Optimal Control of a Queue With High-Low Delay
Announcements: The Significance of the Queue

Refael Hassin
Alexandra Koshman

Department of Statistics and Operations Research, Tel Aviv University
Tel Aviv 69978, Israel

May 22, 2015

Abstract

This article deals with strategic control of price and information in a single-server model. It considers an M/M/1 system with identical customers. There is a single cut-off number, and the level of congestion is said to be low (high) if the queue length is less than (at least) this value. The firm can dynamically change the admission fee according to congestion level. Arriving customers cannot observe queue length, but are informed of the current level of congestion and the admission fee. The article deals with finding the profit-maximizing admission fee using analytical and numerical methods. We observe that such a pricing regime can be used to achieve profit equal to the maximum social welfare in this model and that the proportion of the increase relative to the single price unobservable queue is unbounded. We observe that the profit-maximizing threshold is usually quite small and therefore raise a question as to whether there is a significant difference in profit when customers only join the system when the server is idle. We also investigate this question considering the classical observable model.

1 Introduction

The seminal work on strategic queueing behavior in queues by Naor (1969) assumes an observable M/M/1 system with homogeneous customers arriving at rate $\lambda$, service rate $\mu$, service value $R$ and waiting costs $C$ per unit time. Customers join only when the queue length they observe upon arrival is below a threshold (see chapter 3 of Hassin and Haviv (2003)). Naor shows this behavior not to be socially optimal and that a single price is sufficient to induce socially optimal behavior. Naor also considers a profit-maximizing queue manager imposing a static, queue-length independent, price. Since this system, in general, does not fully extract customer surplus, the achievable profit is lower than the maximum social welfare the system can generate. This can be contrasted with dynamic pricing, as we discuss below.

This contributions of this paper are presented in three parts. In the first two parts we consider a restricted type of dynamic pricing, which we denote high-low pricing. For a threshold $N$, the
admission fee is \( p_L \) if the number of customers in the system, \( n \), is less than \( N \), or is \( p_H \) otherwise. New customers cannot observe the exact queue length, but are informed whether or not the number of customers in the system satisfies \( n < N \) or \( n \geq N \). Note that the special case \( N = 0 \) leads to the unobservable version of Naor’s model, as analyzed by Edelson and Hildebrand (1975), and is equivalent to the other extreme \( N \to \infty \).

In the first part (section 3) we present price mechanisms that guarantee the server a profit equal to the maximum achievable social welfare, which clearly is an upper bound on profits. Three of these mechanisms are known and the fourth follows from our high-low model. We also compare these methods and emphasize the advantages of our new mechanism for the server and customers.

In the second part (section 4) we solve the high-low model when threshold \( N \) is not a decision variable but rather exogenously given. We observe that in most cases a small value, such as \( N = 1 \) or \( N = 2 \), is optimal or close to optimal, and therefore we investigate the advantage of maintaining long queues. These results naturally raise similar questions with regard to the classical observable queue investigated in Naor’s paper. We deal with these questions in the third part of our paper (section 5).

Our high-low model is applicable especially for virtual queues that are not observable to customers. But it also can be applied to some physical queues. For example, a common situation where the threshold arises exogenously is when customers can see the end of the queue only when queue length exceeds \( N \), because the view of the first \( N - 1 \) queue positions is blocked. Such examples are common in amusement parks where the head of the line forms inside a building or waiting hall, but once this area is filled the queue extends outside the building and may be observed before joining. The firm can control this information by monitoring the size of the internal waiting space or simply by blocking queue visibility by other means.

2 Literature review

High-low type queueing models have been the subject of various studies.

Several papers consider high-low delay announcements. Allon, Bassamboo and Gurvich (2011) deal with a queueing model in which the firm chooses a threshold and announces whether or not queue length is below it. Customers cannot verify this information, but they rationally use it in their decision to join or not to join the queue. In contrast, the information provided in our article is always assumed reliable. Altman and Jimenez (2004) deal with a queueing model in which customers are informed whether the queue is longer or shorter than a threshold. However, the admission fee is not changed for each case. Dobson and Pinker (2006) deal with a queueing model in which customers have heterogeneous waiting cost rates. When the system is below threshold, the queue is observable. Otherwise, customers are only informed that the system is congested. The admission fee always stays the same. Hall, Kopalle, and Pyke (2009) consider a single server committed to supplying service to its core customers within a given expected waiting time, and for a fixed price. The server uses excess capacity and admit occasional fill-in customers. Demand of fill-in customers depends on the charged price. The authors compare three options: (i) a constant price, (ii) a constant price up to a threshold and blocking fill-in customers above the threshold, (iii) state-dependent dynamic pricing. In contrast to our results, in their model option (iii) yields higher revenues than option (ii). Le Ny and Tuffin (2007) consider, among other models, a queueing model where a larger charge is imposed when occupancy is above threshold. The difference between our model and theirs is that in our case the admission fee is smaller when the queue is long. Moreover,
the customers in their article always have positive net utility even if they have to wait for a long
time. Maoui, Ayhan and Foley (2009) study a queue with a fixed price for queue lengths below
a threshold and an infinite price above it. In our model this means that \( p_H = \infty \). They assume
there are no customer waiting costs but there are server holding costs instead, and they allow
heterogeneous service valuations as expressed by the demand function.

Mutlu, Alanyali, Starobinski, and Turhan (2012) consider threshold pricing policies, where sec-
ondary customers are blocked if the number of busy servers is above a threshold.

Other models assume that service rate changes when queue length exceeds a threshold. For
example, Dimitrakopoulos and Burnetas (2011) discuss an unobservable model in which the service
rate is kept at a low value when the number of customers in the system is \( \leq T \), and turns to a
high value when the system congestion is above \( T \). They show that the equilibrium strategy is
not always unique, and derive an upper bound on the number of possible equilibria. Li and Jiang
(2013) describe a model in which service capacity is increased if the number of customers is above
a threshold. They deal with the optimal additional service capacity and the base capacity level
from the point of view of profit optimization. Perel and Yechiali (2010) consider a queueing model
with fast and slow phases of service rate. When system is in the slow phase customers become
impatient and leave the queue if the phase is not changed to the fast one within some period of
time. Chan, Yom-Tov, and Escobar use a fluid model to examine conditions where speedup of
service when queue length exceeds a threshold is beneficial even though faster service increases the
need for rework.

An article with a special case resembling our model was written by Economou and Kanta
(2008). They deal with an M/M/1 model in which the waiting space of the system is partitioned
into compartments of fixed size, and before entering the customer is told which compartment he
will enter or the position within the compartment he will have. We can describe our system in
terms of this article by saying we have two compartments, the first with fixed size \( N \), the second
with unlimited size, and an arriving customer is told the compartment he will enter. However,
admission fees do not differ for different queue lengths.

Shi, Shen, Wu and Cheng (2014) consider a Markovian single-server model with breakdowns.
The firm dynamically changes the price between exogenous \( p_1 \) and \( p_2 \), inducing exogenous demand
rates \( \lambda_1 \) and \( \lambda_2 \), respectively. Price is \( p_1 \) if inventory is below a threshold that depends on whether
the server is on or off, and is \( p_2 \) otherwise.

We also raise the question of the significance of maintaining a queue and keeping waiting spaces.
A similar problem is solved by Masarani and Gokturk (1987), who investigate the profit-maximizing
size of the waiting room in a queueing system. They consider an M/M/1/N queue where the server
incurs a cost \( C(N) \) and the buffer’s size \( N \) is a decision variable. The major difference is that the
customers in their case are not delay sensitive.

3 Profit maximization in the single-server queue

Naor (1969) assumes an M/M/1 system with homogeneous risk-neutral customers arriving at rate
\( \Lambda \), service rate \( \mu \), service value \( R \) and waiting costs \( C \) per unit time. The maximal value of social
welfare is attained if customers join the queue only when it is shorter than threshold \( n^* \). Denote by
\( S^* \) the social welfare under the threshold \( n^* \). Obviously, if customers enjoy nonnegative utilities, the
server’s profit cannot be greater than \( S^* \). A server can only attain this profit if arriving customers
join in accordance to the threshold $n^*$ and give all their welfare to the server.

We now describe three known pricing mechanisms that achieve these properties, and add a new method.

Chen and Frank (2001) observe that a profit equal to $S^*$ can be achieved by utilizing dynamic (state-dependent) pricing, i.e., charging $p(n) = R - CW_{<n^*}$ from a customer observing $n < n^*$ customers upon arrival, and a higher price otherwise. This pricing induces socially optimal behavior, the server receives all of the welfare generated by the system, and the net utility of each customer is equal to zero. Despite its advantages, such pricing is usually inconvenient to implement.

Another socially optimal admission model follows from an article by Hassin (1986) that describes an observable last-come first-served model with preemption (LCFS-PR) queueing model. All arriving customers join the queue and the last customer decides whether or not to abandon the queue. Since this customer remains last until served or abandoning the queue, he imposes no externalities and his decision is socially optimal. Hence he balks if and only if his position at the queue is $n^* + 1$, which is the socially optimal decision. All arriving customers have the same expected utility, which is independent of queue length. Therefore, the server can obtain all the social welfare by charging the maximal price they are ready to pay. The main problem in this model is that customers may renege from the queue and return to be the first in line.

A third possibility for achieving $S^*$ follows from work on priority sales done by Adiri and Yechiali (1975) and Alperstein (1988), who showed that a LCFS-PR regime can be obtained through adequate pricing of preemptive priorities while inducing threshold $n^*$ and leaving no customer surplus. An arriving customer buys the lowest priority with no current customer, and balks if all $n^*$ priorities have customers. To achieve this strategy, we set price for priority $i$ to be the expected utility of a customer buying this priority assuming all others behave according to the strategy. This behavior is an equilibrium under the stated strategy: Buying the lowest available priority (or balking when all priorities have at least one present customer) gives zero net expected utility, while any other action gives non-positive net expected utility. The result is a LCFS-PR regime, customer behavior is socially optimal, and server profit attains its upper bound. An advantage of this model is that, although the outcome is LCFS-PR among customers obtaining service, customers may not feel it is unfair because they choose the type of priority to purchase. Also, those paying eventually obtain service and those balking do not incur any costs, whereas under the LCFS-PR regime with a single price the waiting costs of reneging customers are not refunded. More details can be found in Erlichman and Hassin (2013).

We now explain how our model can be used to guarantee a profit equal to the socially optimal value $S^*$ when threshold $N$ is a decision variable. In the observable model with threshold $n^*$, the average customer utility for a customer arriving when $n < n^*$, is $R - CW_{<n^*}$, where $W_{<n^*}$ is the expected waiting time of a joining customer. Therefore, $S^* = \lambda L (n < n^*) (R - CW_{<n^*})$. In our model, the server can set $N = n^*$ and charge price $p_L = R - CW_{<n^*}$ (or a slightly lower price to guarantee that all customers arriving to state $L$ join, i.e., $\lambda_L = \Lambda$) and any sufficiently large $p_H$, so that no customers will join when $n \geq N$. Clearly, this guarantees the server’s profit rate to be $S^*$.

We now compare the four models from the points of view of the server and the customers. From the server’s point of view we compare: the number of different prices; the amount of information provided to customers; the costs of maintaining the system and changing state information; and the communication costs involved in supplying the information.
<table>
<thead>
<tr>
<th>Prices</th>
<th>Information</th>
<th>Costs</th>
<th>Communication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic pricing</td>
<td>$n^*$ admission prices</td>
<td>Observable queue (multiple states)</td>
<td>High switching costs</td>
</tr>
<tr>
<td>LCFS-PR</td>
<td>Single price</td>
<td>A single state</td>
<td>Elimination of renege-and-return</td>
</tr>
<tr>
<td>Priority sales</td>
<td>$n^*$ priority classes</td>
<td>Observable queue (multiple states)</td>
<td>Priority sales</td>
</tr>
<tr>
<td>High-low pricing (Our model)</td>
<td>Single price</td>
<td>Two states</td>
<td>Maintaining two states.</td>
</tr>
</tbody>
</table>

From the customer’s point of view, we compare: **fairness**; customer **surplus**; and the **variance** of the number of other customers served before completing service.

### Customers

<table>
<thead>
<tr>
<th>Prices</th>
<th>Surplus</th>
<th>Variance</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic pricing</td>
<td>Price inequality, but FCFS</td>
<td>Zero</td>
<td>Zero</td>
</tr>
<tr>
<td>LCFS-PR</td>
<td>Possible pay and wait with no service obtained</td>
<td>Zero</td>
<td>Highest</td>
</tr>
<tr>
<td>Priority sales</td>
<td>LCFS-PR but those who join are served. You ”make your choice”.</td>
<td>Zero</td>
<td>High</td>
</tr>
<tr>
<td>High-low pricing (Our model)</td>
<td>FCFS. Equal price at each of the two states, but some customers realize long queues.</td>
<td>Zero</td>
<td>Medium</td>
</tr>
</tbody>
</table>

Our proposed solution has advantages over the other solutions. On one hand, it has a single price and does not have high switching costs, which is convenient for the server. On the other hand, it is fair for the customers since it offers FCFS policy and has low variance.

## 4 The high-low system

Recall that we consider an M/M/1 queueing system with a potential arrival rate $\lambda$ of risk-neutral customers, service rate $\mu$, waiting cost rate $C$, and service value $R$. $N$ is an exogenous constant.

The queue manager sets admission fee $p_L$ if the queue length is smaller than $N$ (i.e., the state is $L$) and $p_H$ otherwise (the state is $H$). An arriving customer is informed whether the state is $L$ or $H$, and decides whether to join the queue or balk.
Given $p_L$, $p_H$, and $N$, denote the equilibrium arrival rate in the state $L$ ($H$) by $\lambda_L$ ($\lambda_H$), respectively. Then, the profit-rate function is:

$$\Pi(p_L, p_H, N) = p_L\lambda_L Pr(n < N) + p_H\lambda_H Pr(n \geq N).$$

(4.1)

Let $\pi_n$ be the probability of $n$ customers in the system. Then

$$\pi_n = \begin{cases} 
\left(\frac{\lambda_L}{\mu}\right)^n \pi_0, & n \leq N; \\
\left(\frac{\lambda_H}{\mu}\right)^N \left(\frac{\lambda_L}{\mu}\right)^{n-N} \pi_0, & n > N.
\end{cases}$$

(4.2)

We reduce the number of parameters by introducing normalized parameters:

$$\nu = \frac{R\mu}{C}, \rho_L = \frac{\lambda_L}{\mu}, \rho_H = \frac{\lambda_H}{\mu} \text{ and } \rho = \frac{\lambda}{\mu}.$$  

We assume that $\nu > 1$, otherwise no one will enter even if the server is idle. Then,

$$\pi_n = \begin{cases} 
\rho_L^n \pi_0, & n \leq N; \\
\rho_L^N \rho_H^{n-N} \pi_0, & n > N.
\end{cases}$$

(4.3)

Now we use $\sum_{n=0}^{\infty} \pi_n = 1$ to obtain

$$\pi_0 = \left(\frac{1 - \rho_L^N}{1 - \rho_L} + \frac{\rho_H^N}{1 - \rho_H}\right)^{-1}.$$  

(4.4)

Let $W_{<N}(\lambda_L, \lambda_H)$ be the expected waiting time of a customer joining the queue at state $L$, i.e., when there are less than $N$ customers in the system. Similarly, denote $W_{\geq N}(\lambda_L, \lambda_H)$. The net utility from joining the queue is $U_L = R - p_L - CW_{<N}$ in the $L$ state, or $U_H = R - p_H - CW_{\geq N}$ otherwise.

Without the loss of generality, we assume that in equilibrium customers are indifferent between joining and balking (this property will always hold under profit maximization). Thus,

$$\begin{cases} 
CW_{<N} = R - p_L; \\
CW_{\geq N} = R - p_H.
\end{cases}$$

(4.5)

Since clearly $W_{<N} < W_{\geq N}$, we conclude that $p_L > p_H$.

Denote $W_i = i+1$ the expected waiting time of a customer when there are $i$ customers already in the system. Then:
The formula for \( W \) resembles the mean of the exponential distribution with parameter \( \theta \) truncated at \( b \), which is defined as:

\[
f(y; \theta) = \begin{cases} 
\theta e^{-\theta y} (1 - e^{-\theta b})^{-1}, & 0 < y \leq b \\
0, & \text{otherwise} 
\end{cases}
\]
From Al-Athari (2008), the mean of this distribution is given by:

\[ \mu(\theta) = \frac{1}{\theta} - \frac{b}{e^{\frac{b}{\theta}} - 1}. \]

Since \( W_{M/M/1} \sim \exp(\mu - \lambda) \), we see that if we set \( \theta = \mu - \lambda_L \) and \( b = \frac{N}{\mu} \), we get a similar formula, but \( e^{b/\theta} \) replaces \( \frac{1}{e^{b/\theta} - 1} \). The difference can be explained by the fact that the truncated exponential distribution is conditioned by the fact that waiting time is less than \( b \), while \( W_{\geq N} \) is conditioned by the fact that expected waiting time is less than \( \frac{N}{\mu} \).

The formula for \( W_{\geq N} \) has a simple interpretation. A customer joining the queue at state \( H \) must wait the \( N \) customers that surely present in the system (including himself), plus the conditional expected waiting time when there are \( N \) customers in the system, i.e., the expected waiting time in an M/M/1 system with arrival rate \( \lambda_H \).

Inserting \( W_{< N} \) and \( W_{\geq N} \) in (4.5), we get the following set of equations:

\[
\begin{align*}
C_{1} \frac{N_{L}^{N+1} - (N+1)\rho_{L}^{N} + 1}{1 - \rho_{L}} &= R - p_L, \\
C_{2} \frac{(N+1) - N\rho_H}{1 - \rho_H} &= R - p_H. 
\end{align*}
\]  

(4.6)

From (4.6) we obtain:

\[
p_L = R - \frac{C N\rho_L^{N+1} - (N+1)\rho_L^{N} + 1}{\mu (1 - \rho_L^N) (1 - \rho_L)} = \frac{C}{\mu} \left( \nu - \frac{N\rho_L^{N+1} - (N+1)\rho_L^{N} + 1}{(1 - \rho_L^N) (1 - \rho_L)} \right),
\]

(4.7)

\[
p_H = R - \frac{C (N+1) - N\rho_H}{1 - \rho_H} = \frac{C}{\mu} \left( \nu - \frac{(N+1) - N\rho_H}{1 - \rho_H} \right).
\]

(4.8)

profit maximization for a given value \( N \) amounts to solving the following problem:

\[
\tilde{\Pi}(N) = \max_{p_L, p_H} \left( \Pr(n < N) p_L \lambda_L + \Pr(n \geq N) p_H \lambda_H \right)
= \max_{p_L, p_H} \left( \Pr(n < N) p_L \lambda_L + (1 - \Pr(n < N)) p_H \lambda_H \right).
\]

Substituting (4.3) and (4.4):

\[
\Pr(n < N) = \sum_{n=0}^{N-1} \pi_n = \sum_{n=0}^{N-1} \left( \frac{1}{1 - \rho_L} \frac{\rho_L^N}{1 - \rho_L} \right) = \frac{(1 - \rho_L) (1 - \rho_H)}{(1 - \rho_L^N) (1 - \rho_H) + \rho_L^N (1 - \rho_L)} = \frac{(1 - \rho_L^N) (1 - \rho_H)}{(1 - \rho_L) (1 - \rho_H) + \rho_L^N (1 - \rho_L)}.
\]

Denote the normalized profit rate function \( \max \Pi(N) := \frac{\tilde{\Pi}(N)}{C} \). Then, Using (4.7) and (4.8) we obtain that maximizing \( \tilde{\Pi}(N) \) is equivalent to solving:

8
\[ \Pi(N) := \max_{\rho_L, \rho_H} \left\{ \frac{(1-\rho_L)(1-\rho_H)}{(1-\rho_L)(1-\rho_H)+\rho_L(1-\rho_L)} \rho_L \left( \nu - \frac{N \rho_L + 1 - (N+1)\rho_H + 1}{(1-\rho_L)(1-\rho_L)} \right) + \right. \\
+ \left. \left(1 - \frac{(1-\rho_H)(1-\rho_H)}{(1-\rho_L)(1-\rho_H)+\rho_L(1-\rho_L)} \right) \rho_H \left( \nu - \frac{(N+1) - N \rho_H}{1-\rho_H} \right) \right\}. \tag{4.9} \]

**Unbounded demand**

We first consider the case \( \Lambda = \infty \). We have \( \nu > 1 \) as the input. We must have \( \rho_H < 1 \) in order to maintain stability of the system, and \( R \geq \frac{(N+1)C}{\mu} \) in order for \( \rho_H \) to be positive.

Clearly, when \( \Lambda = \infty \), \( n^* = 1 \). We now compare the optimal profit obtained when \( N = 1 \) to the unobservable case where \( N \to \infty \) (equivalently, \( N = 0 \)) and show that the gains associated with controlling the queue length can be significant.

With infinite arrival rate and \( N = 1 \) we have \( p_L = R - \frac{C}{\mu} \) and demand is continuously served at rate \( \mu \). Thus, \( \pi(1) = \frac{\mu}{\nu} \left( R - \frac{C}{\mu} \right) = \nu - 1 \).

For the unobservable model we use Edelson and Hildebrand’s (1975) formulas, \( \lambda_L = \mu - \sqrt{\frac{C \mu}{R}} = \mu \left(1 - \sqrt{\frac{1}{\nu}}\right) \), \( p_L = R - \frac{C}{\mu - \lambda_L} = \frac{C}{\mu} \left( \nu - \frac{1}{1 - \rho_L} \right) \), giving

\[ \Pi_{N \to \infty} = \left(1 - \sqrt{\frac{1}{\nu}}\right) \left( \nu - \frac{1}{1 - \rho_L} \right) = \left(1 - \sqrt{\frac{1}{\nu}}\right) \left( \nu - \frac{1}{1 - \left(1 - \sqrt{\frac{1}{\nu}}\right)} \right) = (\sqrt{\nu} - 1)^2. \]

Therefore, when \( N \to \infty \), \( \frac{\pi^{(1)}}{\Pi(N)} \to \frac{\nu - 1}{(\sqrt{\nu} - 1)^2} = \frac{\sqrt{\nu} + 1}{\sqrt{\nu} - 1} \), which grows to \( \infty \) when \( \nu \downarrow 1 \). On the other hand, when \( \nu \) is very large we have a ratio of approximately 1.

**Bounded demand**

We now solve the problem numerically assuming \( \Lambda < \infty \). This means we have one more parameter, \( \rho = \frac{\Lambda}{\mu} \), in addition to \( \nu \). The optimization problem is similar to the infinite \( \Lambda \) case, except that now \( \rho_L \leq \rho \) and \( \rho_H < \min(\rho, 1) \).

The graphs in Figure (1) examine the behavior of \( \Pi(N) \) for various values of \( N, \rho \) and \( \nu \). \( N \) is a discrete variable, and the relevant points of the graph are connected in order to ease understanding of the results.
Figure 1: $\Pi(N)$ for various values of $\rho$ and high values of $\nu$. The values of $\rho$ from bottom to top are 0.3, 0.5, 0.8, 2, 5, 10, 100.

When $\rho$ is small the line is almost flat because very few customers arrive to the system and almost always there are less than $N$ in the system. The profit-rate function is then equal to
\( \hat{\Pi}(N) = \lambda p_L = \lambda \left( R - \frac{C}{\rho} \right) \). Then \( \Pi(N) = \frac{MR}{\nu} - \frac{\lambda}{\rho} = \rho(\nu - 1) \).

We conclude from our numerical study that for each value of \( \rho \):

- For \( N < n^* \), \( \rho_L = \rho \) and \( 0 \leq \rho_H \leq \rho \).
- When \( N = n^* \), \( \rho_L = \rho \) and \( \rho_H = 0 \).
- For \( N > n^* \), \( \rho_L \leq \rho \) and \( \rho_H = 0 \).
- The profit-maximizing threshold is usually small and often equal to 1.

These conclusions are illustrated in Figures 2 and 3. Figure 3 illustrates the behavior of \( n^* \) as a function of \( \nu \). As we saw in Figure (1), when \( \rho \) is low, the profit function is flat, so the jumps in \( n^* \) are insignificant for such values. For slightly higher values of \( \rho \), we see that \( n^* \leq 3 \), and is equal to 1 for high values of \( \rho \).

![Figure 2: \( \Pi(N) \), \( \rho_L \) and \( \rho_H \) for \( 1 \leq N \leq 20 \). The upper graph is \( \Pi(N) \), the middle graph is \( \rho_L \) and the lower one is \( \rho_H \).](image-url)
Figure 3: $n^*$ as a function of $\nu$. 
\[ N = 1 \text{ with } \rho_H = 0 \text{ means creating an M/M/1/1 queue with no waiting room. A customer arriving when the server is busy balks. We now examine how much profit is gained from choosing } N = n^* > 1 \text{ and gaining } \Pi(n^*) = S^*, \text{ instead of letting customers join only when the server is idle.} \]

We observed that when \( N \geq n^* \), the profit-maximizing solution has \( \rho_H = 0 \).

Denote \( \Pi'(N) \) the maximal profit attained when arriving customers join the queue only if there are less than \( N \) customers in the system, when \( N \) is not necessarily optimal. We substitute \( \rho_H = 0 \) in (4.9). Then

\[
\Pi'(N) = \frac{1 - \rho_L^N}{1 - \rho_L^N} \nu \left( \frac{N \rho_L^{N+1} - (N + 1) \rho_L^N + 1}{(1 - \rho_L^N)(1 - \rho_L)} \right). \tag{4.10}
\]

For \( N = 1 \),

\[
\Pi'(1) = \frac{\rho_L}{1 + \rho_L} (\nu - 1). \tag{4.11}
\]

In Figure 4 we observe that the ratio \( \frac{\Pi(n^*)}{\Pi'(1)} = \frac{S^*}{\Pi'(1)} \) is at most 2, as proved in the following proposition:

**Proposition 1:** \( \frac{S^*}{\Pi'(1)} \leq 2 \).

(The proofs to claims in the section are given in the appendix.)

We observe that for practical values of \( \nu \) this ratio is much smaller. Moreover, Figure 5 demonstrates that \( \frac{S^*}{\Pi'(2)} \leq 1.5. \)

Figure 4: The gain associated with threshold 1 vs. \( n^* \).
When waiting positions are costly, we are interested in the marginal value of increasing the waiting room size from $N$ to $N + 1$. The ratio $\frac{\Pi'(N+1)}{\Pi'(N)}$ is given by

$$\frac{\Pi'(N+1)}{\Pi'(N)} = \frac{(1 - \rho_L^{N+1})[\nu(1 - \rho_L^{N+1})(1 - \rho_L) - ((N + 1)\rho_L^{N+2} - (N + 2)\rho_L^{N+1} + 1)]}{(1 - \rho_L^{N+2})[\nu(1 - \rho_L^{N})(1 - \rho_L) - (N\rho_L^{N+1} - (N + 1)^2\rho_L^{N+1} + 1)]}.$$

Figure 6 illustrates the value of this ratio for $1 \leq N \leq 10$. For each value of $N$, we consider all possible values of $\nu$ and $\rho$ and find the highest achievable value of $\frac{\Pi'(N+1)}{\Pi'(N)}$. 

Figure 5: The gain associated with 2 vs. $n^*$ when $n^* > 2$. 

!
We observe from Figure 6 that $\frac{\Pi'(N+1)}{\Pi'(N)} \leq \frac{4}{3}$. We now prove this observation.

**Lemma 1:** $\frac{\Pi'(N)}{\Pi'(1)}$ attains its highest value when $\nu = \infty$ and $\rho_L = 1$.

**Proposition 2:** $\frac{\Pi'(2)}{\Pi'(1)} \leq \frac{4}{3}$.  

We also see that $\frac{\Pi'(3)}{\Pi'(2)} \leq \frac{9}{8}$. Moreover, we see that for higher values of $N$, $\frac{\Pi'(N+1)}{\Pi'(N)} \approx 1$. Thus we conclude that if waiting space is costly, it often would not be optimal to have more than $N = 3$ spaces.

## 5 The observable model

We now compare the maximum achievable profit, $Z_o$, in the observable model with a single static price to the maximum value, $S^*$, that the system is able to generate, which can also be gained by any of the methods presented in Section 3.

We present in Figure 7 plots of the maximum normalized profit $Z'_o = Z_o/C$ and social welfare $S^*$. The functions coincide when both thresholds are 1 as in the case of infinite demand, and they are equal to $\nu - 1$.  

Figure 7: Maximum social welfare and profit in the observable model. The lower graph shows profit and the upper graph is $S^*$.

We present in Figure 8 the ratio $S^*/Z'_v$. 
Our numerical study reveals that the maximal ratio is approximately equal to 1.23 and is attained when $\nu = 5.54$ and $\rho = 0.47$. For large values of $\rho$, the ratio is equal to 1 since $n^* = n_m 1$, where $n_m$ is the profit-maximizing threshold.

We observed that optimal high-low delay announcements do not give a much better result than when setting $N = 1$. This observation raises a similar question about Naor’s observable model. We consider the observable model and compare the optimal profit obtained from Naor’s formula and the profit obtained from setting a threshold of $n = 1$ and price $p = R - \frac{C}{\mu}$. With a threshold 1 the normalized profit is $\frac{\rho}{1 + \rho}(\nu - 1)$.

The graphs in Figure 9 show the maximum profit vs. the profit obtained from using the threshold of 1.
Figure 9: Profit attained using threshold $n_m$ and profit attained when $n = 1$. The upper graph is obtained with $n_m$, the lower graph with $n = 1$ and $p = R - \frac{C}{\mu}$.

Figure 10 shows the ratio of the optimal profit with static pricing to the profit with threshold 1. For every given $\rho$ the ratio is a monotone increasing function of $\nu$. We also see that the highest ration of value 2, is obtained when $\rho = 1$ and $\nu \to \infty$. See also Figure 11.
Figure 10: The ratio of optimal profit and profit when $n = 1$.

Figure 11: The ratio of optimal profit to profit when $n = 1$, $\rho \to 1$. 

Max $Z_o/Z_o' = 1.9606$
We conclude that the best we can get with a single price yields at most twice the profit attained with \( n = 1 \). Of course, this result coincides with our previous observation that the maximal social welfare is at most twice the profit attained with \( n = 1 \). When \( \rho \) is close to 1 and \( \nu \) has a large value, it is worth maintaining a queue.

6 Conclusions

This paper consists of three parts. In the first part we consider a model in which arriving customers are informed whether the queue length is below a threshold \( n^* \) and pay \( p_L \) if they decide to join the queue. Such model attains a profit equal to the maximum welfare that can be generated, which is an upper bound on the achievable profit. We compare it to other mechanism that guarantee the same profit and conclude that it does not have some of their drawbacks.

In the second part we conduct analytical and numerical studies for a more general setting where a threshold \( N \) is exogenous, customers are informed whether the queue length is below (state \( L \)) or above it (state \( H \)), and the firm can only set two prices, one for each of the two states. We conclude that in order to attain optimality, customers shouldn’t enter when the state is \( H \) and all of the customers should enter when the state is \( L \). Moreover, the profit-maximizing \( N \) is in most cases quite small, and often \( N = 1 \). In some cases the gain from \( N = 2 \) is significant, but in most cases the gain associated with higher values is insignificant. In particular, we bound the gain from maintaining a queue and show that it is usually quite limited.

We compare our queueing system to the classical unobservable queueing systems. The optimal profit we attain is always higher, and the ratio of the profits is unbounded.

The fact that maintaining a queue usually doesn’t give a significant improvement, raised interest in checking the same about the classical observable model. We conduct analytical and numerical studies and conclude that maintaining a queue in the observable case also cannot guarantee a much higher profit. We have achieved similar results and conclusions with regard to social welfare.

Proofs:

Proof of Proposition 1:

We need to prove that for every \( N > 1 \):

\[
(1 + \rho_L)[\nu(1 - \rho_L^N)(1 - \rho_L) - (N\rho_L^N + (N + 1)\rho_L^N + 1)] \leq 2, \\
(1 - \rho_L^{N+1})(1 - \rho_L)(\nu - 1)
\]

or equivalently:

\[
(1+\rho_L)(\nu-1)(1-\rho_L^N)(1-\rho_L)+(1-\rho_L^N)(1-\rho_L)-(N\rho_L^{N+1}-(N+1)\rho_L^N+1)] \leq 2(\nu-1)(1-\rho_L^{N+1})(1-\rho_L).
\]

We separate this inequality into two parts, one containing \( \nu - 1 \) and the other containing all the rest, and prove the inequality for each of them. First we claim that:

\[
(1) \ (1 + \rho_L)(\nu - 1)(1 - \rho_L^N)(1 - \rho_L) \leq 2(\nu - 1)(1 - \rho_L^{N+1})(1 - \rho_L)
\]

This inequality is equivalent to:

\[
(1 - \rho_L)[1 - \rho_L^N + \rho_L - \rho_L^{N+1} - 2 + 2\rho_L^{N+1}] \leq 0, \text{ which holds for every } \rho_L \text{ because}
\]

\[
(1 - \rho_L)(1 + \rho_L - \rho_L^N + \rho_L^{N+1}) = -(1 - \rho_L)^2(1 + \rho_L^N) \leq 0.
\]

(2) The second inequality, \((1 + \rho_L)[(1 - \rho_L^N)(1 - \rho_L) - (N\rho_L^{N+1} - (N + 1)\rho_L^N + 1)] \leq 0\), is equivalent
to:
\((1 + \rho_L)(1 - \rho_L^N) + N\rho_L^N(1 - \rho_L) - (1 - \rho_L^N)\leq 0,\)
\((1 - \rho_L)(1 + \rho_L)[1 - \rho_L^N + N\rho_L^N - (1 + \rho_L + \rho_L^2 + \cdots + \rho_L^{N-1})] \leq 0,\)
\((1 - \rho_L)(1 + \rho_L)(\rho_L^N - \rho_L^2)(\rho_L^N - \rho_L^3) + \cdots + (\rho_L^N - \rho_L^N) \leq 0,\)
and \((1 - \rho_L^2)\sum_{i=1}^{N}(\rho_L^N - \rho_L^i) \leq 0,\) which is correct for any value of \(\rho_L.\)

**Proof of Lemma 1:**

We first prove that \(\frac{\Pi'(N)}{\Pi'(1)}\) is a monotone increasing in \(\nu.\) We rewrite \(\frac{\Pi'(N)}{\Pi'(1)}\) as
\[
\frac{\Pi'(N)}{\Pi'(1)} = \frac{\nu(1 + \rho_L)(1 - \rho_L^N)(1 - \rho_L) - (1 + \rho_L)(N\rho_L^{N+1} + (N + 1)\rho_L^N + 1)}{\nu(1 - \rho_L^{N+1})(1 - \rho_L) - (1 - \rho_L^{N+1})(1 - \rho_L)} = \frac{A\nu - B}{C\nu - D}.
\]
We will show that \(\frac{\partial}{\partial \nu} \left( \frac{\Pi'(N)}{\Pi'(1)} \right) = \frac{A(C\nu - D) - C(A\nu - B)}{(C\nu - D)^2} \geq 0.\)
The denominator is clearly positive, so we look at the numerator.
\(A(C\nu - D) - C(A\nu - B) \geq 0\) means that \(BC \geq AD.\) In our terms, this means
\((1 + \rho_L)(N\rho_L^{N+1} - (N + 1)\rho_L^N + 1)(1 - \rho_L^{N+1})(1 - \rho_L) \geq (1 + \rho_L)(1 - \rho_L^N)(1 - \rho_L)^2(1 - \rho_L)^2,\)
or equivalently,
\[(N\rho_L^{N+1} - (N + 1)\rho_L^N + 1)(1 - \rho_L^{N+1})(1 - \rho_L) \geq (1 - \rho_L^N)(1 - \rho_L)(1 - \rho_L^{N+1})(1 - \rho_L).
\]
We rewrite this inequality as
\[(1 - \rho_L^{N+1})(1 - \rho_L)(N\rho_L^{N+1} - (N + 1)\rho_L^N + 1) - (1 - \rho_L^N)(1 - \rho_L) \geq 0,
\]
which is equivalent to
\[(1 - \rho_L^{N+1})(1 - \rho_L)\rho_L[(N - 1)\rho_L^N - N\rho_L^{N-1} + 1] \geq 0.
\]
Clearly, \((1 - \rho_L^{N+1})(1 - \rho_L)\rho_L \geq 0,\) and we want to show that \((N - 1)\rho_L^N - N\rho_L^{N-1} + 1 \geq 0\) for any value of \(\rho_L.\)
The derivative is equal to \(N(N - 1)\rho_L^{N-1} - N(N - 1)\rho_L^{N-2} = N(N - 1)(\rho_L^{N-1} - \rho_L^{N-2});\) It is positive for \(\rho_L > 1\) and negative for \(\rho_L < 1,\) which means that \((N - 1)\rho_L^N - N\rho_L^{N-1} + 1\) is a unimodal function with a unique minimum at \(\rho_L = 1\) where the function is equal to 0.

In particular, we conclude that for any \(\rho_L,\) \(\frac{\Pi'(N)}{\Pi'(1)}\) attains its highest value when \(\nu = \infty.\)

We now prove that when \(\nu = \infty,\) \(\frac{\Pi'(N)}{\Pi'(1)}\) attains its maximal value when \(\rho_L = 1.\) When \(\nu = \infty,\) we get that:
\[
\frac{\Pi'(N)}{\Pi'(1)} = \frac{(1 + \rho_L)(1 - \rho_L^N)(1 - \rho_L)}{(1 - \rho_L^{N+1})(1 - \rho_L)} = \frac{(1 + \rho_L)(1 - \rho_L^N)}{(1 - \rho_L^{N+1})}
\]
and \(\frac{d}{d\rho_L} \frac{\Pi'(N)}{\Pi'(1)} = \frac{N\rho_L^{N+1} - \rho_L^{N+1} - N\rho_L^{N-1} + 1}{\rho_L^{N+2} - 2\rho_L^{N+1} + 1}.
\)

Therefore, \(\lim_{\rho_L \to 1} \frac{N\rho_L^{N+1} - \rho_L^{N+1} - N\rho_L^{N-1} + 1}{\rho_L^{N+2} - 2\rho_L^{N+1} + 1} = 0.\)

In order to prove that \(\rho_L = 1\) maximizes \(\frac{\Pi'(N)}{\Pi'(1)}\), we consider the second derivative.
\[
\frac{d^2}{d\rho_L^2} \frac{\Pi'(N)}{\Pi'(1)} = \frac{\rho_L^{N-2} - 2(N - 1)\rho_L^{N+1} - (N - 1)\rho_L^{N-1} + 1}{\rho_L^{N+2} - 2\rho_L^{N+1} + 1}.
\]
So \(\lim_{\rho_L \to 1} \frac{d^2}{d\rho_L^2} \frac{\Pi'(N)}{\Pi'(1)} = -\frac{N(N - 1)}{3(N + 1)} < 0,\)
meaning that \(\rho_L = 1\) maximizes \(\frac{\Pi'(N)}{\Pi'(1)}\) when \(\nu = \infty.\)

**Proof of Proposition 2:**
\[
\frac{\Pi'(2)}{\Pi'(1)} = \frac{(1 + \rho_L)\left[\nu(1 - \rho_L^2)(1 - \rho_L) - (2\rho_L^3 - 3\rho_L^2 + 1)\right]}{(1 - \rho_L^2)(1 - \rho_L)(\nu - 1)}
\]
\[
= \frac{(1 + \rho_L)\left[\nu(1 - \rho_L^2)(1 - \rho_L) - (1 - \rho_L)(1 - 2\rho_L^2 + \rho_L)\right]}{(1 - \rho_L^3)(1 - \rho_L)(\nu - 1)}
\]
\[
= \frac{(1 + \rho_L)\left[\nu(1 - \rho_L^2) - (1 + \rho_L - 2\rho_L^2)\right]}{(1 - \rho_L^3)(\nu - 1)}.
\]

We use Lemma 1 to find the maximal value of \(\frac{\Pi'(2)}{\Pi'(1)}\). When \(\nu \to \infty\), we get that:

\[
\lim_{\nu \to \infty} \frac{\Pi'(2)}{\Pi'(1)} = \frac{(1 + \rho_L)(1 - \rho_L^2)}{1 - \rho_L^3} = \frac{(1 + \rho_L)(1 - \rho_L)(1 + \rho_L)}{(1 - \rho_L)(1 + \rho_L + \rho_L^2)} = \frac{1 + 2\rho_L + \rho_L^2}{1 + \rho_L + \rho_L^2}.
\]

When \(\rho = 1\), we get that
\[\lim_{\nu \to \infty} \frac{\Pi'(2)}{\Pi'(1)} = \frac{4}{3}\].

References


