemma gives a necessary and sufficient condition for the finiteness of the waiting times. In fact, the condition in Lemma 2.1 ensures that the absorption time of the state zero is finite with probability one. Details of the proof are omitted.

Lemma 2.1 *The waiting times* $\{W_n(\eta_I, \eta_C), n \geq 1\}$ *are finite with probability one if and only if* $\eta_I \rho < 1$. *The moments of waiting times* $\{W_n(\eta_I, \eta_C), n \geq 1\}$ *are finite if and only if* $\eta_I \rho < 1$.

Intuitively, $\eta_I \rho < 1$ implies that the process Q_a, on average, always drifts toward state zero. We shall assume $\eta_I \rho < 1$ in the rest of this section. In Sects. 2.2 and 2.4 the condition $\eta_I \rho < 1$ is replaced by a stronger one $\rho < 1$ to ensure the finiteness of the waiting times of normal customers and arbitrary customers.

For two random variables X and Y, X is stochastically larger than Y if $P\{X < t\} \leq P\{Y < t\}$ for all t. We refer to [14, 20] for basic properties of the stochastically larger order. It is intuitive that $W_n(\eta_I, \eta_C)$ is increasing in η_I and η_C. Based on Eqs. (2.2) and (2.3) and Lemma 2.1, the stochastic monotonicity of $W_n(\eta_I, \eta_C)$ in η_I and η_C can be shown.

Lemma 2.2 *Assume that* $\eta_I \rho < 1$. *If* $(\eta_I, \eta_C) \leq (\eta'_I, \eta'_C)$, *i.e.,* $\eta_I \leq \eta'_I$ *and* $\eta_C \leq \eta'_C$, *then the waiting time* $W_n(\eta'_I, \eta'_C)$ *is stochastically larger than the waiting time* $W_n(\eta_I, \eta_C)$.

Proof By Eq. (2.3), we have

$$\begin{aligned} P\bigl\{W_n(\eta_I, \eta_C) > t\bigr\} &= e^{-(\lambda+\mu)t} \boldsymbol{\alpha}_n \exp\biggl\{\biggl(I + \frac{Q}{\lambda+\mu}\biggr)(\lambda+\mu)t\biggr\}\mathbf{e} \\ &= e^{-(\lambda+\mu)t} \sum_{m=0}^{\infty} \frac{((\lambda+\mu)t)^m}{m!} \boldsymbol{\alpha}_n P^m \mathbf{e}, \end{aligned} \tag{2.4}$$

where I is the identity matrix and $P = I + Q/(\lambda+\mu) = (p_{i,j})_{i,j\geq 1}$, which is a substochastic matrix [19]. To prove the lemma, it is sufficient to show that $P^m\mathbf{e}$ is non-decreasing in η_I and η_C elementwise for $m \geq 0$. It is easy to see that the matrix P is monotone, i.e., the $(k+1)$st row of P dominates the kth row of P for all $k \geq 0$, $\sum_{j=n}^{\infty} p_{k+1,j} \geq \sum_{j=n}^{\infty} p_{k,j}$, for $n \geq 1$ [14]. It is also easy to obtain $P\mathbf{e} = (\lambda/(\lambda+\mu), 1, 1, \ldots)^{\mathrm{T}}$, where "T" is for transpose. Thus, the components of $P\mathbf{e}$ are non-decreasing. Let $P^m\mathbf{e} = (d_{m,1}, d_{m,2}, \ldots)^{\mathrm{T}}$. Suppose that $d_{m,1} \leq d_{m,2} \leq \cdots$, i.e., the elements of $P^m\mathbf{e}$ are in ascending order. For $P^{m+1}\mathbf{e}$, we have, for $j \geq 2$,

$$\begin{aligned} d_{m+1,j} &= p_{j,j-1}d_{m,j-1} + p_{j,j}d_{m,j} + p_{j,j+1}d_{m,j+1} \\ &\ge p_{j,j-1}d_{m,j-2} + p_{j,j}d_{m,j-1} + p_{j,j+1}d_{m,j} \\ &= p_{j-1,j-2}d_{m,j-2} + p_{j-1,j-1}d_{m,j-1} + p_{j-1,j}d_{m,j} \\ &\quad + (p_{j,j} - p_{j-1,j-1})d_{m,j-1} + (p_{j,j+1} - p_{j-1,j})d_{m,j} \\ &= d_{m+1,j-1} + \frac{\lambda\eta_I(1-\eta_C)^{j-1}\eta_C}{\lambda+\mu}(d_{m,j} - d_{m,j-1}) \\ &\ge d_{m+1,j-1}. \end{aligned} \tag{2.5}$$

By induction, $P^m\mathbf{e}$ is monotone for all m.

Suppose that $(\eta_I, \eta_C) \le (\eta'_I, \eta'_C)$. Then P is dominated by P' (i.e., every row of P is dominated by the corresponding row of P'), which can be readily obtained by comparing the vectors $(\mu/(\lambda+\mu), 1-\mu/(\lambda+\mu)-\lambda\eta_I(1-(1-\eta_C)^n)/(\lambda+\mu), \lambda\eta_I(1-(1-\eta_C)^n)/(\lambda+\mu))$ and $(\mu/(\lambda+\mu), 1-\mu/(\lambda+\mu)-\lambda\eta'_I(1-(1-\eta'_C)^n)/(\lambda+\mu), \lambda\eta'_I(1-(1-\eta'_C)^n)/(\lambda+\mu))$. Since $P\mathbf{e} = P'\mathbf{e}$, $P\mathbf{e}$ is dominated by $P'\mathbf{e}$. Suppose that $P^m\mathbf{e}$ is dominated by $(P')^m\mathbf{e}$. Since P is monotone and $P^m\mathbf{e}$ is dominated by $(P')^m\mathbf{e}$, we have $P(P^m\mathbf{e}) \le P(P')^m\mathbf{e}$. Since P is dominated by P' and $(P')^m\mathbf{e}$ is monotone, we have $P(P')^m\mathbf{e} \le P'(P')^m\mathbf{e}$. Then $P^{m+1}\mathbf{e} = P(P^m\mathbf{e}) \le P(P')^m\mathbf{e} \le P'(P')^m\mathbf{e} = (P')^{m+1}\mathbf{e}$. Therefore, the elements of $P^m\mathbf{e}$ are monotone in (η_I, η_C) for all m. By Eq. (2.4), the waiting time is monotone in (η_I, η_C) with respect to the stochastically larger order. This completes the proof of Lemma 2.2. □

Define $w_n^*(s) = E[\exp\{-sW_n(\eta_I, \eta_C)\}]$, $s \ge 0$. For convenience, we define $W_0(\eta_I, \eta_C) = 0$. Then we have $w_0^*(s) = 1$. Conditioning on the next transition of the Markov process Q_a, it is easy to obtain: for $n \ge 1$,

$$\begin{aligned} w_n^*(s) = \frac{\mu + \lambda\eta_I(1-(1-\eta_C)^n)}{s+\mu+\lambda\eta_I(1-(1-\eta_C)^n)} \Bigg[& \frac{\mu w_{n-1}^*(s)}{\mu+\lambda\eta_I(1-(1-\eta_C)^n)} \\ & + \frac{\lambda\eta_I(1-(1-\eta_C)^n)w_{n+1}^*(s)}{\mu+\lambda\eta_I(1-(1-\eta_C)^n)} \Bigg]. \end{aligned} \tag{2.6}$$

By Eq. (2.6), the following expression for $w_n^*(s)$ can be obtained, which is more convenient for analyzing the mean and variance of waiting times.

Lemma 2.3 *Assume that $\eta_I\rho < 1$. The functions $\{w_n^*(s), n \ge 1\}$ satisfy the following equation*:

$$w_n^*(s) = 1 - \frac{s}{\mu}\left[\sum_{m=1}^{n}\sum_{k=m}^{\infty}\left[(\rho\eta_I)^{k-m}\prod_{j=0}^{k-m-1}\left(1-(1-\eta_C)^{m+j}\right)\right]w_k^*(s)\right]. \tag{2.7}$$

Note that $\prod_{j=0}^{-1}(\ldots) = 1$ and $w_0^*(s) = 1$ by convention.

With the expression in Eq. (2.7), we are able to derive formulas for the mean and variance of the waiting time $W_n(\eta_I, \eta_C)$. The results are summarized in the following lemma.

Lemma 2.4 *Assume that $\eta_I\rho < 1$. The first two moments of $W_n(\eta_I, \eta_C)$ are non-decreasing in η_I and η_C for $n \geq 1$. Furthermore, the first two moments of $W_n(\eta_I, \eta_C)$ are given explicitly as follows: $E[W_0(\eta_I, \eta_C)] = E[(W_0(\eta_I, \eta_C))^2] = 0$, and for $n \geq 1$,*

$$\begin{aligned} E\big[W_n(\eta_I, \eta_C)\big] &= \frac{1}{\mu}\sum_{m=1}^{n}\sum_{k=m}^{\infty}(\rho\eta_I)^{k-m}\prod_{j=0}^{k-m-1}\big(1-(1-\eta_C)^{m+j}\big); \\ E\big[W_n^2(\eta_I, \eta_C)\big] & \\ &= \frac{2}{\mu}\sum_{m=1}^{n}\sum_{k=m}^{\infty}(\rho\eta_I)^{k-m}\left(\prod_{j=0}^{k-m-1}\big(1-(1-\eta_C)^{m+j}\big)\right)E\big[W_k(\eta_I, \eta_C)\big]. \end{aligned} \tag{2.8}$$

In addition, the variance of $W_n(\eta_I, \eta_C)$, denoted by $Var[W_n(\eta_I, \eta_C)]$, is increasing in η_I and η_C. For fixed η_C, the functions $E[W_n(\eta_I, \eta_C)]$, $E[(W_n(\eta_I, \eta_C))^2]$, and $Var[W_n(\eta_I, \eta_C)]$ are convex in η_I.

Proof By Lemma 2.2, $W_n(\eta_I, \eta_C)$ becomes stochastically larger if η_I and η_C are increasing. Consequently, the first two moments of $W_n(\eta_I, \eta_C)$ are non-decreasing in η_I and η_C. Expressions in Eq. (2.8) are obtained from Eq. (2.7) by routine calculations. The last part of the lemma is obtained from Lemmas A.1 and A.2 in the Appendix. This completes the proof of Lemma 2.4. □

Using expressions in Eq. (2.8), the mean and variance of $W_n(\eta_I, \eta_C)$ can be calculated. However, a computational method based on Eq. (2.8) can be numerically instable and time consuming. Existing approximation methods can be used to compute the mean and variance more efficiently (e.g., [6, 15]). Based on matrix-analytic methods, an efficient algorithm for computing the first two moments of $W_n(\eta_I, \eta_C)$ is developed in the Appendix.

2.2 Waiting time of a normal customer

We now consider the waiting time $W_{[N]}(\eta_I, \eta_C)$ of a (arbitrary) normal customer. Recall that a normal customer always joins the queue at the end. Assuming $\rho < 1$ and conditioning on the number of customers in the system at the arrival epoch, by Eq. (2.1), we obtain

$$P\big\{W_{[N]}(\eta_I, \eta_C) < t\big\} = 1-\rho+\sum_{n=1}^{\infty}(1-\rho)\rho^n P\big\{W_n(\eta_I, \eta_C) < t\big\}, \quad t > 0. \tag{2.9}$$

By Eq. (2.9), we obtain

$$\begin{aligned} E\big[W_{[N]}(\eta_I, \eta_C)\big] &= \sum_{n=1}^{\infty}(1-\rho)\rho^n E\big[W_n(\eta_I, \eta_C)\big] \\ &= \frac{1}{\mu}\sum_{k=0}^{\infty}(\rho\eta_I)^k\left[\sum_{m=1}^{\infty}\rho^m\prod_{j=0}^{k-1}\big(1-(1-\eta_C)^{m+j}\big)\right]; \end{aligned} \tag{2.10}$$

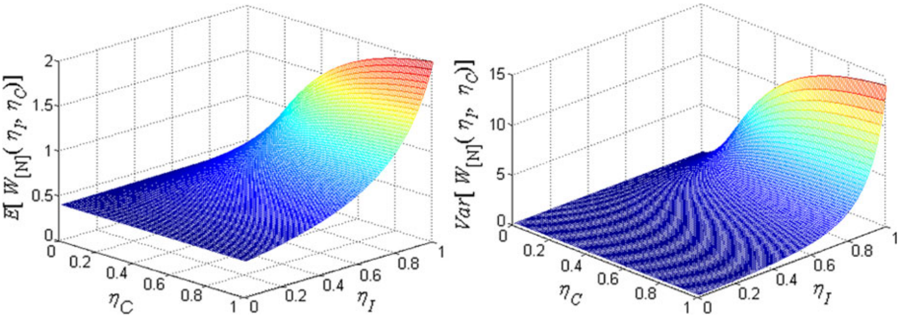


Fig. 2.1 The mean and variance of $W_{[\mathrm{N}]}(\eta_I, \eta_C)$ for $0 \le \eta_I, \eta_C \le 1$

$$Var\big[W_{[\mathrm{N}]}(\eta_I, \eta_C)\big] = \sum_{n=1}^{\infty}(1-\rho)\rho^n E\big[\big(W_n(\eta_I, \eta_C)\big)^2\big] - \big(E\big[W_{[N]}(\eta_I, \eta_C)\big]\big)^2.$$

By Lemmas 2.2 and 2.4, and Eqs. (2.9) and (2.10), $W_{[\mathrm{N}]}(\eta_I, \eta_C)$, $E[W_{[\mathrm{N}]}(\eta_I, \eta_C)]$, and $Var[W_{[\mathrm{N}]}(\eta_I, \eta_C)]$ are characterized as follows.

Proposition 2.1 *If $\rho < 1$, the random variable $W_{[\mathrm{N}]}(\eta_I, \eta_C)$ is non-decreasing in η_I and η_C with respect to the stochastically larger order. In addition, we have*

- (i) *The function $E[W_{[\mathrm{N}]}(\eta_I, \eta_C)]$ is non-decreasing in η_I and η_C. For fixed η_C, the function $E[W_{[\mathrm{N}]}(\eta_I, \eta_C)]$ is convex in η_I.*
- (ii) *The function $Var[W_{[\mathrm{N}]}(\eta_I, \eta_C)]$ is non-decreasing in η_I and η_C. For fixed η_C, the function $Var[W_{[\mathrm{N}]}(\eta_I, \eta_C)]$ is convex in η_I.*

We remark that (i) $E[W_{[N]}(\eta_I, \eta_C)]\mu$ gives the mean number of customers served during the waiting period of a normal customer (this applies to other types of customers as well) and (ii) numerical results indicate that $E[W_{[\mathrm{N}]}(\eta_I, \eta_C)]$ is neither a convex nor a concave function in η_C.

In the rest of this section we analyze the mean and variance of $W_{[\mathrm{N}]}(\eta_I, \eta_C)$ numerically. The following example demonstrates the structure of both functions $E[W_{[\mathrm{N}]}(\eta_I, \eta_C)]$ and $Var[W_{[\mathrm{N}]}(\eta_I, \eta_C)]$.

Example 2.1 We consider a queueing model with $\lambda = 8$ and $\mu = 10$. Then $\rho = 0.8 < 1$. The mean and variance of $W_{[\mathrm{N}]}(\eta_I, \eta_C)$ are plotted in Fig. 2.1 for $0 \le \eta_I$, $\eta_C \le 1$.

Figure 2.1 shows that the mean and variance of the waiting time of a normal customer are non-decreasing in both η_I and η_C, which is consistent with Proposition 2.1. Intuitively, for fixed η_C, if η_I increases, more customers will interject. Consequently, the waiting time of a normal customer will increase. For fixed η_I, if η_C increases, the chance for more customers cutting in before a given customer is greater. Consequently, the waiting time of a normal customer will increase. Figure 2.1 also shows that the variance of the waiting time is non-decreasing in both η_I and η_C. It is well known that the variance of the waiting time is minimized if customers are all served on the FCFS basis. If η_I or η_C increases, the service order is drifting away from FCFS. Consequently, the variance of the waiting time of a normal customer increases. The variance is maximized at $\eta_I = \eta_C = 1$, which corresponds to the $M/M/1$ queue with an LCFS service discipline. Numerical experiments are conducted for examples with different arrival rate λ and service rate μ. For all examples tested, $E[W_{[\mathrm{N}]}(\eta_I, \eta_C)]$ and $Var[W_{[\mathrm{N}]}(\eta_I, \eta_C)]$ are similar to those plotted in Fig. 2.1.

The plots in Fig. 2.1 show that the mean and variance of the waiting time of a normal customer can increase drastically due to interjection. They also show that the increase is most significant at boundary points. Next, we conduct a sensitivity analysis on the mean waiting time $E[W_{[\mathrm{N}]}(\eta_I, \eta_C)]$. By routine calculations, Eq. (2.10) leads to the following results.

Proposition 2.2 *Assume that* $\rho < 1$. *For the boundary points* $(\eta_I, 0)$, $(\eta_I, 1)$, $(0, \eta_C)$, *and* $(1, \eta_C)$, *we have*

$$\begin{aligned} \left.\frac{\partial E[W_{[\mathrm{N}]}(\eta_I, \eta_C)]}{\partial \eta_I}\right|_{\eta_I=0} &= \frac{\rho^2 \eta_c}{\mu(1-\rho)(1-\rho(1-\eta_C))}; \\ \left.\frac{\partial E[W_{[N]}(\eta_I, \eta_C)]}{\partial \eta_I}\right|_{\eta_I=1} &= \frac{(1-\rho)}{\mu} \sum_{n=1}^{\infty}\sum_{k=1}^{\infty} \rho^{n+k} k \sum_{m=1}^{n} \prod_{j=0}^{k-1} \bigl(1-(1-\eta_C)^{m+j}\bigr), \end{aligned} \tag{2.11}$$

and

$$\begin{aligned} \left.\frac{\partial E[W_{[\mathrm{N}]}(\eta_I, \eta_C)]}{\partial \eta_C}\right|_{\eta_C=0} &= \frac{\rho^2 \eta_I}{\mu(1-\rho)^2}; \\ \left.\frac{\partial E[W_{[\mathrm{N}]}(\eta_I, \eta_C)]}{\partial \eta_C}\right|_{\eta_C=1} &= \frac{\rho^2 \eta_I}{\mu(1-\rho\eta_I)}. \end{aligned} \tag{2.12}$$

By Eqs. (2.11) and (2.12), we observe that the derivative of $E[W_{[\mathrm{N}]}(\eta_I, \eta_C)]$ is larger, sometime significantly, at $\eta_C = 0$ than that at $\eta_C = 1$. The implication is that a small increase in the tolerance level may lead to a significant increase in the waiting time of a normal customer. Thus, it is important to make sure that no one tolerates interjection, i.e., to keep $\eta_C = 0$.

In summary, both the mean and variance of $W_{[\mathrm{N}]}(\eta_I, \eta_C)$ are increasing in η_I and η_C. Therefore, reducing the chance for customer interjections is always beneficial to normal customers.

2.3 Waiting time of an interjecting customer

Denote by $W_{[I]}(\eta_I, \eta_C)$ the waiting time of an (arbitrary) interjecting customer. If an interjecting customer finds n customers waiting in queue upon arrival, the customer takes position j with probability $(1-\eta_C)^{j-1}\eta_C$, $1 \le j \le n$, and position $n+1$ with probability $(1-\eta_C)^n$. Assuming $\rho < 1$ and conditioning on the number of customers in the system at the arrival epoch, we obtain

$$\begin{aligned}
&E\big[\exp\big\{-sW_{[I]}(\eta_I,\eta_C)\big\}\big]\\
&\quad = 1-\rho+\sum_{n=1}^{\infty}(1-\rho)\rho^n\left[\sum_{j=1}^{n-1}(1-\eta_C)^{j-1}\eta_C w_j^*(s)+(1-\eta_C)^{n-1}w_n^*(s)\right]\\
&\quad = 1-\rho+\big(1-\rho(1-\eta_C)\big)\sum_{n=1}^{\infty}\rho^n(1-\eta_C)^{n-1}w_n^*(s).
\end{aligned}\tag{2.13}$$

If $\rho \geq 1$, the (total) queue length is infinite, an interjecting customer takes position j in the queue with probability $(1-\eta_C)^{j-1}\eta_C$, $1 \leq j < \infty$. Assuming $\eta_I\rho < 1$, $\eta_c < 1$, and $\rho \geq 1$, we have

$$E\big[\exp\big\{-sW_{[I]}(\eta_I,\eta_C)\big\}\big]=\sum_{n=1}^{\infty}(1-\eta_C)^{n-1}\eta_C w_n^*(s).\tag{2.14}$$

By Lemma 2.4 and Eqs. (2.13) and (2.14), the first two moments of $W_{(I)}(\eta_I,\eta_C)$ can be obtained as follows, if $\eta_I\rho < 1$:

$$\begin{aligned}
&E\big[W_{[I]}(\eta_I,\eta_C)\big]\\
&\quad=\begin{cases}(1-\rho(1-\eta_C))\sum_{n=1}^{\infty}\rho^n(1-\eta_C)^{n-1}E[W_n(\eta_I,\eta_C)], & \rho<1;\\ \sum_{n=1}^{\infty}\eta_C(1-\eta_C)^{n-1}E[W_n(\eta_I,\eta_C)], & \eta_C>0 \text{ and } \rho\geq 1.\end{cases}\\
&\quad=\begin{cases}\frac{1}{\mu(1-\eta_C)}\sum_{k=0}^{\infty}(\rho\eta_I)^k[\sum_{m=1}^{\infty}(\rho(1-\eta_C))^m\prod_{j=0}^{k-1}(1-(1-\eta_C)^{m+j})],\\ \quad\rho<1;\\ \frac{1}{\eta_C\mu}+\frac{\eta_C}{\mu}\sum_{n=1}^{\infty}\sum_{k=1}^{\infty}(\rho\eta_I)^k(1-\eta_C)^{n-1}[\sum_{m=1}^{n}\prod_{j=0}^{k-1}(1-(1-\eta_C)^{m+j})],\\ \quad\eta_C>0 \text{ and } \rho\geq 1.\end{cases}
\end{aligned}\tag{2.15}$$

$$\begin{aligned}
&Var\big[W_{[I]}(\eta_I,\eta_C)\big]\\
&\quad=\begin{cases}(1-\rho(1-\eta_C))\sum_{n=1}^{\infty}\rho^n(1-\eta_C)^{n-1}E[(W_n(\eta_I,\eta_C))^2]\\ \quad-(E[W_{[I]}(\eta_I,\eta_C)])^2, \quad \rho<1;\\ \sum_{n=1}^{\infty}\eta_C(1-\eta_C)^{n-1}E[(W_n(\eta_I,\eta_C))^2]-(E[W_{[I]}(\eta_I,\eta_C)])^2,\\ \quad\eta_C>0 \text{ and } \rho\geq 1.\end{cases}
\end{aligned}$$

The random variable $W_{[I]}(\eta_I,\eta_C)$ and the functions $E[W_{[I]}(\eta_i,\eta_c)]$ and $Var[W_{[I]}(\eta_i,\eta_c)]$ are characterized as follows.

Proposition 2.3 *If $\rho\eta_I < 1$, the random variable $W_{[I]}(\eta_I,\eta_C)$ is non-decreasing in η_I with respect to the stochastically larger order. In addition, we have:*

(i) *The function $E[W_{[I]}(\eta_I,\eta_C)]$ is non-decreasing in η_I and non-increasing in η_C. For fixed η_C, the function $E[W_{[I]}(\eta_I,\eta_C)]$ is convex in η_I.*

(ii) *The function $Var[W_{[I]}(\eta_I,\eta_C)]$ is non-decreasing in η_I. For fixed η_C, the function $Var[W_{[I]}(\eta_I,\eta_C)]$ is convex in η_I.*

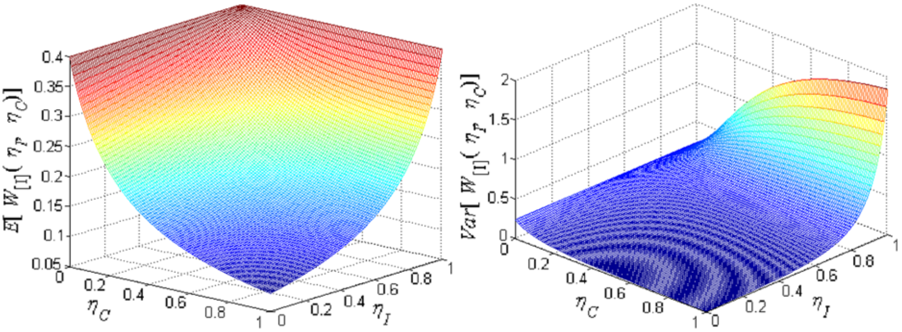


Fig. 2.2 The mean and variance of $W_{[I]}(\eta_I, \eta_C)$ for $0 \le \eta_I, \eta_C \le 1$

Proof By the explicit expressions of $E[W_{[I]}(\eta_I, \eta_C)]$ in Eq. (2.15), it is easy to see that $E[W_{[I]}(\eta_I, \eta_C)]$ is non-decreasing and convex in η_I. By Eq. (2.17) and Propositions 2.1 and 2.4, $E[W_{[I]}(\eta_I, \eta_C)]$ has to be non-increasing in η_C. (Note: Proposition 2.4 in Sect. 2.4 is used in this proof.) Part (ii) is obtained by Lemma 2.4 and Eqs. (2.13) and (2.14). This completes the proof of Proposition 2.3. □

Example 2.2 (Example 2.1 continued) We consider the queueing model with $\lambda = 8$ and $\mu = 10$. The mean and variance of $W_{[I]}(\eta_I, \eta_C)$ can be computed and are depicted in Fig. 2.2.

Figure 2.2 shows that increasing η_I will increase the mean and variance of the waiting time of an interjecting customer. Intuitively, for fixed η_C, if η_I increases, more customers interject. Then the waiting time of any customer in queue increases. On the other hand, for fixed η_I, increasing η_C will put an interjecting customer in a position closer to the head of the queue. Thus, the waiting time of an interjecting customer decreases. This is different from the waiting time of a normal customer who joins the queue at the end.

Figure 2.2 also shows that the relationship between $Var[W_{[I)}(\eta_I, \eta_C)]$ and η_C is more complicated. If η_I is close to zero, the variance is decreasing in η_C. If η_I is close to one, the variance is increasing in η_C. For a moderate η_I, $Var[W_{[I]}(\eta_I, \eta_C)]$, as a function of η_C, may not be monotone. The relationship between $E[W_{[I]}(\eta_I, \eta_C)]$, $Var[W_{[I]}(\eta_I, \eta_C)]$, and η_C indicates that a higher tolerance level of interjection is not always beneficial to interjecting customers, particularly if the percentage of interjecting customers is high.

Similar to that of $E[W_{[\mathrm{N}]}(\eta_I, \eta_C)]$ and $Var[W_{[\mathrm{N}]}(\eta_I, \eta_C)]$, explicit results can be obtained for $E[W_{[I]}(\eta_I, \eta_C)]$ and $Var[W_{[I]}(\eta_I, \eta_C)]$ at the boundary points by routine calculations. Details are omitted.

In contrast to $W_{[\mathrm{N}]}(\eta_I, \eta_C)$, $W_{[I]}(\eta_I, \eta_C)$ can be finite when the queueing system is unstable (i.e., $\rho \ge 1$). The next example illustrates the mean and variance of $W_{[I]}(\eta_I, \eta_C)$ if $\rho \ge 1$.

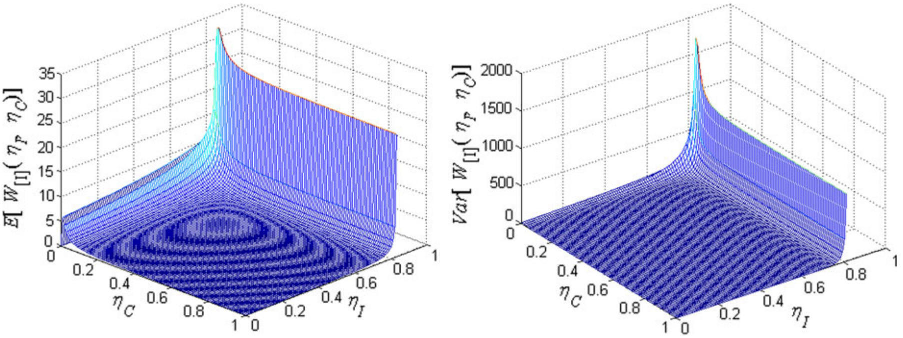


Fig. 2.3 The mean and variance of $W_{[I]}(\eta_I, \eta_C)$ if $\rho = 1.2$ for $0 \le \eta_I, \eta_C \le 1$

Example 2.3 We consider a queueing model with $\lambda = 12$ and $\mu = 10$. Although $\rho = 1.2 > 1$, if $\eta_I \rho < 1$, the waiting time of an interjecting customer is still finite (Lemma 2.1).

Figure 2.3 demonstrates that $E[W_{[I]}(\eta_I, \eta_C)]$ depends on η_I and η_C in a way similar to the case with $\rho < 1$. However, $\mathit{Var}[W_{[I]}(\eta_I, \eta_C)]$ behaves differently. For this case, the variance is always decreasing in η_C.

In summary, the mean of $W_{[I]}(\eta_I, \eta_C)$ is non-decreasing in η_I and non-increasing in η_C. The variance of $W_{[I]}(\eta_I, \eta_C)$ is non-decreasing in η_I. The variance of $W_{[I]}(\eta_I, \eta_C)$ is non-increasing in η_C if η_I is close to zero, and non-decreasing in η_C if η_I is close to one. For a unstable queue, the variance of $W_{[I]}(\eta_I, \eta_C)$ is always non-increasing in η_C.

2.4 Waiting time of an arbitrary customer

Let $W_{[\mathrm{A}]}(\eta_I, \eta_C)$ be the waiting time of an arbitrary customer. According to the definition, an arbitrary arrival is a normal customer with probability $1 - \eta_I$ and an interjecting customer with probability η_I. If $\rho < 1$, by conditioning on the type of the arbitrary arrival, we have

$$\begin{aligned} P\{W_{[\mathrm{A}]}(\eta_I, \eta_C) < t\} = (1 - \eta_I) P\{W_{[\mathrm{N}]}(\eta_I, \eta_C) < t\} \\ + \eta_I P\{W_{[I]}(\eta_I, \eta_C) < t\}. \end{aligned} \tag{2.16}$$

In terms of $E[W_{[\mathrm{N}]}(\eta_I, \eta_C)]$, $\mathit{Var}[W_{[\mathrm{N}]}(\eta_I, \eta_C)]$, $E[W_{[I]}(\eta_I, \eta_C)]$, and $\mathit{Var}[W_{[I]}(\eta_I, \eta_C)]$, we have

$$\begin{aligned} E\big[W_{[\mathrm{A}]}(\eta_I, \eta_C)\big] &= (1 - \eta_I) E\big[W_{[\mathrm{N}]}(\eta_I, \eta_C)\big] + \eta_I E\big[W_{[I]}(\eta_I, \eta_C)\big]; \\ \mathit{Var}\big[W_{[\mathrm{A}]}(\eta_I, \eta_C)\big] &= (1 - \eta_I) Var\big[W_{[\mathrm{N}]}(\eta_I, \eta_C)\big] + \eta_I \mathit{Var}\big[W_{[I]}(\eta_I, \eta_C)\big] \\ &\quad + \eta_I (1 - \eta_I)\big(E\big[W_{[\mathrm{N}]}(\eta_I, \eta_C)\big] - E\big[W_{[I]}(\eta_I, \eta_C)\big]\big)^2. \end{aligned} \tag{2.17}$$

Since the mean number of customers in queue is independent of (η_I, η_C), by Little's law, the mean waiting time is independent of (η_I, η_C). Therefore, we have

Proposition 2.4 *If* $\rho < 1$, *we have* $E[W_{[\mathrm{A}]}(\eta_I, \eta_C)] = E[W_{[\mathrm{A}]}(0, 0)] = \rho/(\mu - \lambda)$.

Proof The result can also be proved directly as follows. By Eqs. (2.10) and (2.15), we obtain

$$
\begin{aligned}
&E\big[W_{[\mathrm{A}]}(\eta_I,\eta_C)\big]\\
&\quad=\frac{1}{\mu}\sum_{k=0}^{\infty}(\rho\eta_I)^k\left[\sum_{m=1}^{\infty}\rho^m\big(1-\eta_I+\eta_I(1-\eta_C)^{m-1}\big)\prod_{j=0}^{k-1}\big(1-(1-\eta_C)^{m+j}\big)\right]\\
&\quad=\frac{\rho}{\mu-\lambda}+\frac{1}{\mu}\sum_{k=0}^{\infty}\eta_I^{k+1}\left[\sum_{m=1}^{\infty}\rho^{m+k}\big((1-\eta_C)^{m-1}-1\right.\\
&\qquad\left.+\rho\big(1-(1-\eta_C)^{m+k}\big)\big)\prod_{j=0}^{k-1}\big(1-(1-\eta_C)^{m+j}\big)\right].
\end{aligned}
\tag{2.18}
$$

The coefficient of the term η_I^{k+1}, for $k \geq 0$, can be evaluated as follows:

$$
\begin{aligned}
&\sum_{m=1}^{\infty}\rho^{m+k}\left(\big((1-\eta_C)^{m-1}-1\big)\prod_{j=0}^{k-1}\big(1-(1-\eta_C)^{m+j}\big)+\rho\prod_{j=0}^{k}\big(1-(1-\eta_C)^{m+j}\big)\right)\\
&\quad=\sum_{m=1}^{\infty}\rho^{m+k+1}\prod_{j=0}^{k}\big(1-(1-\eta_C)^{m+j}\big)-\sum_{m=1}^{\infty}\rho^{m+k}\prod_{j=-1}^{k-1}\big(1-(1-\eta_C)^{m+j}\big)\\
&\quad=\sum_{m=1}^{\infty}\rho^{m+k+1}\prod_{j=0}^{k}\big(1-(1-\eta_C)^{m+j}\big)-\sum_{m=2}^{\infty}\rho^{m+k}\prod_{j=-1}^{k-1}\big(1-(1-\eta_C)^{m+j}\big)\\
&\quad=\sum_{m=1}^{\infty}\rho^{m+k+1}\prod_{j=0}^{k}\big(1-(1-\eta_C)^{m+j}\big)-\sum_{m=1}^{\infty}\rho^{m+1+k}\prod_{j=-1}^{k-1}\big(1-(1-\eta_C)^{m+1+j}\big)\\
&\quad=\sum_{m=1}^{\infty}\rho^{m+k+1}\prod_{j=0}^{k}\big(1-(1-\eta_C)^{m+j}\big)-\sum_{m=1}^{\infty}\rho^{m+k+1}\prod_{j=0}^{k}\big(1-(1-\eta_C)^{m+j}\big)\\
&\quad=0.
\end{aligned}
\tag{2.19}
$$

In Eq. (2.19), the second equality is obtained by using $(1-\eta_C)^0 = 1$. Therefore, $E[W_{[\mathrm{A}]}(\eta_I,\eta_C)] = \rho/(\mu-\lambda)$. This completes the proof of Proposition 2.4. □

The variance of $W_{[\mathrm{A}]}(\eta_I,\eta_C)$ is exemplified by the following example.

Example 2.4 (Examples 2.1 and 2.2 continued) We consider the queueing model with $\lambda = 8$ and $\mu = 10$. The $Var[W_{[\mathrm{A}]}(\eta_I,\eta_C)]$ can be computed and is plotted in Fig. 2.4.

Figure 2.4 shows that the variance of $W_{[\mathrm{A}]}(\eta_I,\eta_C)$ is increasing in both η_I and η_C. Intuitively, increasing η_I or η_C leads to higher variability in the position of a customer (regardless of its type) in the queue. Thus, the variance of its waiting time increases.

In summary, while the mean waiting time of an arbitrary customer is not affected by customer interjection, the variability of its waiting time is increasing if there are more customers interjecting or if customers in queue are more tolerable to interjection.

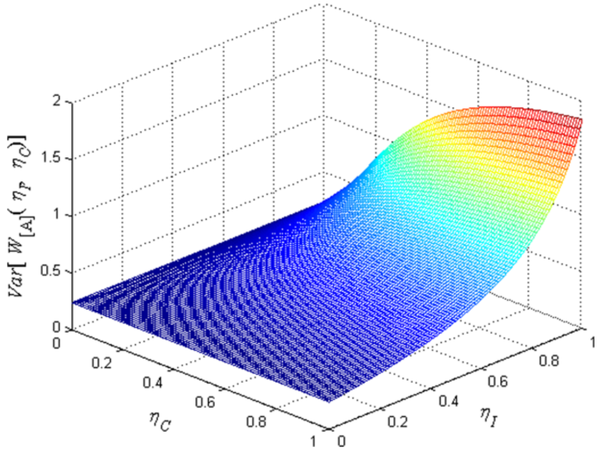


Fig. 2.4 The variance of $W_{[A]}(\eta_I, \eta_C)$ for $0 \le \eta_I$, $\eta_C \le 1$

2.5 Comparison of queueing models

In this subsection we compare the waiting times for following three queues.

(i) The classical $M/M/1$/FCFS queue. Let $W_{[C]}$ be the waiting time of an arbitrary customer.
(ii) The priority $M/M/1$ queue with two types of customers and a non-preemption service discipline. Let $W_{[H]}$ be the waiting time of a high priority customer, $W_{[L]}$ the waiting time of a low priority customer, and $W_{[HLA]}$ the waiting time of an arbitrary customer. The percentage of high priority customers is η_I.
(iii) The $M[2]/M/1$ queue with customer interjections defined in Sect. 2.1

We focus on the means and variances of the waiting times. For the first two queueing models, explicit formulas for the means and variances of the waiting times can be found in the literature. We refer to [4] for details.

Example 2.5 (Example 2.1 continued) We assume that all three queues have the same (total) arrival rate $\lambda = 8$ and service rate $\mu = 10$. For the priority queue, we assume that the percentage of high priority customer is $\eta_I = 0.3$. For the queue with interjections, we assume that $\eta_I = 0.3$, and $\eta_C = 0.1, 0.2, \ldots, 0.9, 1.0$. The means and variances of waiting times are given in the following two tables.

While the mean waiting time of an arbitrary customer is the same for all three queues, the corresponding variances are different, except for the case with $\eta_C = 0$. The low priority customers and the normal customers have different means and variances of waiting times, except for $\eta_C = 1$. For the low priority customers, their queue positions may spread out over the entire queue. Thus, the variance of the low priority class is greater than that of the normal class. The high priority customers and the interjecting customers have different means waiting time except for $\eta_C = 1$, and different variances for all η_C. For the high priority class, since the service discipline within the class is FCFS, the variance of its waiting time can be significantly smaller that of the interjecting class.

Table 2.1 Means and variances of waiting times for $\eta_I = 0.3$: I

$E[W_{[C]}]$	$Var[W_{[C]}]$	$E[W_{[L]}]$	$Var[W_{[L]}]$	$E[W_{[H]}]$	$Var[W_{[H]}]$	$E[W_{[HLA]}]$	$Var[W_{[HLA]}]$
0.4	0.24	0.5263	0.4593	0.1053	0.0166	0.4	0.3637

Table 2.2 Means and variances of waiting times for $\eta_I = 0.3$: II

η_C	0	0.2	0.4	0.6	0.8	1.0
$E[W_{[N]}]$	0.4	0.4652	0.4945	0.5104	0.5201	0.5263
$Var[W_{[N]}]$	0.24	0.3661	0.4228	0.4430	0.4534	0.4593
$E[W_{[I]}]$	0.4	0.2480	0.1795	0.1424	0.1199	0.1053
$Var[W_{[I]}]$	0.24	0.1066	0.0591	0.0396	0.0302	0.0254
$E[W_{[A]}]$	0.4	0.4	0.4	0.4	0.4	0.4
$Var[W_{[A]}]$	0.24	0.3052	0.3345	0.3504	0.3601	0.3663

In summary, Tables 2.1 and 2.2 demonstrate that the priority queues are fundamentally different from the interjecting queues. The difference is highlighted by the variance, but it is rooted in the difference between the service disciplines.

3 An *MMAP*[2]/*PH*/1 queue with customer interjections

In this section we consider an *MMAP*[2]/*PH*/1 queueing model in which customers arrive according to a marked Markovian arrival process [2, 8] and the service times have a *PH*-distribution [16]. The structure of the *MMAP*[2]/*PH*/1 queue with customer interjections is the same as the one defined in Sect. 2. Recall that η_C is the probability that a customer in queue allows the interjection of a new arrival.

We assume that the marked Markovian arrival process *MMAP*[2] has a matrix representation $\{D_0, D_1, D_2\}$, where D_0, D_1, and D_2 are $m_a \times m_a$ matrices, D_1 is for the arrival rates of normal customers, and D_2 is for the arrival rates of interjecting customers. Let $D = D_0 + D_1 + D_2$. Then D is the infinitesimal generator of the underlying (continuous time) Markov chain of the arrival process. Let $\boldsymbol{\theta}$ be the vector satisfying $\boldsymbol{\theta} D = 0$ and $\boldsymbol{\theta}\mathbf{e} = 1$. We assume that D is irreducible. Then $\boldsymbol{\theta}$ is unique and is a stochastic vector. Let $\lambda_1 = \boldsymbol{\theta} D_1 \mathbf{e}$ and $\lambda_2 = \boldsymbol{\theta} D_2 \mathbf{e}$. Then λ_1 and λ_2 are the arrival rates of normal customers and interjecting customers, respectively. If $D_1 = (1 - \eta_I)\lambda$ and $D_2 = \eta_I \lambda$, the arrival process is the same as the one used in Sect. 2.

The service times of all customers are i.i.d.r.v.s with a phase type distribution $(\boldsymbol{\beta}, S)$, where $\boldsymbol{\beta}$ is a stochastic vector of size m_s and S is a *PH*-generator of size m_s. Let $\mu = -(\boldsymbol{\beta} S^{-1}\mathbf{e})^{-1}$, which is the service rate. The service times are independent of the customer arrival process. If $\boldsymbol{\beta} = 1$ and $S = -\mu$, the service time has an exponential distribution with parameter μ.

Define $\rho = (\lambda_1 + \lambda_2)/\mu$ as the traffic intensity of the queueing system. We assume $\rho < 1$ to ensure system stability.

3.1 Stationary distributions of queue lengths

Let $I_a(t)$ be the phase of the underlying Markov process of the *MMAP*[2] at time t and $I_s(t)$ be the phase of the service at time t (if any), and zero, otherwise. Recall that $q(t)$ is the queue length at time t (including any customer in service). Again, the queue length is not affected by interjections (i.e., not affected by η_c and (D_1, D_2) as long as $D_1 + D_2$ is fixed). Then $\{(q(t), I_a(t), I_s(t)), t \geq 0\}$ represents the system state at time t and is a quasi-birth-and-death process, which is the same as the classical *MAP*/*PH*/1 queue with an arrival process $(D_0, D_1 + D_2)$ [12]:

$$Q_q = \begin{pmatrix} D_0 & (D_1+D_2)\otimes\boldsymbol{\beta} & & & & \\ I\otimes\mathbf{S}^0 & D_0\otimes I + I\otimes S & (D_1+D_2)\otimes I & & & \\ & I\otimes\mathbf{S}^0\boldsymbol{\beta} & D_0\otimes I + I\otimes S & (D_1+D_2)\otimes I & & \\ & & I\otimes\mathbf{S}^0\boldsymbol{\beta} & D_0\otimes I + I\otimes S & (D_1+D_2)\otimes I & \\ & & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{3.1}$$

where $\mathbf{S}^0 = -S\mathbf{e}$ and "$\otimes$" denotes the Kronecker product of two matrices. According to [16], under the condition $\rho < 1$, the steady state distribution $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots)$ of $\{(q(t), I_a(t), I_s(t)), t \geq 0\}$ exists and is given by

$$\boldsymbol{\pi}_n = \boldsymbol{\pi}_1 R^{n-1}, \quad n \geq 1, \tag{3.2}$$

where $(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1)$ can be obtained by solving linear system:

$$\begin{aligned} (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1)\begin{pmatrix} D_0 & (D_1+D_2)\otimes\boldsymbol{\beta} \\ I\otimes\mathbf{S}^0 & D_0\otimes I + I\otimes S + R(I\otimes\mathbf{S}^0\boldsymbol{\beta}) \end{pmatrix} &= 0; \\ \boldsymbol{\pi}_0 + \boldsymbol{\pi}_1(I-R)^{-1}\mathbf{e} &= 1, \end{aligned} \tag{3.3}$$

and R is a matrix which is the minimal nonnegative solution to

$$(D_1 + D_2)\otimes I + R(D_0\otimes I + I\otimes S) + R^2\big(I\otimes\mathbf{S}^0\boldsymbol{\beta}\big) = 0. \tag{3.4}$$

Denote by $(\boldsymbol{\pi}_{a,0}, \boldsymbol{\pi}_{a,1}, \boldsymbol{\pi}_{a,2}, \ldots)$ the joint distribution of the number of customers in the system just prior to the arrival of an arbitrary customer and the phases of the arrival and service right after the arrival of the customer. Similarly, we define $(\boldsymbol{\theta}_{1,0}, \boldsymbol{\theta}_{1,1}, \boldsymbol{\theta}_{1,2}, \ldots)$ and $(\boldsymbol{\theta}_{2,0}, \boldsymbol{\theta}_{2,1}, \boldsymbol{\theta}_{2,2}, \ldots)$ the queue length distribution seen by an arbitrary normal customer and an arbitrary interjecting customer, respectively. By a standard probabilistic argument, we obtain

$$\begin{aligned} \boldsymbol{\pi}_{a,n} &= \begin{cases} \boldsymbol{\pi}_0(D_1+D_2)(\boldsymbol{\pi}_0(D_1+D_2)\mathbf{e} + \boldsymbol{\pi}_1(I-R)^{-1}((D_1+D_2)\otimes I)\mathbf{e})^{-1}, \\ \quad n = 0 \\ \boldsymbol{\pi}_n((D_1+D_2)\otimes I)(\boldsymbol{\pi}_0(D_1+D_2)\mathbf{e} \\ \quad + \boldsymbol{\pi}_1(I-R)^{-1}((D_1+D_2)\otimes I)\mathbf{e})^{-1}, \quad n \geq 1; \end{cases} \\ \theta_{1,n} &= \begin{cases} \boldsymbol{\pi}_0 D_1(\boldsymbol{\pi}_0 D_1\mathbf{e} + \boldsymbol{\pi}_1(I-R)^{-1}(D_1\otimes I)\mathbf{e})^{-1}, & n = 0; \\ \boldsymbol{\pi}_n(D_1\otimes I)(\boldsymbol{\pi}_0 D_1\mathbf{e} + \boldsymbol{\pi}_1(I-R)^{-1}(D_1\otimes I)\mathbf{e})^{-1}, & n \geq 1; \end{cases} \\ \theta_{2,n} &= \begin{cases} \boldsymbol{\pi}_0 D_2(\boldsymbol{\pi}_0 D_2\mathbf{e} + \boldsymbol{\pi}_1(I-R)^{-1}(D_2\otimes I)\mathbf{e})^{-1}, & n = 0; \\ \boldsymbol{\pi}_n(D_2\otimes I)(\boldsymbol{\pi}_0 D_2\mathbf{e} + \boldsymbol{\pi}_1(I-R)^{-1}(D_2\otimes I)\mathbf{e})^{-1}, & n \geq 1. \end{cases} \end{aligned} \tag{3.5}$$

3.2 Mean and variance of waiting times

To analyze the waiting time of a normal or an interjecting customer, similar to Sect. 2, we first study the waiting time W_n of the n-th customer in queue initially. For that purpose, we consider the absorption time of the following Markov chain:

$$Q_a = \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \\ \vdots \end{array} \left(\begin{array}{cccccc} 0 & 0 & & & & \\ \mathbf{e} \otimes \mathbf{S}^0 & Q_{1,1} & Q_{1,2} & & & \\ & Q_{2,1} & Q_{2,2} & Q_{2,3} & & \\ & & Q_{3,2} & Q_{3,3} & Q_{3,4} & \\ & & & \ddots & \ddots & \ddots \end{array} \right) \tag{3.6}$$

where $\mathbf{S}^0 = -S\mathbf{e}$, "$\otimes$" denotes the Kronecker product of two matrices, and the matrix transition blocks are given by $Q_{i,i-1} = I \otimes (\mathbf{S}^0\boldsymbol{\beta})$, for $i = 2, 3, \dots$, and $Q_{i,i} = (D - (1 - (1 - \eta_C)^i)D_2) \oplus S$, $Q_{i,i+1} = (1 - (1 - \eta_C)^i)D_2 \otimes I$, for $i = 1, 2, \dots$. Recall that η_c is the probability that a customer in queue allows the interjection of a new arrival. Note that the state zero of Q_a is an absorption state. The waiting time of the nth customer in queue is the absorption time of state zero of Markov chain Q_a, given that Q_a is initially in level n.

To find the absorption time of the above Markov process, we first consider the fundamental period (i.e., the first passage time from level $n + 1$ to level n). Denote by $g^*_{n,i,j}(s)$ the L.S. transform of the first passage time of Q_a from the level n to the level $n - 1$ by entering into state $(n - 1, j)$, given that the state was (n, i) initially. Let $G^*_n(s) = (g^*_{n,i,j}(s))$, an $(m_a m_s) \times 1$ matrix for $n = 1$ and an $(m_a m_s) \times (m_a m_s)$ matrix for $n \geq 2$. Routinely, we have the following equations:

$$\begin{aligned} G^*_1(s) &= \big(sI - (D - \eta_C D_2) \oplus S\big)^{-1}\big(\mathbf{e} \otimes \mathbf{S}^0 + (\eta_C D_2 \otimes I)G^*_2(s)G^*_1(s)\big); \\ G^*_n(s) &= \big(sI - \big(D - \big(1 - (1 - \eta_C)^n\big)D_2\big) \oplus S\big)^{-1} \\ &\quad \cdot \big(I \otimes \mathbf{S}^0\boldsymbol{\beta} + \big(1 - (1 - \eta_C)^n\big)(D_2 \otimes I)G^*_{n+1}(s)G^*_n(s)\big), \quad n \geq 2. \end{aligned} \tag{3.7}$$

Let $G_n = \lim_{s \to 0+} G^*_n(s)$, $n \geq 1$, $\mathbf{u}^{(1)}_1 = -\lim_{s \to 0+} dG^*_1(s)/ds$, and $\mathbf{u}^{(1)}_n = -\lim_{s \to 0+} dG^*_n(s)\mathbf{e}/ds$, $n \geq 2$. By (3.7), it is easy to obtain:

$$\begin{aligned} G_n &= \begin{cases} -((D - (1 - (1 - \eta_C)^n)D_2) \oplus S \\ \quad + (1 - (1 - \eta_C)^n)(D_2 \otimes I)G_{n+1})^{-1}(\mathbf{e} \otimes \mathbf{S}^0), \quad n = 1; \\ -((D - (1 - (1 - \eta_C)^n)D_2) \oplus S \\ \quad + (1 - (1 - \eta_C)^n)(D_2 \otimes I)G_{n+1})^{-1}(I \otimes \mathbf{S}^0\boldsymbol{\beta}); \quad n \geq 2; \end{cases} \\ \mathbf{u}^{(1)}_n &= -\big(\big(D - \big(1 - (1 - \eta_C)^n\big)D_2\big) \oplus S + \big(1 - (1 - \eta_C)^n\big)(D_2 \otimes I)G_{n+1}\big)^{-1} \\ &\quad \cdot \big(\mathbf{e} + \big(1 - (1 - \eta_C)^n\big)(D_2 \otimes I)\mathbf{u}^{(1)}_{n+1}\big), \quad n \geq 1. \end{aligned} \tag{3.8}$$

Note that, since the absorption probability to level zero is one, we must have $G_1 = \mathbf{e}$ and $G_n\mathbf{e} = \mathbf{e}$, for $n \geq 2$, which is useful for checking computational accuracy.

Let $G_n^{(1)} = -\lim_{s\to 0+} dG_n^*(s)/ds$, $n \geq 1$, $\mathbf{u}_1^{(2)} = \lim_{s\to 0+} d^2G_1^*(s)/ds^2$, and $\mathbf{u}_n^{(2)} = \lim_{s\to 0+} d^2G_n^*(s)\mathbf{e}/ds^2$, $n \geq 2$. By Eq. (3.7), it is easy to obtain, for $n \geq 1$,

$$\begin{aligned} G_n^{(1)} &= -\Big(\big(D - \big(1-(1-\eta_C)^n\big)D_2\big) \oplus S + \big(1-(1-\eta_C)^n\big)(D_2 \otimes I)G_{n+1}\Big)^{-1} \\ &\quad \cdot \Big(\big(1-(1-\eta_C)^n\big)(D_2 \otimes I)G_{n+1}^{(1)} + I\Big)G_n; \\ \mathbf{u}_n^{(2)} &= -\Big(\big(D - \big(1-(1-\eta_C)^n\big)D_2 \oplus S \\ &\quad + \big(\big(1-(1-\eta_C)^n\big)D_2 \otimes I\big)G_{n+1}\Big)^{-1} \\ &\quad \cdot \Big(\Big(\big(1-(1-\eta_C)^n\big)D_2 \otimes I\Big)\mathbf{u}_{n+1}^{(2)} \\ &\quad + 2\big(I + \big(1-(1-\eta_C)^n\big)(D_2 \otimes I)G_{n+1}^{(1)}\big)\mathbf{u}_n^{(1)}\Big). \end{aligned} \tag{3.9}$$

Let $G_\infty = \lim_{n\to\infty} G_n$, $G_\infty^{(1)} = \lim_{n\to\infty} G_n^{(1)}$, $\mathbf{u}_\infty^{(1)} = \lim_{n\to\infty} \mathbf{u}_n^{(1)}$, and $\mathbf{u}_\infty^{(2)} = \lim_{n\to\infty} \mathbf{u}_n^{(2)}$. These matrices and vectors can be found by using the following equations:

$$\begin{aligned} &I \otimes \big(\mathbf{S}^0\boldsymbol{\beta}\big) + \big((D - D_2) \oplus S\big)G_\infty + (D_2 \otimes I)G_\infty^2 = 0; \\ &\big((D - D_2) \oplus S + (D_2 \otimes I)G_\infty\big)G_\infty^{(1)} + (D_2 \otimes I)G_\infty^{(1)}G_\infty = G_\infty; \\ &\mathbf{u}_\infty^{(1)} = -\big((D - D_2) \oplus S + (D_2 \otimes I)(G_\infty + I)\big)^{-1}\mathbf{e}; \\ &\mathbf{u}_\infty^{(2)} = -2\big((D - D_2) \oplus S + (D_2 \otimes I)(G_\infty + I)\big)^{-1}\big(I + (D_2 \otimes I)G_\infty^{(1)}\big)\mathbf{u}_\infty^{(1)}. \end{aligned} \tag{3.10}$$

Denote by $w_{n,j}^*(s)$ the L.S. transform of the absorption time of the level zero from the state (n, j). Let $\mathbf{w}_n^*(s) = (w_{n,j}^*(s))$, an $(m_a m_s) \times 1$ matrix for $n \geq 1$. It is easy to see that

$$\mathbf{w}_n^*(s) = G_n^*(s)G_{n-1}^*(s)\cdots G_1^*(s) = G_n^*(s)\mathbf{w}_{n-1}^*(s). \tag{3.11}$$

Denote by $w_{n,j}^{(1)}$ and $w_{n,j}^{(2)}$ the first and second moments of the absorption time of the level zero from the state (n, j). Let $\mathbf{w}_n^{(1)} = (w_{n,j}^{(1)})$ and $\mathbf{w}_n^{(2)} = (w_{n,j}^{(2)})$, an $(m_a m_s) \times 1$ matrix for $n \geq 1$. By Eq. (3.11), we obtain

$$\begin{aligned} &\mathbf{w}_1^{(1)} = \mathbf{u}_1^{(1)}; \qquad \mathbf{w}_n^{(1)} = \mathbf{u}_n^{(1)} + G_n\mathbf{w}_{n-1}^{(1)}, \quad n \geq 2; \\ &\mathbf{w}_1^{(2)} = \mathbf{u}_1^{(2)}; \qquad \mathbf{w}_n^{(2)} = \mathbf{u}_n^{(2)} + 2G_n^{(1)}\mathbf{w}_{n-1}^{(1)} + G_n\mathbf{w}_{n-1}^{(2)}, \quad n \geq 2. \end{aligned} \tag{3.12}$$

By conditioning on the system state at the arrival epoch and using Eq. (3.4), the mean and variance of the waiting time $W_{[\mathrm{N}]}$ of an arbitrary normal customer can be obtained as

$$\begin{aligned} E[W_{[\mathrm{N}]}] &= \frac{1}{\boldsymbol{\pi}_0 D_1\mathbf{e} + \boldsymbol{\pi}_1(I - R)^{-1}(D_1 \otimes I)\mathbf{e}} \sum_{n=1}^{\infty} \boldsymbol{\pi}_1 R^{n-1}(D_1 \otimes I)\mathbf{w}_n^{(1)}; \\ Var[W_{[\mathrm{N}]}] &= \frac{1}{\boldsymbol{\pi}_0 D_1\mathbf{e} + \boldsymbol{\pi}_1(I - R)^{-1}(D_1 \otimes I)\mathbf{e}} \sum_{n=1}^{\infty} \boldsymbol{\pi}_1 R^{n-1}(D_1 \otimes I)\mathbf{w}_n^{(2)} \\ &\quad - \big(E[W_{[N]}]\big)^2. \end{aligned} \tag{3.13}$$

Similarly, the mean and variance of the waiting time $W_{[\mathrm{I}]}$ of an arbitrary interjecting customer can be obtained as

$$E[W_{[\mathrm{I}]}] = \frac{\boldsymbol{\pi}_1(I+\eta_c R(I-R)^{-1})}{\boldsymbol{\pi}_0 D_2\mathbf{e}+\boldsymbol{\pi}_1(I-R)^{-1}(D_2\otimes I)\mathbf{e}}\sum_{n=1}^{\infty}(1-\eta_c)^{n-1}R^{n-1}(D_2\otimes I)\mathbf{w}_n^{(1)};$$

$$Var[W_{[\mathrm{I}]}] = \frac{\boldsymbol{\pi}_1(I+\eta_c R(I-R)^{-1})}{\boldsymbol{\pi}_0 D_2\mathbf{e}+\boldsymbol{\pi}_1(I-R)^{-1}(D_2\otimes I)\mathbf{e}} \times\sum_{n=1}^{\infty}(1-\eta_c)^{n-1}R^{n-1}(D_2\otimes I)\mathbf{w}_n^{(2)} - \big(E[W_{[I]}]\big)^2. \tag{3.14}$$

The mean and variance of the waiting time $W_{[\mathrm{A}]}$ of an arbitrary customer can be obtained as follows:

$$E[W_{[\mathrm{A}]}] = \frac{(\boldsymbol{\pi}_1(I-R)^{-1}(D_1\otimes I)\mathbf{e})E[W_{[\mathrm{N}]}]+(\boldsymbol{\pi}_1(I-R)^{-1}(D_2\otimes I)\mathbf{e})E[V_{[\mathrm{I}]}]}{\boldsymbol{\pi}_0(D_1+D_2)\mathbf{e}+\boldsymbol{\pi}_1(I-R)^{-1}((D_1+D_2)\otimes I)\mathbf{e}};$$

$$\begin{aligned}Var[W_{[\mathrm{A}]}] &= \frac{(\boldsymbol{\pi}_1(I-R)^{-1}(D_1\otimes I)\mathbf{e})Var[W_{[\mathrm{N}]}]+(\boldsymbol{\pi}_1(I-R)^{-1}(D_2\otimes I)\mathbf{e})Var[W_{[\mathrm{I}]}]}{\boldsymbol{\pi}_0(D_1+D_2)\mathbf{e}+\boldsymbol{\pi}_1(I-R)^{-1}((D_1+D_2)\otimes I)\mathbf{e}};\\ &+\frac{(\boldsymbol{\pi}_1(I-R)^{-1}(D_1\otimes I)\mathbf{e})(\boldsymbol{\pi}_1(I-R)^{-1}(D_2\otimes I)\mathbf{e})}{(\boldsymbol{\pi}_0(D_1+D_2)\mathbf{e}+\boldsymbol{\pi}_1(I-R)^{-1}((D_1+D_2)\otimes I)\mathbf{e})^2}\\ &\times\big(E[W_{[\mathrm{N}]}]-E[W_{[\mathrm{I}]}]\big)^2.\end{aligned} \tag{3.15}$$

By Little's law, the mean of $W_{[\mathrm{A}]}$ is constant for η_C, D_1, and D_2 with a fixed D_1+D_2.

Proposition 3.1 *Assume $\rho<1$. Then we have*

$$\sum_{n=1}^{\infty}(n-1)\boldsymbol{\pi}_{a,n}\mathbf{e}=\boldsymbol{\pi}_1 R(I-R)^{-2}\mathbf{e}=(\lambda_1+\lambda_2)E[W_{[\mathrm{A}]}]. \tag{3.16}$$

Equivalently, we have $\sum_{n=1}^{\infty}n\boldsymbol{\pi}_{a,n}\mathbf{e}=\boldsymbol{\pi}_1(I-R)^{-2}\mathbf{e}=(\lambda_1+\lambda_2)(E[W_{[\mathrm{A}]}]+1/\mu)$.

The relationship in Proposition 3.1 is useful for checking computational accuracy.

3.3 Numerical examples

Computation of the means and variances of waiting times can be done as follows:

Step 1: Compute λ_1, λ_2, μ, and ρ. If $\rho<1$, go to Step 2. Otherwise, stop;
Step 2: Compute $\boldsymbol{\pi}_0$, $\boldsymbol{\pi}_1$, and R by using Eqs. (3.3) and (3.4);
Step 3: Compute G_∞, $G_\infty^{(1)}$, $\mathbf{u}_\infty^{(1)}$, and $\mathbf{u}_\infty^{(2)}$ using Eq. (3.10);
Step 4: Choose N sufficiently large and set $G_N=G_\infty$, $G_N^{(1)}=G_\infty^{(1)}$, $\mathbf{u}_N^{(1)}=\mathbf{u}_\infty^{(1)}$, and $\mathbf{u}_N^{(2)}=\mathbf{u}_\infty^{(2)}$;
Step 5: Compute G_n, $G_n^{(1)}$, $\mathbf{u}_n^{(1)}$, and $\mathbf{u}_n^{(2)}$ for $n=N-1, N-2, \ldots, 2$, and 1, using Eqs. (3.8) and (3.9);

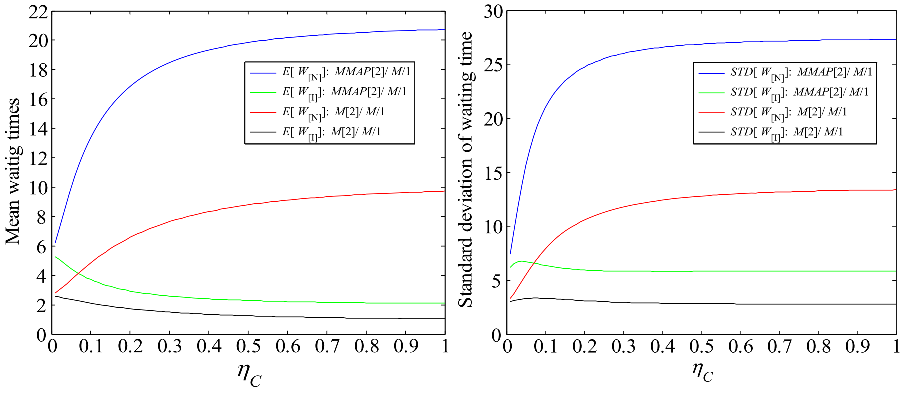


Fig. 3.1 The means and standard deviations of $W_{[\mathrm{N}]}$ and $W_{[\mathrm{I}]}$

Step 6: Compute $\mathbf{w}_n^{(1)}$ and $\mathbf{w}_n^{(2)}$ for $n = 1, 2, \ldots, N-1$, and N, using Eq. (3.12);
Step 7: Compute the mean and variance of waiting times using Eqs. (3.13), (3.14), and (3.15). Use Eq. (3.16) to check the accuracy of results.

We would like to remark that, since the queue length distribution decays geometrically, the convergence of the above computational method is guaranteed. In addition, Proposition 3.1 can be used for choosing a proper N.

Extensive numerical experiments have been carried out. Numerical results indicate that all the theoretical and numerical results obtained in Sect. 2 hold for more general cases. We do not repeat the numerical analysis conducted in Sect. 2. Instead, we use examples to explore the relationship between the waiting times, the burstiness of the customer arrival process, and the variation of the service times.

Example 3.1 We consider two queueing systems with customer interjections. The first queueing system, to be called $MMAP[2]/M/1$, has the following parameters:

$$D_0 = \begin{pmatrix} -0.8 & 0.8 \\ 0 & -5 \end{pmatrix}, \qquad D_1 = \begin{pmatrix} 0 & 0 \\ 0.5 & 0.4 \end{pmatrix}, \qquad D_2 = \begin{pmatrix} 0 & 0 \\ 0.1 & 4 \end{pmatrix}, \tag{3.17}$$

and $\boldsymbol{\beta} = 1$ and $S = -3.2$. For this queueing system, $\lambda_1 = 0.5143$, $\lambda_2 = 2.3429$, $\mu = 3.2$, and $\rho = 0.8929$. It is easy to see that the arrivals of interjecting customers are clustered (in phase two).

The second queueing system, to be called $M[2]/M/1$, has parameters: $D_1 = 0.5143$, $D_2 = 2.3429$, $\boldsymbol{\beta} = 1$, and $S = -3.2$.

It is clear that the two queueing systems have the same (average) arrival rate and service rate. In Fig. 3.1, the means and standard deviations (*STD*) of waiting times are plotted for an arbitrary normal customer.

Figure 3.1 shows that both the means and variances of the waiting times of a normal customer and an interjecting customer in the $MMAP[2]/M/1$ queue can be significantly greater than that in the $M[2]/M/1$ queue. Intuitively, in the $MMAP[2]/M/1$ queue, the arrivals of interjecting customers are clustered, which causes the increase

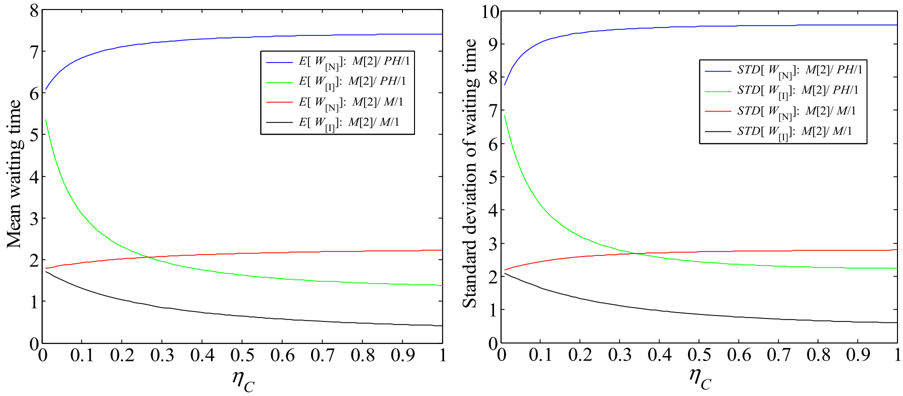


Fig. 3.2 The means and standard deviations of $W_{[\mathrm{N}]}$ and $W_{[\mathrm{I}]}$

in the means and variances of the waiting times of an arbitrary normal customer and an arbitrary interjecting customer.

Example 3.2 We consider two queueing systems with customer interjections. The first queueing system, to be called $M[2]/PH/1$, has the following parameters: $D_0 = -2$, $D_1 = 1.5$, $D_2 = 0.5$,

$$\boldsymbol{\beta} = (0.9, 0.1), \qquad S = \begin{pmatrix} -10 & 1 \\ 0 & -0.6 \end{pmatrix}. \tag{3.18}$$

For this queueing system, $\lambda_1 = 1.5$, $\lambda_2 = 0.5$, $\mu = 2.459$, and $\rho = 0.8133$. The service time has the *DFR* (decreasing failure rate) property and the coefficient of variation of the service time is $cv = 2.3818$.

The second queueing system, to be called $M[2]/M/1$, has parameters: $D_0 = -2$, $D_1 = 1.5$, $D_2 = 0.5$, $\boldsymbol{\beta} = 1$, and $S = -2.459$. It is well known that the coefficient of variation of an exponential distribution is $cv = 1$.

Similar to Fig. 3.1, Fig. 3.2 demonstrates that means and standard variations of the waiting times are significantly different for the two queueing systems. The differences are caused by the variations in the service times.

In summary, numerical examples in this section indicate that the relationship between the mean and variance of waiting times and system parameters obtained in Sect. 2 still hold for more complicated queueing systems. Numerical results also demonstrate that the mean and variance of waiting times will increase if the customer arrival process is bursty or if the service time is more variable.

4 Conclusions and discussion

This paper analyzed the $M/M/1$ queue with customer interjections. The means and variances of the waiting times of a normal customer, an interjecting customer, and an arbitrary customer are found through the LST (L.S. transform) method and matrix-analytic methods. It is shown that the intuitive relationship between the means and variances of waiting times and the two system parameters η_I and η_C, stated in Sect. 1, are indeed true. The results and observations imply that reducing the society tolerance level of interjection and the percentage of interjecting customers is always beneficial to the whole society (arbitrary customers), especially to customers playing by the rule (normal customers). By using a more general model *MMAP*[2]/*PH*/1, it is shown that the results still hold. Numerical results demonstrate that the means and variances of waiting times increase if the arrival process is bursty or if the service time is more variable.

The queueing model of interest can be extended in several directions.

(1) The level of tolerance on interjection can be significantly different between normal customers and interjecting customers: $\eta_{C,N}$ for normal customers and $\eta_{C,I}$ for interjecting customers. The analysis of such a queueing model is quite challenging since information on the types of all customers in the queue must be kept and updated.

(2) The level of tolerance on interjection can be different for customers in different positions in the queue. Assume that $\eta_{C,j}$ is the tolerance level of the customer (regardless of its type) in position j in the queue, for $j \geq 1$. On the one hand, explicit formulas for performance measures become more complex. On the other hand, the computation method developed in the Appendix can be extended for this case. Thus, the model can be analyzed numerically.

Acknowledgements The authors would like to thank two anonymous reviewers for their insightful comments and suggestions.

Appendix: An efficient algorithm for the *M*/*M*/1 case

The computation approach taken here is similar to the one used in Sect. 3. Since this is a special case, some limits are obtained explicitly and the approximations to moments of waiting times are significantly simpler. Based on the first passage time analysis [16], we develop a stable algorithm for computing the mean and variance of $W_n(\eta_I, \eta_C)$. Denote by T_n the first passage time for a customer in position n to move to position $n-1$ in queue, which is the first passage time from level n to $n-1$ for the Markov chain Q_a defined in Eq. (2.2).

Similar to the proof of Lemma 2.2, it can be shown that T_n is stochastically larger in n (also see [3]). Thus, $E[T_n]$ and $E[T_n^2]$ are increasing in n. Let $T_\infty = \lim_{n\to\infty} T_n$. The variable T_∞ can be considered as the length of the busy period of an *M*/*M*/1 queue with arrival rate $\lambda\eta_I$ and service rate μ. It is readily seen that

$$\begin{aligned} E\bigl[e^{-sT_n}\bigr] &= \frac{\mu + \lambda\eta_I(1-(1-\eta_C)^n)E[e^{-sT_{n+1}}]E[e^{-sT_n}]}{s+\mu+\lambda\eta_I(1-(1-\eta_C)^n)}, \quad n \geq 1; \\ E\bigl[e^{-sT_\infty}\bigr] &= \frac{\mu + \lambda\eta_I(E[e^{-sT_\infty}])^2}{s+\mu+\lambda\eta_I}. \end{aligned} \tag{A.1}$$

$E[T_\infty]$ and $E[T_\infty^2]$ are finite and are given by:

$$\begin{aligned} E[T_\infty] &= \frac{1}{\mu - \lambda\eta_I}, \\ E\big[T_\infty^2\big] &= \frac{2E[T_\infty](1+\lambda\eta_I E[T_\infty])}{\mu-\lambda\eta_I} = \frac{2\mu}{(\mu-\lambda\eta_I)^3}. \end{aligned} \tag{A.2}$$

The following relationships of $\{T_n, n \geq 1\}$ can be shown routinely:

$$\begin{aligned} E[T_n] &= \frac{1}{\mu} + \frac{\lambda\eta_I(1-(1-\eta_C)^n)}{\mu}E[T_{n+1}] \\ &= \frac{1}{\mu}\lim_{K\to\infty}\left(\sum_{k=0}^{K-1}(\rho\eta_I)^k\left(\prod_{j=0}^{k-1}\big(1-(1-\eta_C)^{n+j}\big)\right)\right. \\ &\quad \left. + (\rho\eta_I)^K\left(\prod_{j=0}^{K-1}\big(1-(1-\eta_C)^{n+j}\big)\right)E[T_{n+K}]\right) \\ &= \frac{1}{\mu}\lim_{K\to\infty}\sum_{k=0}^{K-1}(\rho\eta_I)^k\left(\prod_{j=0}^{k-1}\big(1-(1-\eta_C)^{n+j}\big)\right) \\ &\quad + \frac{1}{\mu}\lim_{K\to\infty}(\rho\eta_I)^K\left(\prod_{j=0}^{K-1}\big(1-(1-\eta_C)^{n+j}\big)\right)E[T_{n+K}] \\ &= \frac{1}{\mu}\sum_{k=0}^{\infty}(\rho\eta_I)^k\left(\prod_{j=0}^{k-1}\big(1-(1-\eta_C)^{n+j}\big)\right). \end{aligned} \tag{A.3}$$

The second term on the third line of Eq. (A.3) is zero since $\rho\eta_I < 1$ and $E[T_\infty]$ is finite. Similarly, we can obtain

$$\begin{aligned} E\big[T_n^2\big] &= \frac{2E[T_n] + \lambda\eta_I(1-(1-\eta_C)^n)(E[T_{n+1}^2] + 2E[T_n]E[T_{n+1}])}{\mu} \\ &= \frac{2}{\mu}\sum_{k=0}^{\infty}(\rho\eta_I)^k E[T_{n+k}]\big(1+\lambda\eta_I\big(1-(1-\eta_C)^{n+k}\big) \\ &\quad \times E[T_{n+k+1}]\big)\left(\prod_{j=0}^{k-1}\big(1-(1-\eta_C)^{n+j}\big)\right); \\ Var[T_n] &= \left(\frac{1}{\mu} + \frac{\lambda\eta_I(1-(1-\eta_C)^n)E[T_{n+1}]}{\mu}\right)^2 + \frac{\lambda\eta_I(1-(1-\eta_C)^n)}{\mu}E\big[T_{n+1}^2\big]. \end{aligned} \tag{A.4}$$

By Eqs. (A.3) and (A.4), it is easy to obtain the following result.

Lemma A.1 *The functions $E[T_n]$, $E[T_n^2]$, and $Var[T_n]$ are non-decreasing functions of η_I and η_C. In addition, the three functions are convex in η_I.*

Since $W_n(\eta_I, \eta_C) = T_1 + T_2 + \cdots + T_n$, we have

$$E\big[W_n(\eta_I, \eta_C)\big] = E[T_1] + E[T_2] + \cdots + E[T_n] = E\big[W_{n-1}(\eta_I, \eta_C)\big] + E[T_n];$$

$$\begin{aligned} E\big[W_n^2(\eta_I, \eta_C)\big] &= \sum_{j=1}^{n} E\big[T_j^2\big] + 2 \sum_{1\le i<j\le n} E[T_i]E[T_j] \\ &= E\big[W_{n-1}^2(\eta_I, \eta_C)\big] + E\big[T_n^2\big] + 2E\big[W_{n-1}(\eta_I, \eta_C)\big]E[T_n]. \end{aligned} \tag{A.5}$$

Since the variance of $W_n(\eta_I, \eta_C)$ is given by $Var[W_n(\eta_I, \eta_C)] = \sum_{i=1}^{n} Var[T_n]$, we obtain the following lemma.

Lemma A.2 *The functions $E[W_n(\eta_I, \eta_C)]$ and $Var[W_n(\eta_I, \eta_C)]$ are non-decreasing in both η_I and η_C. In addition, the two functions are convex in η_I.*

To approximate the mean and variance of W_n, choose sufficiently large N and set $E[T_N] = E[T_\infty]$ and $E[T_N^2] = E[T_\infty^2]$. Then $E[T_n]$ and $E[T_n^2]$ can be computed by using the formulas in Eqs. (A.3) and (A.4) for $n \le N$. Using the formulas in Eq. (A.5), $E[W_n]$ and $E[W_n^2]$ can be computed for $n \ge 1$.

Note that, by the monotonicity property of $\{T_n, n \ge 0\}$, we have $E[W_n] \le nE[T_\infty]$ and $E[W_n^2] \le nE[T_\infty^2] + n(n-1)(E[T_\infty])^2$. Thus, we can choose large enough N so that the error in computing the first two moments of waiting times, such as $E[W_{[\mathrm{N}]}(\eta_I, \eta_C)]$ and $E[(W_{[\mathrm{N}]}(\eta_I, \eta_C))^2]$, can be smaller than any given positive number.

References

1. Afeche, P., Mendelson, H.: Pricing and priority auctions in queueing systems with a generalized delay cost structure. Manag. Sci. **50**, 869–882 (2004)
2. Asmussen, S., Koole, G.: Marked point processes as limits of Markovian arrival streams. J. Appl. Probab. **30**, 365–372 (1993)
3. Chen, H., Yao, D.D.: Fundamentals of Queueing Networks. Springer, New York (2001)
4. Cohen, J.W.: The Single Server Queue. North-Holland, Amsterdam (1982)
5. Gordon, E.S.: New problems in queues: social injustice and server production management. Ph.D thesis, MIT (1987)
6. Grassmann, W.: Means and variances of time averages in Markovian environments. Eur. J. Oper. Res. **31**, 132–139 (1987)
7. Hassin, R., Haviv, Mo.: To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems. Kluwer Academic, Boston (1999)
8. He, Q.-M., Neuts, M.: Markov chains with marked transitions. In: Stochastic Processes and Their Applications, vol. 74, pp. 37–52 (1998)
9. Kleinrock, L.: Optimal bribing for queue position. Oper. Res. **15**, 304–318 (1967)
10. Kleinrock, L.: Queueing Systems, vol. I: Theory. Wiley, New York (1975)
11. Larson, R.C.: Perspectives on queues: social justice and the psychology of queueing. Oper. Res. **35**(6), 895–905 (1987)
12. Latouche, G., Ramaswami, V.: Introduction to Matrix Analytic Methods in Stochastic Modelling. ASA & SIAM, Philadelphia (1999)
13. Lui, F.T.: An equilibrium queueing model of bribery. J. Polit. Econ. **93**, 760–781 (1985)
14. Marshall, A.W., Olkin, I., Arnold, B.: Inequalities: Theory of Majorization and Its Applications. Springer, New York (2010)
15. Melamed, B., Yadin, M.: Randomization procedures in the computation of cumulative-time distributions over discrete state Markov processes. Oper. Res. **32**, 926–944 (1984)

16. Neuts, M.F.: Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach. Johns Hopkins University Press, Baltimore (1981)
17. Neuts, M.F.: Generalizations of the Pollaczek-Khinchin integral method in the theory of queues. Adv. Appl. Probab. **18**, 952–990 (1986)
18. Ross, S.: Stochastic Processes. Wiley, New York (1983)
19. Seneta, E.: Non-negative Matrices and Markov Chains. Springer, New York (2006)
20. Shaked, M., Shanthikumar, J.G.: Stochastic Orders. Springer, New York (2006)
21. Takagi, H.: Queueing Analysis: A Foundation of Performance Evaluation, vol. 1: Vacation and Priority Systems, Part 1. Elsevier, Amsterdam (1990)
22. Takine, T., Hasegawa, T.: The workload in the MAP/G/1 queue with state-dependent services its application to a queue with preemptive resume priority. Stoch. Models **10**, 183–204 (1994)
23. Whitt, W.: Deciding which queue to join: some counterexample. Oper. Res. **34**, 55–62 (1983)
24. Whitt, W.: The amount of overtaking in a network of queues. Networks **14**, 411–426 (1984)
25. Zhao, Y., Grassmann, W.K.: Queueing analysis of a jockeying model. Oper. Res. **43**, 520–529 (1995)