Queueing Systems 37, 337–347, 2001

Inequalities for Queues with a Learning Server

EROL A. PEKÖZ* and MICHAEL LAPRÉ pekoz@bu.edu
Operations Management Department, School of Management, Boston University, 595 Commonwealth Ave, Boston, MA 02215, USA

Received 12 July 1999; Revised 26 June 2000

Abstract. We study a multi-class queue with a "learning" server who becomes stochastically faster with each subsequent customer served of the same type in a row, and returns to some baseline speed each time he switches to a different type of customer. We show under some conditions that customer waiting time is larger (in the increasing convex ordering sense) with server learning than in a queue with iid service times having the same marginal service distribution as the learning server. An easy to evaluate inequality for the mean stationary waiting time is derived from this in the case of Poisson arrivals, and results in more general settings are given. The primary tool used in the proofs is the supermodularity of the delay in queue as a function of previous service times.

1. Introduction

In this article we first consider a $GI/G/1$ queue with iid inter-arrival times and a service distribution which depends on the "state of learning" of the server. Letting X_n and I_n respectively denote the service time and the "state of learning" corresponding to the nth service performed, we suppose that when $I_n = i$ the service time X_n is independently generated so as to have cumulative distribution function $F_i(\cdot)$. By this we mean

$$P(X_n \leqslant x \mid I_n = i) = F_i(x),$$

and that $X_1, \ldots, X_n$ are conditionally independent given the values $I_1, \ldots, I_n$. Our interest here is the case where I_n is the stationary Markov chain with transition probabilities given by

$$P(I_{n+1} = i + 1 \mid I_n = i) = p = 1 - P(I_{n+1} = 0 \mid I_n = i), \quad i = 0, 1, 2, \ldots,$$

started according to its stationary probabilities

$$\pi_i \equiv P(I_1 = i) = (1 - p)p^i, \quad i = 0, 1, 2, \ldots.$$

The motivation for this model is in its application to study a first-come-first-served multi-class queue where the server temporarily learns to work faster with experience. Here an arriving customer independently is of type i, $i = 1, 2, \ldots N$, with probability p_i. The server's "state of learning" increases by one with each successive customer

* Corresponding author.

served of the same type in a row, but returns to zero each time the server switches over to serve a customer of a different type than the one previously served. When the server's "state of learning" equals i, service times are distributed having cdf $F_i(\cdot)$. We suppose that $F_i(x)$ is nondecreasing in i for every x, indicating that a higher state of learning results in stochastically smaller service times. In the case of equally likely customer types with $p \equiv p_1 = p_2 = \cdots = p_N$, the next customer is of the same type as the previous customer with probability p and so the "state of learning" evolves like the Markov chain I_n. In the case of non-equal p_i the "state of learning" will be stochastically larger than I_n using $p = \min_i p_i$, and we will see that this allows lower bounds on waiting times to be obtained.

In some sense this model could be called a "learning and forgetting" server who learns to be faster with experience but forgets everything each time he switches to a new type of customer. Or this can be viewed as a natural generalization of a multiclass queue with setup times, where we allow the setup work to be spread over several customer service times. We opt to call this model the "learning server queue". This model appears to be new; we have not been able to find other study in the literature of queues with this type of Markov-modulated service distribution.

In this article we study the waiting time W_n in queue for the nth customer, and develop stochastic comparison results which can be used to obtain lower bounds for moments of W_n and gain insight about the behavior of this system. We also give results for some natural generalizations of the model.

There have been recent efforts to compare the performance of "dedicated" servers able only to perform a single type of task with the performance of "flexible" servers able to perform more than one type of task. This is related to the issue of deciding when to "pool" queues in a service center, and when to have "dedicated" queues. The recent [6] compares average waiting times for dedicated and flexible servers in the $M/PH/1$ setting. Though it is not considered there, with flexible servers the issue of modeling switching between tasks and learning from experience naturally becomes important. A model where service times can depend on some other factor is useful for this.

There has been some study of queues where the service distribution can depend on some other factor. In [5] a class of $M/M/1$ queueing models are studied in which the service time of a customer depends on the number of customers served in the current busy period. In [3,11–13] models are studied where the service times depend on the size of the queue. See also [8] for a different model of dependence. There has also been some related study of queues with dependent inter-arrival times in [1,2]. On the issue of comparing specialized and dedicated servers is the study [4] of competition among specialized servers. Our model is different from the above models in that we allow the service time to depend on the number of customers served in a row of the same type. Approaches used for the other models do not seem to be applicable to study our model, and our approach does not seem to be obviously applicable to the other models. In short, we have not been able to find other study of our model for a learning-server, or results for waiting times with dependent service times under such general settings.

In section 2 we present the main results and give a corollary for the stationary learning queue with Poisson arrivals. In section 3 we present the proofs of the results, in section 4 we give some results for some more general settings, and in section 5 we give a summary of our conclusions.

2. Main results

Below we use the terms "increasing"and "decreasing" to mean, respectively, "non-decreasing" and "non-increasing". Our main result involves comparing the learning server queue defined above with a related $GI/GI/1$ queue, which we refer to as the corresponding "iid queue". This iid queue has the same iid inter-arrival times as the learning queue and has iid service times which have the same marginal distribution as the learning queue. In this "iid queue" the service time X_n^* for the nth customer is independent of everything else and has the same unconditional distribution as X_n, and will hence be distributed according to a mixture of the F_i mixed according to the stationary distribution for I_n. This means that the X_n^* are iid with

$$X_n^* =_{\text{st}} X_n$$

and hence

$$P(X_n^* \leqslant x) = \sum_{i=0}^{\infty} \pi_i F_i(x) = P(X_n \leqslant x). \tag{1}$$

Note that we are not comparing the learning server to the obviously inferior server who always serves according to the cdf F_0, but to a server having the same unconditional service distribution as the server in the learning server queue. In some sense this iid queue can be considered a reasonably close approximation to the learning server queue, as the service times will have the same distribution. The only difference is that service times in the learning server queue will be correlated, whereas in the iid queue they will be independent. This iid queue has the advantage that it can be analyzed using the usual formulas of queueing theory.

Our main result we prove below says that the nth customer's waiting time W_n for the learning server queue is larger than the nth customer's waiting time W_n^* for the corresponding iid queue in the increasing and convex ordering sense (a random variable X is smaller than Y in the *increasing convex ordering*, denoted by $X \leqslant_{\text{icx}} Y$, if $E[f(X)] \leqslant E[f(Y)]$ for all increasing and convex functions f for which the expectations exist). This is summarized in the following theorem.

Theorem 2.1. Let $I_1, I_2, \ldots$ be a stationary Markov chain defined by $p = P(I_{n+1} = i+1 \mid I_n = i) = 1 - P(I_{n+1} = 0 \mid I_n = i)$, $i = 0, 1, 2, \ldots$. Let $F_i(\cdot)$, $i = 0, 1, 2, \ldots$, be cumulative distribution functions so that $F_i(x)$ is increasing in i for every x. Let W_n be the nth customer's waiting time for a single-server queue (the "learning server queue") with general iid inter-arrival times, and service times $X_1, X_2, \ldots$ so that for

each n we have $P(X_n \leqslant x \mid I_n = i) = F_i(x)$ and that $X_1, \ldots, X_n$ are conditionally independent given the values $I_1, \ldots, I_n$. Finally, let W_n^* be the nth customer's waiting time for a $GI/GI/1$ queue (the corresponding "iid queue") with the same iid inter-arrival times and iid service times $X_1^*, X_2^*, \ldots$ each having $X_i^* =_{\text{st}} X_i$, for $i = 1, 2, \ldots$. Then $W_n^* \leqslant_{\text{icx}} W_n$.

We would like to stress again that clearly the stationary waiting time in the learning server queue tends to be smaller than the stationary waiting time in an iid queue where the service distribution is F_0, and our result does not contradict this and is more subtle in that the iid queue we use has the same unconditional service distribution as the learning server queue.

This result has a natural interpretation in the context of simulating a queue using historical data. Suppose an engineer is interested in estimating the average stationary delay experienced by customers in a single-server queue. Suppose the engineer has measured the historical service and inter-arrival time distributions using historical empirical data. The engineer decides to estimate the stationary customer delay by simulating a $GI/GI/1$ queue having iid service and inter-arrival time distributions following these same historical distributions, and will measure the customer delay from the simulation. Now suppose in actuality, unknown to the engineer, the service times in the original system were not iid but there was a "learning" server (by "learning server" we mean the server becomes stochastically faster with each successive similar type of task in a row, and returns to a baseline level whenever the current task is different from the previous task). Assume that the inter-arrival times were iid. Our main result essentially says that the simulation will tend to *underestimate* the actual delay experienced in the original system (in the increasing convex ordering sense). Our result gives some insight on what happens if you assume that service times were created by an iid process, when in actuality they were created by a "learning server" process.

The intuition we have for our result is that the presence of this type of "learning" server creates a positive correlation between successive service times. This positive correlation causes extra variability in the sum of the required service times of customers seen in queue by newly arriving customers, and this extra variability tends to increase queueing delays. To detect the presence of a "learning" server from historical data, without having data on the customer types, one presumably would need to look at the correlations between successive service times. Our result is interesting in that it says that if you neglect to check for the effects of learning, actual experienced delays tend to be larger than you might expect if you assumed the same service times were generated in an iid fashion.

From theorem 2.1 we also get the following corollary, which gives an easy to evaluate lower bound for the mean waiting time in a learning-server queue with Poisson arrivals. It follows easily using formulas for iid queues.

Corollary 2.1. Using the definitions from theorem 2.1, let $W = \lim_{n\to\infty} W_n$ be the stationary waiting time for the learning-server queue with Poisson rate λ arrivals. Let

$\mu_i^{(1)}$ and $\mu_i^{(2)}$ respectively be the first and second moments of the service distribution having cdf $F_i(\cdot)$. Then

$$E[W] \geqslant \frac{\lambda m^{(2)}}{2(1-\lambda m^{(1)})},$$

where

$$m^{(k)} = \sum_{i \geqslant 0} p^i (1-p) \mu_i^{(k)}, \quad k = 1, 2.$$

Proof. In the notation of theorem 2.1, the corresponding iid queue will be a stationary $M/GI/1$ queue with stationary waiting time W^* having mean

$$E[W^*] = \frac{\lambda m^{(2)}}{2(1-\lambda m^{(1)})}$$

(see [9, p. 383]) where $m^{(1)}$ and $m^{(2)}$ are the first two moments of the service distribution. By (1) we have

$$m^{(k)} = \sum_{i=0}^{\infty} \pi_i \mu_i^{(k)}, \quad k = 1, 2,$$

and the corollary follows by sending $n \to \infty$ in theorem (2.1) and using the fact that $W^* \leqslant_{\text{icx}} W \Rightarrow E[W^*] \leqslant E[W]$. □

It is interesting to note that both the theorem and corollary above also hold in the case where $F_i(x)$ is decreasing in i for every fixed x, which might be called a "forgetting server queue". In this setting the server becomes slower with each successive customer of the same type served in a row. In this case successive service times will still be positively correlated, and it is not surprising that the inequality in theorem 2.1 still holds in the same direction.

3. Proof of the main results

To prove the main results we first need some definitions and lemmas.

Definition 3.1.

1. A function $f : \mathcal{R}^n \to \mathcal{R}$ is called supermodular if, for all $x, y \in \mathcal{R}^n$,

$$f(x \vee y) + f(x \wedge y) \geqslant f(x) + f(y),$$

where $x \vee y$ ($x \wedge y$) denotes the componentwise maximum (minimum) of x, y.

2. A random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is said to be smaller than the random vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ in the *supermodular stochastic order* (denoted by $\mathbf{X} \leqslant_{\text{sm}} \mathbf{Y}$) if $E[f(\mathbf{X})] \leqslant E[f(\mathbf{Y})]$ for all supermodular functions f for which the expectations exist.

Our approach to proving theorem 2.1 can be summarized as follows. We first show that the vector of service times for the learning server queue is larger in the supermodular ordering than the vector of service times for the corresponding iid queue, and then we show that the waiting time of the nth customer in the queue is an increasing supermodular function of the previous customers' service times. The theorem follows essentially from these two facts. We present these steps in some lemmas.

Lemma 3.1. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and $\mathbf{X}^* = (X_1^*, X_2^*, \ldots, X_n^*)$ be the vectors of service times of the first n customers respectively in the learning server queue and the corresponding iid queue described in theorem 2.1. Then $\mathbf{X}^* \leqslant_{\text{sm}} \mathbf{X}$.

Proof. The proof will be by induction. The case $n = 1$ holds trivially. Next for $n > 1$ let $f : \mathcal{R}^n \to \mathcal{R}$ be supermodular. We will first define on the same probability space coupled n-dimensional vectors $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$, and $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)$ so that

$$\mathbf{X} =_{\text{st}} \mathbf{Y} =_{\text{st}} \mathbf{Z} \tag{2}$$

and either

$$Y_i \leqslant Z_i, \quad \text{for } i = 1, 2, \ldots, n, \tag{3}$$

or

$$Y_i \geqslant Z_i, \quad \text{for } i = 1, 2, \ldots, n \tag{4}$$

holds almost surely. This means the values of $\mathbf{Y}$ are either all above, or all below the corresponding values of $\mathbf{Z}$.

In order to do this we start with the "state of learning" vector $\mathbf{I} = (I_1, I_2, \ldots, I_n)$ given in the theorem and create a vector $\mathbf{J} = (J_1, J_2, \ldots, J_n)$ coupled to it so that

$$\mathbf{J} =_{\text{st}} \mathbf{I} \tag{5}$$

as follows. First, let J_1 be, independent of all else, distributed according to I_1. Then for each value of $i = 2, 3, \ldots, n$, if $I_i = I_{i-1} + 1$ then let $J_i = J_{i-1} + 1$ and otherwise we have $I_i = 0$ and we then let $J_i = 0$. Clearly this gives (5) and also that I_1 is independent of $J_2, \ldots, J_n$, and that either

$$I_i \leqslant J_i, \quad \text{for } i = 1, 2, \ldots, n, \tag{6}$$

or

$$I_i \geqslant J_i, \quad \text{for } i = 1, 2, \ldots, n \tag{7}$$

holds almost surely.

Next we create $\mathbf{Y}$ and $\mathbf{Z}$ from these. First, let $G(k, x) = \inf\{y\colon F_k(y) \geqslant x\}$, and note that this will give $G(k, x) = F_k^{-1}(x)$ where defined. Then let U_i, $i = 1, 2, \ldots, n$, be iid uniform (0,1) random variables, and let $Y_i = G(I_i, U_i)$ and $Z_i = G(J_i, U_i)$. It can be easily verified that this gives (2) since for any value of x, where $F_k(y) = x$ we have that $P(G(k, U_i) \leqslant y) = P(U_i \leqslant x) = x$.

Since $F_i(x)$ is increasing in i for every x, this implies $G(i, x)$ is decreasing in i for every x, and we can use (6) and (7) to deduce that either (3) or (4) must hold. (It is interesting to note that in the case where $F_i(x)$ is decreasing in i for fixed x either (3) or (4) must still hold because then $G(i, x)$ will be increasing.)

By the fact that either (3) or (4) hold and the definition of the supermodular function it holds almost surely that

$$f(Y_1, Y_2, \ldots, Y_n) + f(Z_1, Z_2, \ldots, Z_n) \geqslant f(Y_1, Z_2, \ldots, Z_n) + f(Z_1, Y_2, \ldots, Y_n)$$

and therefore (using (2))

$$\begin{aligned} E\big[f(X_1, X_2, \ldots, X_n)\big] &\geqslant E\big[f(Z_1, Y_2, \ldots, Y_n)\big] = E\big[E\big[f(Z_1, Y_2, \ldots, Y_n)\big] \mid Z_1\big] \\ &\geqslant E\big[E\big[f\big(Z_1, X_2^*, \ldots, X_n^*\big)\big] \mid Z_1\big] = E\big[f\big(X_1^*, X_2^*, \ldots, X_n^*\big)\big], \end{aligned}$$

where the third line follows from the induction hypothesis, the fact that a supermodular function with the first coordinate fixed is also a supermodular function in the remaining coordinates, and the fact that by construction Z_1 is independent of $Y_2, \ldots, Y_n$. □

We next need the following lemma.

Lemma 3.2. The function f_n defined recursively as follows is increasing and supermodular for each n. For fixed $y_1, y_2, \ldots$ let $f_1(x_1) = 0$ and for $n = 1, 2, \ldots$ let $f_{n+1}(x_1, x_2, \ldots, x_{n+1}) = \max(0, f_n(x_1, x_2, \ldots, x_n) - y_n + x_n)$.

Proof. The proof is by induction. We assume $y_1, \ldots, y_n$ are fixed. Clearly f_1 is supermodular and increasing. Lemma 2.6.1(b) in [14] states that if f, g are supermodular then so is $f + g$, and lemma 2.2 in [1] states that if f is supermodular then $\max(f, 0)$ is supermodular. If we assume f_n is supermodular and increasing, and we easily see that $-y_n + x_n$ is supermodular and increasing in x_n, we use the two facts above to conclude that f_{n+1} is supermodular and increasing. □

Note 1. Note that (see [10, p. 333]) if $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$ respectively are service times and inter-arrival times, the waiting time of the nth customer equals $f_n(x_1, x_2, \ldots, x_n)$. The lemma thus states that when the inter-arrival times are considered fixed, the waiting time of the nth customer is a supermodular function of the previous service times. In [1] a similar argument is used to show that the waiting time of the nth customer, when the service times are considered fixed, is a supermodular function of the previous inter-arrival times.

One final lemma is needed.

Lemma 3.3. Suppose $\mathbf{X} \leqslant_{\text{sm}} \mathbf{Y}$, and let $f : \mathcal{R}^n \to \mathcal{R}$ be any increasing and supermodular function. Then $f(\mathbf{X}) \leqslant_{\text{icx}} f(\mathbf{Y})$.

Proof. See [1, lemma 2.3]. □

We are now ready to prove theorem 2.1.

Proof of theorem 2.1. Suppose the interarrival times are $Y_1, Y_2, \ldots, Y_n$. By note 1 above we have that

$$W_n = f_n(X_1, X_2, \ldots, X_n)$$

and

$$W_n^* = f_n(X_1^*, X_2^*, \ldots, X_n^*)$$

when we plug $(y_1, \ldots, y_n) = (Y_1, \ldots, Y_n)$ into the increasing and supermodular function f_n defined in lemma 3.2.

We can therefore use lemma 3.2 along with lemma 3.3 conditional on $(Y_1, \ldots, Y_n)$ to get

$$E[c(W_n^*) \mid Y_1 = y_1, \ldots, Y_n = y_n] \leqslant E[c(W_n) \mid Y_1 = y_1, \ldots, Y_n = y_n]$$

for any increasing convex function $c(\cdot)$. Unconditioning gives the theorem. □

Note 2. In the case of non-equally likely customer types having probabilities $p_1, \ldots, p_N$ the number of customers K_n of the same type served in a row prior to the nth customer (the "state of learning" of the server) will not be a Markov chain. In this case the argument of theorem 2.1 does not seem to go through. Even if the state space is expanded to include the customer type (so the resultant "state of learning" will be a Markov chain) it does not seem possible to obtain a coupling so that (6) or (7) holds.

We can, however, get a lower bound on W_n', the waiting time of the nth customer in this non-equally likely case, by applying theorem 2.1 using $p = \min_i p_i$. In this case $I_n \leqslant_{\text{st}} K_n$ and thus $W_n \leqslant_{\text{st}} W_n'$ (and hence also $W_n \leqslant_{\text{icx}} W_n'$) and we therefore from the theorem can obtain the lower bound $W_n^* \leqslant_{\text{icx}} W_n'$.

4. Some generalizations

We refer to the model in theorem 2.1 as "Model 1" and next give some results for two more general single-server queueing models. Before we state the results we need a definition from [7].

Definition 4.1. A random vector $\mathbf{I} = (I_1, I_2, \ldots, I_n)$ is said to be conditionally increasing in sequence (CIS) if, for $j = 2, 3, \ldots, n$, we have that $P(I_i > x \mid I_1 = i_1, I_2 = i_2, \ldots, I_{j-1} = i_{j-1})$ is increasing in $i_1, i_2, \ldots, i_{j-1}$ for all x. A sequence $I_1, I_2, \ldots$ is a CIS sequence if the vector $(I_1, I_2, \ldots, I_n)$ is CIS for each n.

Example. The sequence of the number of customers of the same type served in a row in the learning server queue above can be easily seen to be a CIS sequence when the

customer types are equally likely, but is not necessarily CIS when the customer types are not equally likely.

Model 2. Consider a queue where X_n, Y_n are the service time and inter-arrival time of the nth customer. Let $(I_n)_{n=1,2,\ldots}$ and $(J_n)_{n=1,2,\ldots}$ be independent stationary CIS sequences and suppose that the service time and inter-arrival time distributions depend on these so that

$$(X_n \mid I_n = i) =_{\text{st}} S_i$$

and

$$(Y_n \mid J_n = i) =_{\text{st}} A_i,$$

where $P(S_i \leqslant x)$ is either increasing or decreasing in i for all x, and $P(A_i \leqslant x)$ is either increasing or decreasing in i for all x. We also suppose that the $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are independent given $(I_1, \ldots, I_n)$ and $(J_1, \ldots, J_n)$. Let W_n be the waiting time for the nth customer in this queue. Next define an 'iid queue" by letting the nth customer have service time distributed so that $X_n^* =_{\text{st}} X_n$ and inter-arrival time distributed so that $Y_n^* =_{\text{st}} Y_n$, and be independent of all else. Let W_n^* be the waiting time for the nth customer in this "iid queue".

Theorem 4.1. For model 2,

$$W_n^* \leqslant_{\text{icx}} W_n.$$

Note 3. Model 1 can be viewed as a special case of model 2. We list the results separately because model 1 has obvious practical significance, whereas model 2 (and model 3 below) do not, but however are natural generalizations of model 1.

Model 3. Consider a queue where X_n, Y_n are the service time and inter-arrival time of the nth customer. Let $(I_n)_{n=1,2,\ldots}$ be a stationary CIS sequence and suppose that

$$(X_n, Y_n \mid I_n = i) =_{\text{st}} (S_i, A_i),$$

where $P(S_n - A_n \leqslant x)$ is either increasing or decreasing in n for all x. Let W_n be the waiting time for the nth customer in this queue. Next define an 'iid queue" with $X_n^* =_{\text{st}} X_n$ and $Y_n^* =_{\text{st}} Y_n$, and be independent of all else. Let W_n^* be the waiting time for the nth customer in this "iid queue".

Theorem 4.2. For model 3,

$$W_n^* \leqslant_{\text{icx}} W_n.$$

Lemma 4.1. The functions g_n, and h_n defined recursively below are increasing and supermodular for each n:

1. For fixed $x_1, x_2, \ldots$ let $g_1(y_1) = 0$ and for $n = 1, 2, \ldots$ let

$$g_{n+1}(y_1, y_2, \ldots, y_{n+1}) = \max\bigl(0, g_n(y_1, y_2, \ldots, y_n) + y_n + x_n\bigr). \tag{8}$$

2. Let $h_1(x_1) = 0$ and for $n = 1, 2, \ldots$ let

$$h_{n+1}(x_1, x_2, \ldots, x_{n+1}) = \max\bigl(0, h_n(x_1, x_2, \ldots, x_n) + x_n\bigr). \tag{9}$$

Proof of lemma 4.1. The proof of is the same as in lemma 3.2. □

Proof of theorems 4.1 and 4.2. These follow from lemma 3.2 using a very similar argument from which theorem 2.1 follows from lemmas 3.2 and 3.1. For theorem 4.1 we apply the argument to the service times first (using the function f from lemma 3.2), and then to the negative of the inter-arrival times using the function g from lemma 4.1. For theorem 4.2 we apply the argument once using the function h from lemma 4.1. □

5. Summary

In this paper we study a multi-class queueing model where service times depend on the number of customers of the same type served in a row prior to the current customer. This can be considered in some sense a model for a "learning server" queue, or a model where switching or setup times for moving between customer classes can be spread over the course of several customer service times. Such models are useful in evaluating the decision to have either "pooled" or "dedicated" queues at a service station. Our main result here can be viewed as the idea that, under some conditions, using historical service time marginal data and simulating such queues as if they had iid service times leads to an *underestimate* of actual queueing delays. This gives an argument for the importance of more accurately modeling such phenomena.

Acknowledgements

We would like to thank an Associate Editor and two referees for their generous comments which have lead to a greatly improved version of our paper.

References

[1] N. Bäuerle, Monitonicity results for $MR/GI/1$ queues, J. Appl. Probab. 34 (1997) 514–524.
[2] N. Bäuerle and T. Rolski, A monotonicity result for the workload in Markov-modulated queues, J. Appl. Probab. 35 (1998) 741–747.
[3] J.H. Dshalalow and L. Tadj, A queueing system with random server capacity and multiple control, Queueing Systems 14(3/4) (1993) 369–384.
[4] S. Gilbert and Z.K. Weng, Incentive effects favor non-consolidating queues in a service system: The principal-agent perspective, Part 1 of 2, Manag. Sci. 44(12) (1998) 1662–1670.
[5] H. Li, Y. Zhu, P. Yang and S. Madhavapeddy, On $M/M/1$ queues with a smart machine, Queueing Systems 24(1/4) (1996) 23–36.

[6] A. Mandelbaum and M. Reiman, On pooling in queuing networks, Manag. Sci. 44(7) (1998).
[7] L. Meester and G. Shanthikumar, Regularity of stochastic processes: A theory based on directional convexity, Probab. Engrg. Inform. Sci. 7 (1993) 343–360.
[8] G. Pestalozzi, A queue with Markov-dependent service times, J. Appl. Probab. 5 (1968) 461–466.
[9] S.M. Ross, *Introduction to Probability Models*, 5th ed. (Academic Press, San Diego, CA, 1993).
[10] S.M. Ross, *Stochastic Processes* (Wiley, New York, 1996).
[11] S.M. Ross, J.G. Shanthikumar and X. Zhang, Some pitfalls of black box queue inference: The case of state-dependent server queues, Probab. Engrg. Inform. Sci. 7 (1993) 149–157.
[12] J.G. Shanthikumar, On a single-server queue with state-dependent service, Naval Res. Logist. Quart. 26(2) (1979) 305–309.
[13] T. Takine and T. Hasegawa, The workload in the $MAP/G/1$ queue with state-dependent services: Its application to a queue with preemptive resume priority, Comm. Statist. Stochastic Models 10(1) (1994) 183–204.
[14] D. Topkis, *Supermodularity and Complementarity* (Princeton Univ. Press, New York, 1998).