Tel-Aviv University

The Raymond and Beverly Sackler Faculty of
Exact Sciences
The Department of Statistics and Operations
Research

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

Thesis Submitted Towards the Degree of Master
of Science in Operations Research

Alexandra Koshman-Kaz

Supervisor:
Prof. Rafi Hassin

February 2015

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

Abstract

This article deals with strategic control of 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 rather than being informed about the congestion level. 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 Λ, service rate μ, 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. 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, and computes the optimal 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.

Our article deals with a restricted type of dynamic 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 (see Edelson and Hildebrand (1975) and chapter 3 of Hassin and Haviv (2003)), and is equivalent to the other extreme $N \to \infty$.

In this article we first present several 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 model. We also compare these methods and emphasize the advantages of our new mechanism for the server and customers. We also solve our 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. We therefore investigate the advantage of maintaining longer queues. We do the same with Naor's model and obtain, in both cases, theoretical and experimental bounds on the added profit associated with maintaining longer queues.

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.

Section 2 provides an overview of related articles. In section 3 we discuss price mechanisms such that server profit is equal to maximum social welfare.

In sections 4 and 5 we solve the profit maximization problem assuming N is exogenous. We question the significance of keeping a queue and compare optimal profit to the profit attained when $N = 1$.

In section 6 we compare our model to the classical observable model. Sections 7 and 8 consider the observable model and question the significance of a queue regarding profit maximization and social welfare. We summarize our findings in section 9.

2 Literature overview

Topics related to this article have been investigated and discussed in various studies. We refer to several articles that deal with high-low dynamic control of the queue.

Several papers consider delay announcements of the type we consider. For example, 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 an M/M/1 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 heterogenuous 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. Such queueing regime doesn't achieve maximum achievable social welfare. Hall, Kopalle, and Pyke (2009) consider a Markovian model with a single server committed to supplying service to its *core customers* within a given expected waiting time W_0, and for a fixed price. The arrival rate of these customers is fixed. The server can use excess capacity and admit occasional *fill-in customers* as long as it keeps its commitment to core customers. Fill-in customers are price-sensitive but not delay-sensitive and their demand is $\lambda_f(p_f)$, where p_f the charged price. The authors compare three options: (i) a constant price p_f independent of the state of the queue, (ii) a constant price up to a threshold and blocking fill-in customers above the threshold (both price and threshold are decision variables), (iii) general state-dependent dynamic pricing. Le Ny and Tuffin (2007) compare three M/M/1 queues. The first one having a fixed admission fee; the second has a larger charge being imposed when occupancy is above threshold; the third is a threshold-queue with hysteresis for switching between low and high prices. The difference between our model and the second queue type here 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.

Other models assume that service rate changes when queue length exceeds a threshold. For example, Dimitrakopoulos and Burnetas (2011) discuss an unobservable M/M/1 model in which service rate switches between low and high depending on system congestion. 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. They also examine the arrival rate which maximizes the overall customer welfare and the equilibrium. Li and Jiang (2013) describe a model in which additional service capacity is available. Their model also includes impatient customers. Adding service capacity depends on the number of customers in the system; service capacity is increased if the number of customers is above a given 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 an M/M/c 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.

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.

Several articles change the queue managing tactics based on the waiting time of an arriving customer. Kim and Hwang (2009) describe a case in which callers pay less for calls exceeding some given length. Specifically, if an arriving customer has to wait more than some waiting time threshold t, the waiting cost per time unit C is reduced starting from t, but no change in admission fee is considered. Maoui, Ayhan and Foley (2009) study an M/M/1 queue with a limited or unlimited buffer size. They set a fixed price for queue lengths below a threshold N and an infinite price above it. In our model this means that $p_H = \infty$. Their demand model is different from ours because 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.

Shi, Shen, Wu and Cheng (2014) consider a Markovian single server model with breakdowns and repairs, and price sensitive customers. The firm dynamically changes the price between exogenous $p_1 > p_2$, inducing exogenous demand rates $\lambda_1 < \lambda_2$, respectively. Also, the production rate is dynamically controlled. 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. The firm accumulates inventory and unmet demand is backordered. The production rate is maximal if the inventory is below a threshold and 0 otherwise, i.e., a base stock policy.

Wang and Barron (1994) consider an unobservable queueing system in which, due to communication costs, only a limited number of announcements is available. Arriving customers are being informed about the interval the queue length is located in, for example, between 10 and 20. Our system can be described in terms of this article by saying we have two intervals, $[1, N-1]$ and $[N, \infty]$.

In our article 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 they have a single admission fee which doesn't depend on the queue length. Moreover, the customers in their case are not delay sensitive.

3 Profit maximization in the single server queue

Consider an M/M/1 system with homogeneous customers arriving at rate Λ, service rate μ, service value R and waiting costs C per unit time. The maximal value of social welfare is attained its if customers join the queue only when it is shorter than threshold n^*. Denote by S^* the social welfare under the optimal threshold n^*. It is clear the server's profit cannot be greater than S^*. A server can 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 our 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 $p(n) = R - C \cdot \frac{n+1}{\mu}$ 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 LCFS-PR queueing model. All arriving customers join the queue and the last customer decides whether or not to abandon the queue. Since the last 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. Note that 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 regime, customer behavior is socially optimal, and server profit attains its upper bound. An advantage of this model is that, although the outcome is again LCFS 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 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 $R - C \cdot W_{<n^*}$, where $W_{<n^*}$ is the expected waiting time of a joining customer. Therefore, $S^* = \Lambda \cdot \Pr(n < n^*)(R - C \cdot W_{<n^*})$. In our model, the server can set $N = n^*$ and charge price $p_L = R - C \cdot W_{<n^*}$ (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 different **admission fees** the server has to announce.
- The **information** the server provides an arriving customer regrading his position in line.
- The **costs** of maintaining the queueing system and changing the state information.
- The **communication** costs. The amount of information about queue length the server provides an arriving customer.

Server

	Prices	Information	Costs	Communication
Dynamic Pricing	Multiple admission prices (n^* prices)	Observable queue (multiple states)	High switching costs	High
LCFS-PR	Single price	One state - no information supplied to customers.	Elimination of renege-and-return.	-
Priority Sales	Priority pricing menu (n^* prices)	Observable queue (multiple states)	Priority sales	High
High-Low Pricing (Our Model)	Single price	Two states	Maintaining two states.	Very low

From the customer's point of view, we compare:

- The **fairness** of the queueing system for an arriving customer.
- The **surplus** a customer is left with after receiving a service and leaving the system.
- The **variance** of the number of customers an arriving customer must wait to complete their service before he is served.

Customers

	Fairness	Surplus	Variance	Remarks
Dynamic Pricing	Price inequality, but FCFS	Zero	Zero	
LCFS-PR	Possible pay and wait with no service obtained	Zero	Highest	For social optimum it is sufficient of an arriving customer doesn't join the end if the queue, but for a single price it must be LCFS-PR.
Priority Sales	Those who join are eventually served. You "make your choice". LCFS as a result.	Zero	High	
High-Low Pricing (Our Model)	FCFS. All pay equal price at each of the two states. But some customers realize long queues.	Zero	Medium	For social optimum all that matters is that all join at L and are blocked at H.

We see that our model has advantages over the other systems. 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 Equilibrium and profit maximization solutions in a high-low system with a given threshold N

We consider an M/M/1 queueing system. Risk neutral customers arrive with potential arriving rate Λ. Service rate is μ. For each customer, the waiting cost is C per unit time, and service value is R. We assume that N is an externally given constant.

The admission fee is defined by:

$$p = \begin{cases} p_L, & n < N \\ p_H, & n \geq N. \end{cases}$$

The queue manager sets admission fees p_L and p_H. An arriving customer is informed whether the state is L (i.e., $n < N$) or H (i.e., $n \geq N$), and decides whether to join the queue or balk. The utility from balking is 0, and the utility from joining the queue is R minus C times the time spent in the system.

For given admission fees p_L,p_H and threshold N, the arrival process in the state L (H) is Poisson with equilibrium rates λ_L (λ_H), respectively. The profit rate function is:

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

In this section we investigate the equilibrium rates λ_L, λ_H and the profit maximizing admission fees p_L, p_H as a function of the threshold N.

For fixed $\lambda_L, \lambda_H, \mu$ and N, we have a birth and death process such that the birth rate is λ_L if $n < N$, or λ_H otherwise. If $n > 0$, the death rate is equal to μ.

Let π_n be the probability of n customers in the system. Then

$$\pi_n = \begin{cases} \left(\frac{\lambda_L}{\mu}\right)^n \cdot \pi_0, & n \leq N \\ \left(\frac{\lambda_L}{\mu}\right)^N \cdot \left(\frac{\lambda_H}{\mu}\right)^{n-N} \cdot \pi_0, & n > N. \end{cases} \tag{4.1}$$

We reduce the number of parameters by introducing the following 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. We get that:

$$\pi_n = \begin{cases} \rho_L^n \cdot \pi_0, & n \leq N \\ \rho_L^N \cdot \rho_H^{n-N} \cdot \pi_0, & n > N. \end{cases} \tag{4.2}$$

Now we use $\sum_{n=0}^{\infty} \pi_n = 1$ and get the explicit expression for π_0.

$$\sum_{n=0}^{\infty} \pi_n = \pi_0 + \rho_L \cdot \pi_0 + \ldots + \rho_L^N \cdot \pi_0 + \rho_L^N \cdot \rho_H \cdot \pi_0 + \rho_L^N \cdot \rho_H^2 \cdot \pi_0 + \ldots =$$
$$= \pi_0 \left(\left(1 + \rho_L + \rho_L^2 + \ldots + \rho_L^{N-1} \right) + \rho_L^N \left(1 + \rho_H + \rho_H^2 + \ldots \right) \right) =$$
$$= \pi_0 \left(\frac{1-\rho_L^N}{1-\rho_L} + \frac{\rho_L^N}{1-\rho_H} \right) = 1$$

Therefore,

$$\pi_0 = \left(\frac{1-\rho_L^N}{1-\rho_L} + \frac{\rho_L^N}{1-\rho_H} \right)^{-1}. \tag{4.3}$$

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 $U_L := R - p_L - C \cdot W_{<N}$ in the L state, or $U_H := R - p_H - C \cdot W_{\geq N}$ 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,

$$\begin{cases} C \cdot W_{<N} = R - p_L \\ C \cdot W_{\geq N} = R - p_H. \end{cases} \tag{4.4}$$

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

From the equations in (4.4), we extract the equilibrium values of λ_L and λ_H.

Denote $W_i = \frac{i+1}{\mu}$ the expected waiting time of an arriving customer when there are i customers already in the system. Then:

$$
\begin{aligned}
W_{<N} &= \sum_{i=0}^{N-1} \frac{\pi_i}{\sum_{j=0}^{N-1} \pi_j} \cdot W_i = \sum_{i=0}^{N-1} \frac{\pi_i}{\sum_{j=0}^{N-1} \pi_j} \cdot \frac{i+1}{\mu} = \frac{1}{\mu \sum_{j=0}^{N-1} \rho_L^j \pi_0} \sum_{i=0}^{N-1} \rho_L^i \pi_0 \cdot (i+1) \\
&= \frac{1}{\mu} \cdot \frac{1}{\frac{1-\rho_L^N}{1-\rho_L}} \cdot \sum_{i=0}^{N-1} \rho_L^i (i+1) = \frac{1}{\mu} \cdot \frac{1-\rho_L}{1-\rho_L^N} \cdot \rho_L^{-1} \cdot \sum_{i=0}^{N-1} \rho_L^{i+1} (i+1) \\
&= \frac{1}{\mu} \cdot \frac{1-\rho_L}{1-\rho_L^N} \cdot \rho_L^{-1} \cdot \rho_L \cdot \frac{1-(N+1)\rho_L^N + N \cdot \rho_L^{N+1}}{(1-\rho_L)^2} \\
&= \frac{1}{\mu} \cdot \frac{1}{1-\rho_L^N} \cdot \frac{N \cdot \rho_L^{N+1} - (N+1) \cdot \rho_L^N + 1}{1-\rho_L}
\end{aligned}
$$

$$
\begin{aligned}
W_{\geq N} &= \sum_{i=N}^{\infty} \frac{\pi_i}{\sum_{j=N}^{\infty} \pi_j} \cdot W_i = \sum_{i=N}^{\infty} \frac{\pi_i}{\sum_{j=N}^{\infty} \pi_j} \frac{i+1}{\mu} = \frac{1}{\mu \sum_{j=N}^{\infty} \rho_L^N \rho_H^{j-N} \pi_0} \cdot \sum_{i=N}^{\infty} \rho_L^N \rho_H^{i-N} \pi_0 (i+1) \\
&= \frac{1}{\mu \sum_{j=N}^{\infty} \rho_H^{j-N}} \sum_{i=N}^{\infty} \rho_H^{i-N} (i+1) = \frac{1}{\mu} \cdot \frac{1}{\frac{1}{1-\rho_H}} \cdot \sum_{i=0}^{\infty} \rho_H^{i+N+1} \cdot (i+N+1) \cdot \rho_H^{-(N+1)} \\
&= \frac{1}{\mu} \cdot (1-\rho_H) \cdot \rho_H^{-(N+1)} \sum_{i=1}^{\infty} \rho_H^{i+N} \cdot (i+N) = \frac{1}{\mu} (1-\rho_H) \rho_H^{-(N+1)} \cdot \rho_H^{N+1} \cdot \frac{(N+1) - N \cdot \rho_H}{(1-\rho_H)^2} \\
&= \frac{1}{\mu} \cdot \frac{(N+1) - N \cdot \rho_H}{1-\rho_H}
\end{aligned}
$$

All of the aforementioned calculations hold when $\rho_L \neq 1$ and $\rho_H < 1$. When $\rho_L = 1$, we get $W_{<N} = \frac{N+1}{2\mu}$.

There are three interesting special cases:

1. $W_{<1} = \frac{1}{\mu} \cdot \frac{1}{1-\rho_L} \cdot \frac{\rho_L^2 - 2\rho_L + 1}{1-\rho_L} = \frac{1}{\mu}$, which is the expected waiting time for a customer arriving to an empty system.

2. $W_{\geq 0} = \frac{1}{\mu} \cdot \frac{1}{1-\rho_H} = \frac{1}{\mu - \lambda_H}$, which is the expected waiting time in an unobservable queue with arriving rate λ_H.

3. $W_{<\infty} = \frac{1}{\mu} \cdot \frac{1}{1-\rho_L} = \frac{1}{\mu - \lambda_L}$, which is the expected waiting time in an unobservable queue with arriving rate λ_L.

$W_{<N}$ can be rewritten as $W_{<N} = \frac{1}{\mu - \lambda_L} - \frac{N}{\mu} \cdot \frac{1}{\rho_L^{-N} - 1}$. This formulation resembles the mean of the exponential distribution with parameter θ 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:

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

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 $e^{1-\rho_L}$ replaces $\frac{1}{\rho_L}$. 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 $W_{<N}$ is conditioned by the fact that *expected* waiting time is less than $\frac{N}{\mu}$.

The formula of $W_{\geq N}$ can be rewritten as $W_{\geq N} = \frac{1}{\mu} \cdot N + \frac{1}{\mu - \lambda_H}$, which means 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 $\geq N$ customers in the system, i.e., the expected waiting time in an M/M/1 system with arrival rate λ_H. These are exactly the conditions needed to describe a truncated exponential distribution.

Inserting $W_{<N}$ and $W_{\geq N}$ in (4.4), we get the following set of equations:

$$\begin{cases} C \cdot \frac{1}{\mu} \cdot \frac{1}{1-\rho_L^N} \cdot \frac{N \cdot \rho_L^{N+1} - (N+1) \cdot \rho_L^N + 1}{1-\rho_L} = R - p_L \\ C \cdot \frac{1}{\mu} \cdot \frac{(N+1) - N \cdot \rho_H}{1-\rho_H} = R - p_H. \end{cases} \tag{4.5}$$

Since we cannot express λ_L using p_L, we instead express p_L and p_H using λ_L and λ_H. From (4.5) we obtain:

$$p_L = R - \frac{C}{\mu} \cdot \frac{N \cdot \rho_L^{N+1} - (N+1) \cdot \rho_L^N + 1}{\left(1-\rho_L^N\right)(1-\rho_L)} = \frac{C}{\mu} \cdot \left(\nu - \frac{N \cdot \rho_L^{N+1} - (N+1) \cdot \rho_L^N + 1}{\left(1-\rho_L^N\right)(1-\rho_L)} \right), \tag{4.6}$$

$$p_H = R - \frac{C}{\mu} \cdot \frac{(N+1) - N\rho_H}{1-\rho_H} = \frac{C}{\mu} \cdot \left(\nu - \frac{(N+1) - N\rho_H}{1-\rho_H} \right). \tag{4.7}$$

The server faces the following optimization problem:

$$\tilde{\Pi}(p_L, p_H, N) := p_L \lambda_L \cdot Pr(n < N) + p_H \lambda_H \cdot Pr(n \geq N).$$

We will now examine the behavior of the results of this maximization problem for different values using numerical methods. We use p_L and p_H from (4.6) and (4.7) and obtain a function $\tilde{\Pi}(\lambda_L, \lambda_H, N)$, and carry out an optimization of $\tilde{\Pi}$ over λ_L and λ_H.

5 Solving and analyzing the profit maximization problem

In order to maximize the profit rate function for a given value N we solve the following optimization problem

$$\begin{aligned}\max\tilde{\Pi}(N) &= \max_{p_L,p_H}\left(\Pr(n<N)\cdot\ p_L\lambda_L+\Pr(n\geq N)\cdot p_H\lambda_H\right)\\ &= \max_{p_L,p_H}\left(\Pr(n<N)\cdot\ p_L\lambda_L+(1-\Pr(n<N))\cdot p_H\lambda_H\right).\end{aligned}$$

We substitute (4.6), and (4.7)

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

and obtain

$$\max\tilde{\Pi}(N)=\max_{\lambda_L,\ \lambda_H}\Big\{\frac{(1-\rho_L^N)(1-\rho_H)}{(1-\rho_L^N)(1-\rho_H)+\rho_L^N(1-\rho_L)}\ \cdot C\cdot\rho_L\cdot\left(\nu-\frac{N\cdot\rho_L^{N+1}-(N+1)\cdot\rho_L^N+1}{(1-\rho_L^N)(1-\rho_L)}\right)+$$

$$+\left(1-\frac{(1-\rho_L^N)(1-\rho_H)}{(1-\rho_L^N)(1-\rho_H)+\rho_L^N(1-\rho_L)}\right)\cdot C\cdot\rho_H\cdot\left(\nu-\frac{(N+1)-N\cdot\rho_H}{1-\rho_H}\right)\Big\}.$$

Denote the normalized profit rate function $\Pi(N):=\frac{\tilde{\Pi}(N)}{C}$. Then maximizing $\tilde{\Pi}(N)$ is equivalent to solving:

$$\Pi(N):=\max_{\rho_L,\ \rho_H}\Big\{\frac{(1-\rho_L^N)(1-\rho_H)}{(1-\rho_L^N)(1-\rho_H)+\rho_L^N(1-\rho_L)}\cdot\rho_L\cdot\left(\nu-\frac{N\cdot\rho_L^{N+1}-(N+1)\rho_L^N+1}{(1-\rho_L^N)(1-\rho_L)}\right)+$$

$$+\left(1-\frac{(1-\rho_L^N)(1-\rho_H)}{(1-\rho_L^N)(1-\rho_H)+\rho_L^N(1-\rho_L)}\right)\cdot\rho_H\cdot\left(\nu-\frac{(N+1)-N\cdot\rho_H}{1-\rho_H}\right)\Big\}\ . \tag{5.1}$$

This maximization problem attains the maximal social welfare S^*, which is the solution to Naor's formula for the expected social welfare in the observable model:

$$S_o=\lambda R\cdot\frac{1-\rho^n}{1-\rho^{n+1}}-C\cdot\Big[\frac{\rho}{1-\rho}-\frac{(n+1)\rho^{n+1}}{1-\rho^{n+1}}\Big],$$

where $n=\lfloor x\rfloor$, and x is the solution to $\frac{R\mu}{C}=x+\frac{\rho}{(1-\rho)^2}\cdot[x(1-\rho)-(1-\rho^x)]$.

We can translate the formula of S_o to the terms of ν and ρ and get:

$$S_o'=\frac{S_o}{C}=\rho\cdot\nu\cdot\frac{1-\rho^n}{1-\rho^{n+1}}-\Big[\frac{\rho}{1-\rho}-\frac{(n+1)\rho^{n+1}}{1-\rho^{n+1}}\Big]. \tag{5.2}$$

5.1 $\Lambda = \infty$

We solve the maximization problem (5.1) numerically for the case $\Lambda = \infty$. We have $\nu > 1$ as the input. Since $\Lambda = \infty$, ρ_L can have any value, while ρ_H must be < 1 in order to prevent explosion of the system. In order for ρ_H to be greater than 0, the condition $R \geq \frac{(N+1)C}{\mu}$ need to be met, i.e., $N \leq \nu - 1$.

When $\Lambda = \infty$, and N is a decision variable, it is clearly optimal to always have only one customer in the system. Substituting $N = 1$ in (4.6), we see optimal p_L equal to $(\nu - 1)$.

The graphs in Figure 1 examine the behavior of $\Pi(N)$ for $1 \leq N \leq 20$ and different values of ν.

When $N = 0$ (or equivalently $N \to \infty$), we have a simple unobservable $M/M/1$ queueing system. Since we are dealing with an infinite Λ, it is clear that $\Lambda \geq \mu - \sqrt{\frac{C\mu}{R}}$, so using this and the formulas presented by Edelson and Hildebrand (1975), we have $\lambda_L = \mu - \sqrt{\frac{C\mu}{R}}$, i.e., $\rho_L = 1 - \sqrt{\frac{1}{\nu}}$. It is also clear that $W_{<N} = \frac{1}{\mu - \lambda_L}$.

From the first equation in (4.4), we see that $p_L = R - \frac{C}{\mu - \lambda_L}$.

We see that $\tilde{\Pi}_{N\to\infty} = \tilde{\Pi}(0) = \lambda_L \cdot p_L = C \cdot \rho_L \left(\nu - \frac{1}{1-\rho_L}\right)$, i.e.,

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

Of course, our numerical results match this formula. For example, when $\nu = 4$, $\Pi_{N\to\infty} = 1$.

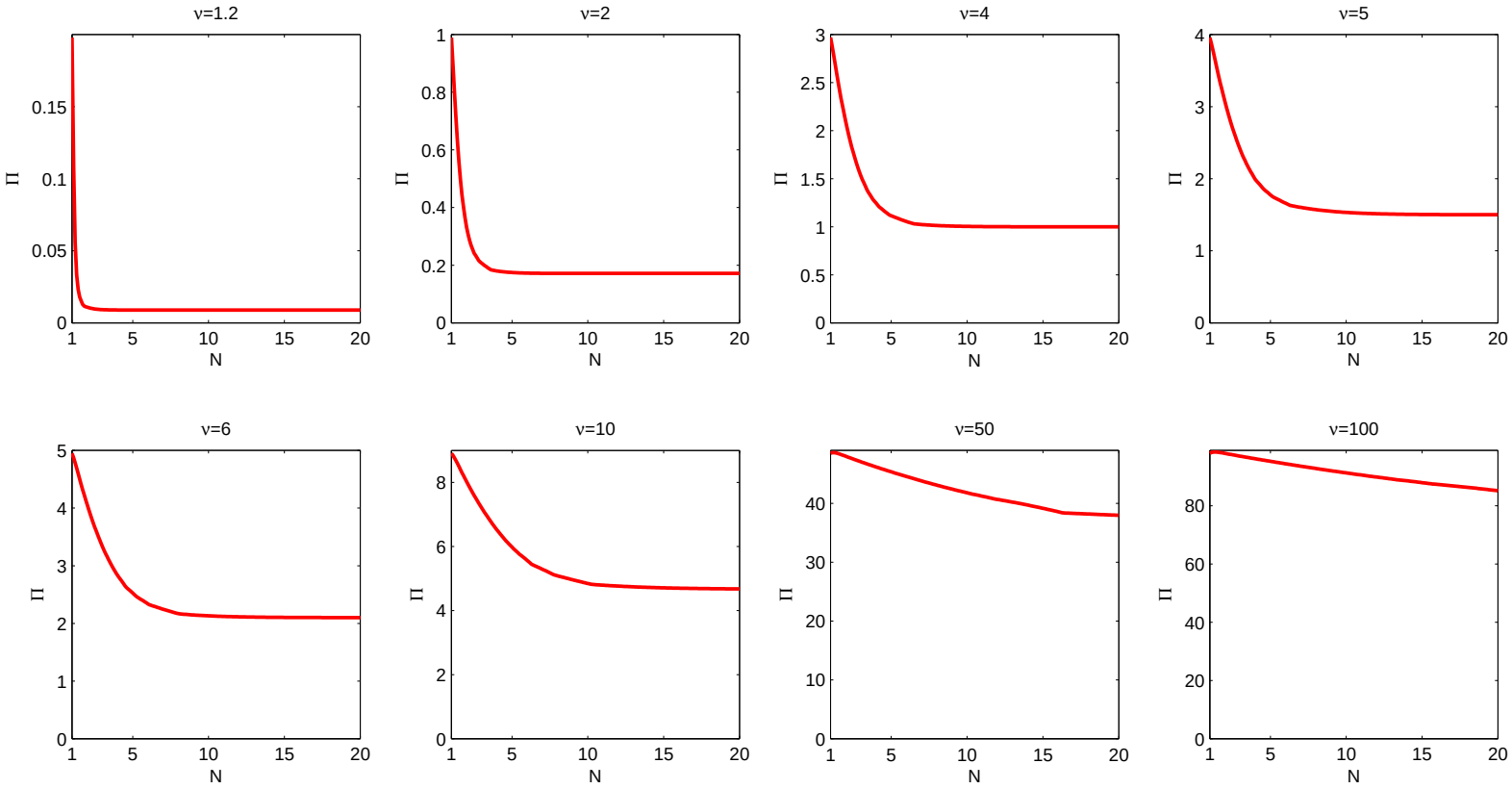

Figure 1: The behavior of Π for different values of ν and $1 \leq N \leq 20$.

5.1.1 Comparison of $N = 0$ and $N = 1$

Now we consider the ratio between values of $\Pi(N)$ when $N = 1$, i.e., the optimal value, and $N \to \infty$. When $N = 1$, $\rho_L = \infty$ and $\rho_H = 0$. From (5.1), we see that when $\rho_L \to \infty$,

$$\Pi(1) = \frac{1}{1-\rho_L} \cdot \rho_L \cdot (\nu - 1) = \frac{\rho_L}{1+\rho_L} \cdot (\nu - 1) \to \nu - 1$$

The ratio $\frac{\Pi_{N=1}}{\Pi_{N\to\infty}}$ is therefore equal to $\frac{\nu-1}{(\sqrt{\nu}-1)^2} = \frac{\sqrt{\nu}+1}{\sqrt{\nu}-1}$, which is not bounded and grows to ∞ when $\nu \downarrow 1$. On the other hand, when ν is very large we have a ratio of approximately 1.

5.2 $\Lambda < \infty$

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

The graphs in Figure 2 examine the behavior of $\Pi(N)$ for various values of N and ρ and for low values of ν. N is a discrete variable, and the relevant points of the graph are connected in order to ease understanding of the results.

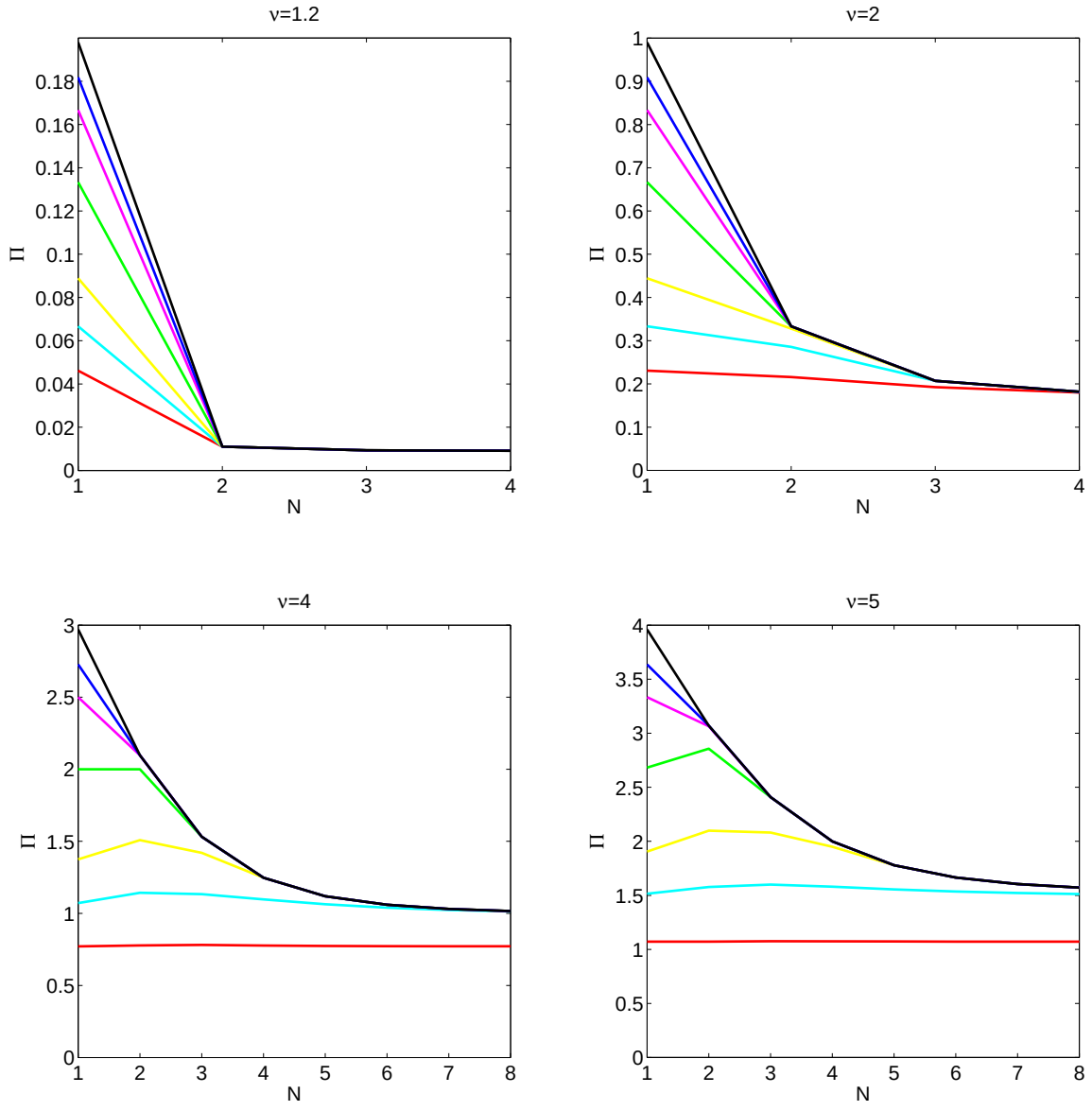

Figure 2: The behavior of $\Pi(N)$ for various values of ρ and low values of ν. The values of ρ from low to high are 0.3, 0.5, 0.8, 2, 5, 10, 100.

The graphs in Figure 3 examine the behavior of $\Pi(N)$ for various values of N and ρ and for high values of ν.

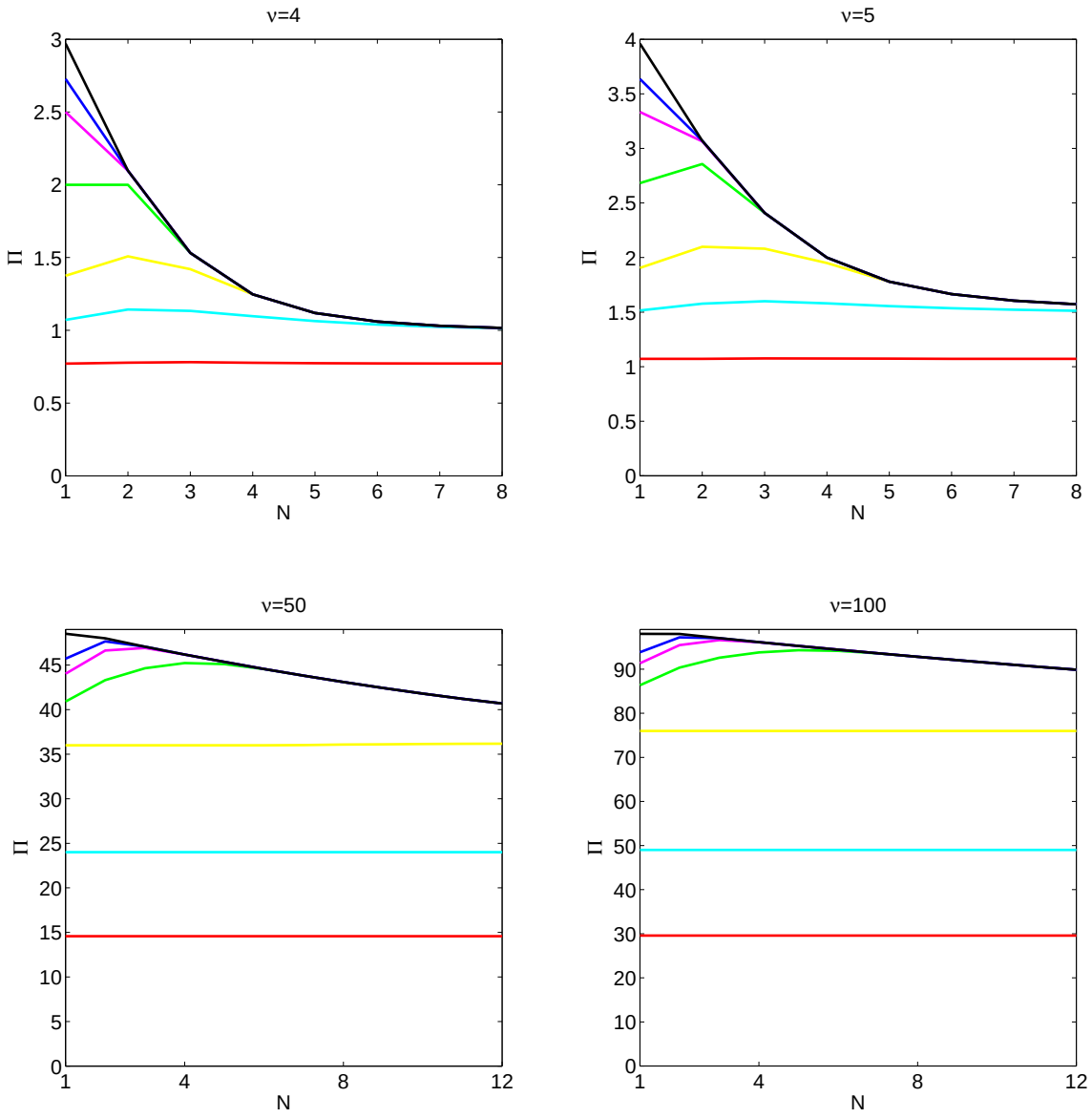

Figure 3: The behavior of $\Pi(N)$ for various values of ρ and high values of ν. The values of ρ from low to high are 0.3, 0.5, 0.8, 2, 5, 10, 100.

We see that for lower values of ρ the line is almost flat and as ρ grows, it gets closer to the results of $\Lambda = \infty$.

When $N \to \infty$, we have the unobservable model, so that $\Pi_{N\to\infty} = \rho\left(\nu - \frac{1}{1-\rho}\right)$.
Of course, our graphs match this formula. For example, let's take $\rho = 0.3$ and $\nu = 8$. Then $\Pi_{N\to\infty} = 2$, and this matches the result we see in Figures 2 and 3.

The flat lines that correspond to small values of ρ can be explained by the fact that when ρ is small, very few customers arrive to the queueing system, so almost always there are less than N in the system. An arriving customer always pays p_L, and thus we have a simple unobservable queue. The profit rate function is then equal to $\tilde{\Pi}(N) = \lambda \cdot p_L = \lambda\left(R - \frac{C}{\mu}\right)$. Then $\Pi(N) = \frac{\lambda R}{C} - \frac{\lambda}{\mu} = \rho(\nu - 1)$. For example, when $\nu = 100$ and $\rho = 0.3$, we have $\Pi(N) = 0.3 \cdot 100 = 30$. Recall that the maximum of $\Pi(N)$ is attained with $N = n^*$, when n^* is the threshold that maximizes the social welfare in Naor's model.

Moreover, we get from our numerical study that for each value of ρ:

- 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 $\rho_H = 0$.

For example, consider the case when $\nu = 5$ and $\rho = 0.6$. Then $n^* = 3$. However, if N is set to 1, $\rho_H = 0.4$. Likewise, $\rho_H = 0.1$ for $N = 2$. Figure 4 illustrates this example.

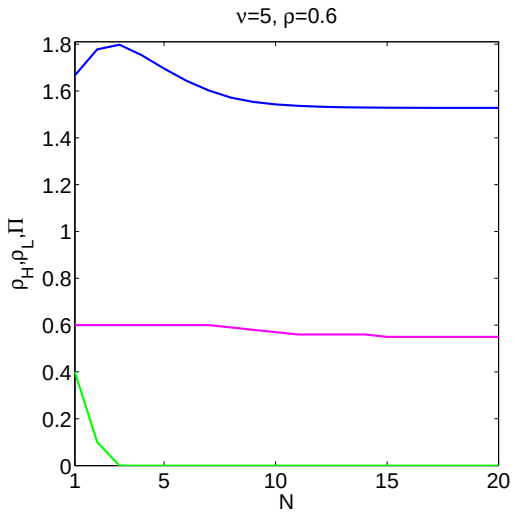

Figure 4: The behavior of $\Pi(N)$, ρ_L and ρ_H for $1 \leq N \leq 20$. The upper graph is $\Pi(N)$, the middle graph is ρ_L and the lower one is ρ_H.

Using these insights, let's assume $\rho_L = \rho$ and $\rho_H = 0$, and denote

$$g(N) := \Pi(\rho, 0, N) = \frac{1-\rho^N}{1-\rho^{N+1}} \cdot \rho \cdot \left(\nu - \frac{N \cdot \rho^{N+1} - (N+1)\rho^N + 1}{(1-\rho^N)(1-\rho)}\right),$$

where N is treated as a continuous variable.

From our numerical studies, $g(N)$ is a unimodal function so that the following conditions will be sufficient for optimality of N: (1)$g(N) < g(N+1)$ and (2) $g(N-1) > g(N)$. This means that:

$$(1) \quad \nu < \frac{\frac{N\rho^{N+1} - (N+1)\rho^N + 1}{(1-\rho^{N+1})(1-\rho)} - \frac{(N+1)\rho^{N+2} - (N+2)\rho^{N+1} + 1}{(1-\rho^{N+2})(1-\rho)}}{\frac{1-\rho^N}{1-\rho^{N+1}} - \frac{1-\rho^{N+1}}{1-\rho^{N+2}}} := h(N)$$

$$(2) \quad \nu > \frac{\frac{(N-1)\rho^N - N\rho^{N-1} + 1}{(1-\rho^N)(1-\rho)} - \frac{N\rho^{N+1} - (N+1)\rho^N + 1}{(1-\rho^{N+1})(1-\rho)}}{\frac{1-\rho^{N-1}}{1-\rho^N} - \frac{1-\rho^N}{1-\rho^{N+1}}} := h(N-1)$$

Let $\hat{\nu}$ be the solution(in most cases unique) to $\hat{\nu}$. We solve $\nu = h(\hat{\nu})$, find $\hat{\nu}$, and then $n^* = \lfloor \hat{\nu} \rfloor$.

6 The gain from maintaining a queue

We notice the profit maximizing threshold is usually very low and often equal to 1. Figure 5 illustrates the behavior of n^* as a function of ν. As we saw in Figures 2 and 3, when ρ is low, the profit function is flat, so the jumps in n^* are insignificant for such values. For slightly higher values of ρ, we see that $n^* \leq 3$, and is equal to 1 for high values of ρ.

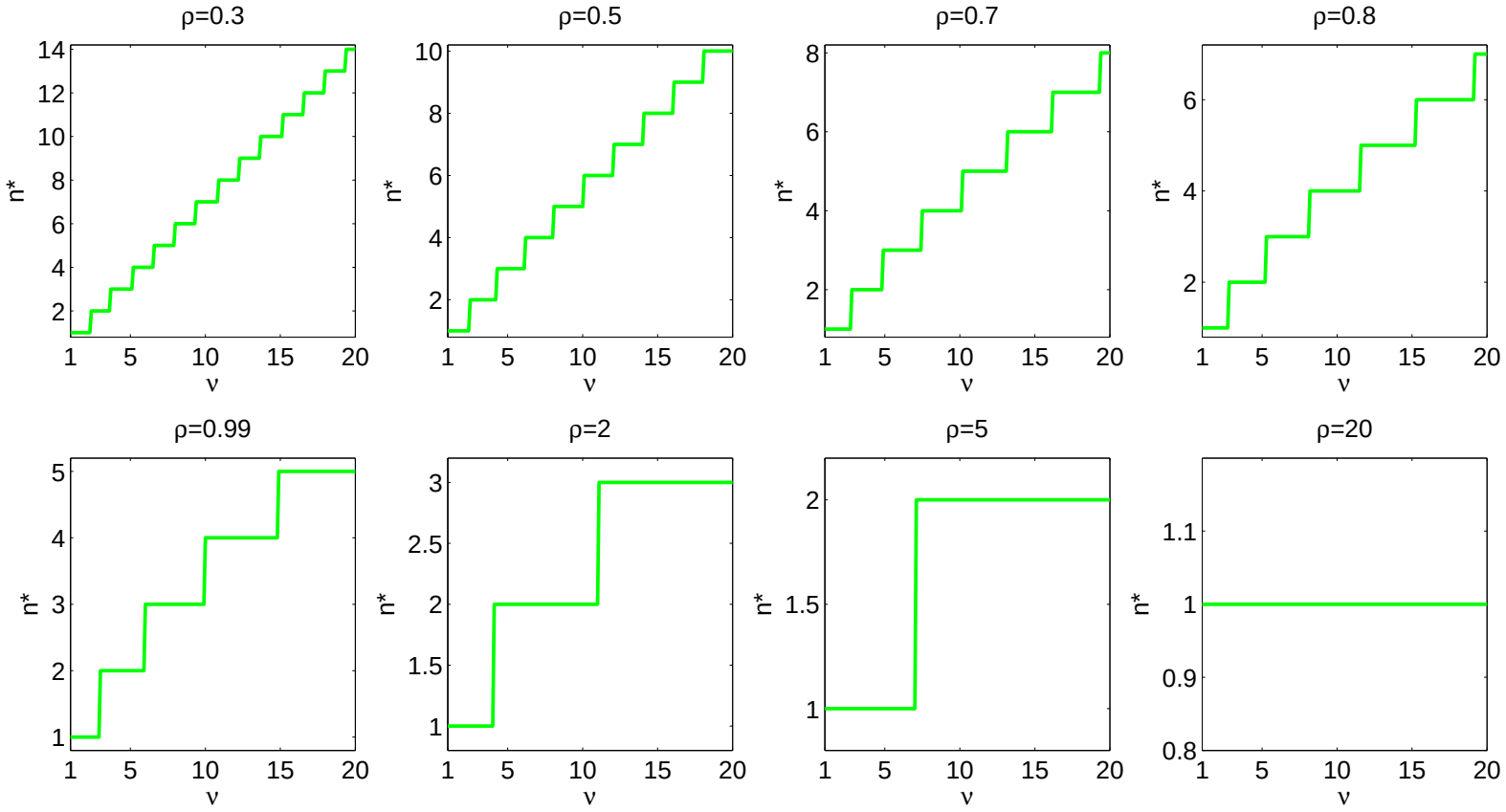

Figure 5: The behavior of n^* as a function of ν.

Recall that $N = 1$ with $\rho_H = 0$ means creating an M/M/1/1 queue with no waiting room. A customer arriving when the server is busy balks and never returns. We therefore check how much profit is gained from choosing $N = n^* > 1$ and gaining $\Pi(n^*) = S^*$, instead of letting customers join only when the server is idle.

We have already mentioned 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 (5.1). Then $\Pi'(N) = \frac{1-\rho_L^N}{1-\rho_L^{N+1}} \cdot \rho_L \cdot \left(\nu - \frac{N \cdot \rho_L^{N+1} - (N+1)\rho_L^N + 1}{(1-\rho_L^N)(1-\rho_L)}\right)$. For $N = 1$, $\Pi'(1) = \frac{\rho_L}{1+\rho_L} \cdot (\nu - 1)$.

Let's try to bound the ratio $\frac{\Pi(n^*)}{\Pi'(1)} = \frac{S^*}{\Pi'(1)}$ to see how much we can gain from choosing $N > 1$. In Figure 6 we see that this ratio is at most 2, and that for practical values of ν it is much smaller.

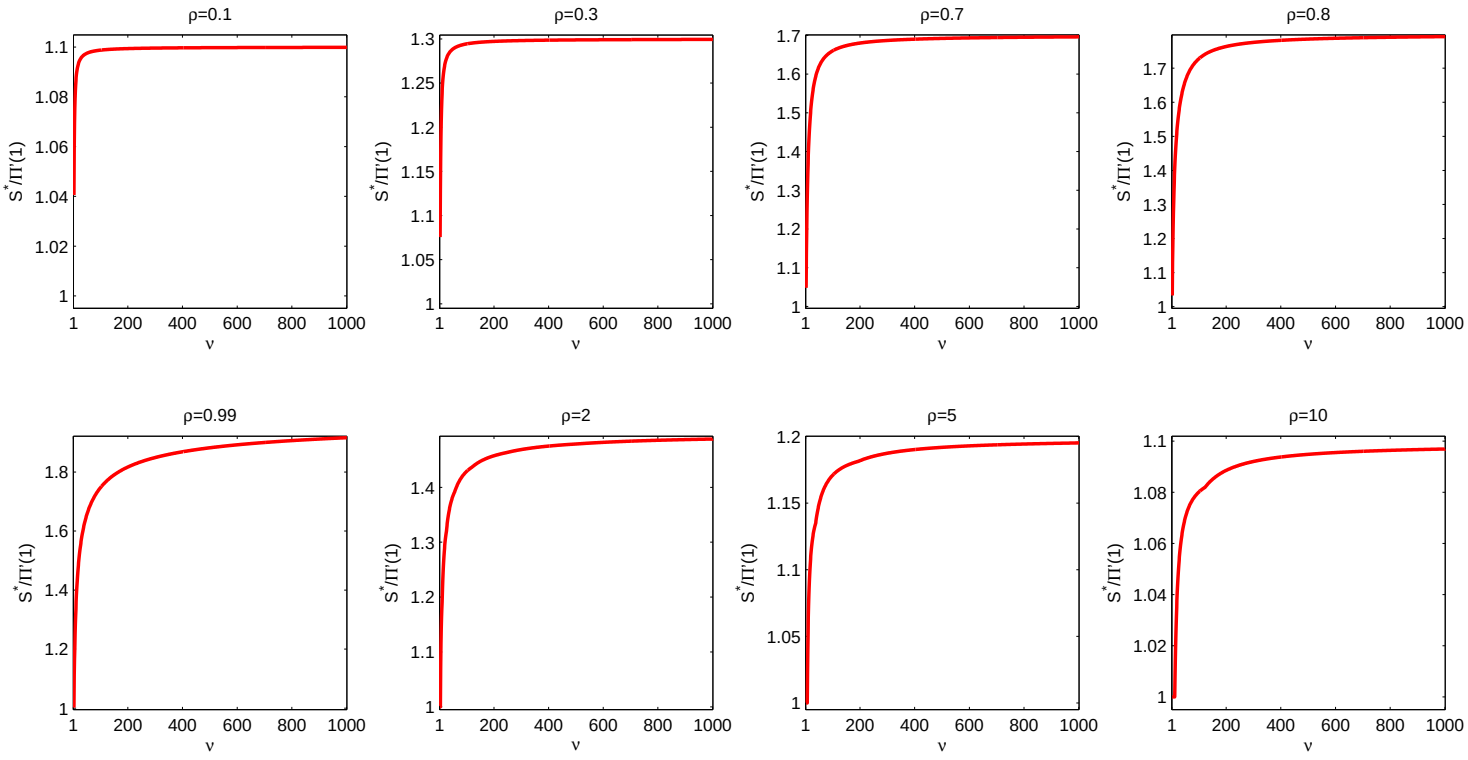

Figure 6: The gain associated with setting a threshold of 1 instead of n^*.

We now prove this numerical result.

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

Proof. We need to prove that for every $N > 1$:

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

or equivalently that:

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

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 :

$(1-\rho_L)\cdot[1-\rho_L^N+\rho_L-\rho_L^{N+1}-2+2\rho_L^{N+1}]\leq 0$, which means that

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

(2) The second inequality,

$(1+\rho_L)\cdot[(1-\rho_L^N)(1-\rho_L)-(N\cdot\rho_L^{N+1}-(N+1)\cdot\rho_L^N+1)]\leq 0,$

is equivalent to:

$(1+\rho_L)[(1-\rho_L^N)(1-\rho_L)+N\cdot\rho_L^N(1-\rho_L)-(1-\rho_L^N)]\leq 0,$

$(1-\rho_L)(1+\rho_L)[1-\rho_L^N+N\cdot\rho_L^N-(1+\rho_L+\rho_L^2+\ldots+\rho_L^{N-1})]\leq 0,$

$(1-\rho_L)(1+\rho_L)[(\rho_L^N-\rho_L)+(\rho_L^N-\rho_L^2)+\ldots+(\rho_L^N-\rho_L^N)]\leq 0,$

and

$(1-\rho_L^2)\cdot\sum_{i=1}^{N}(\rho_L^N-\rho_L^i)\leq 0$, which is correct for any value of ρ_L. □

Moreover, Figure 7 illustrates that $\frac{S^*}{\Pi'(2)} \leq 1.5$.

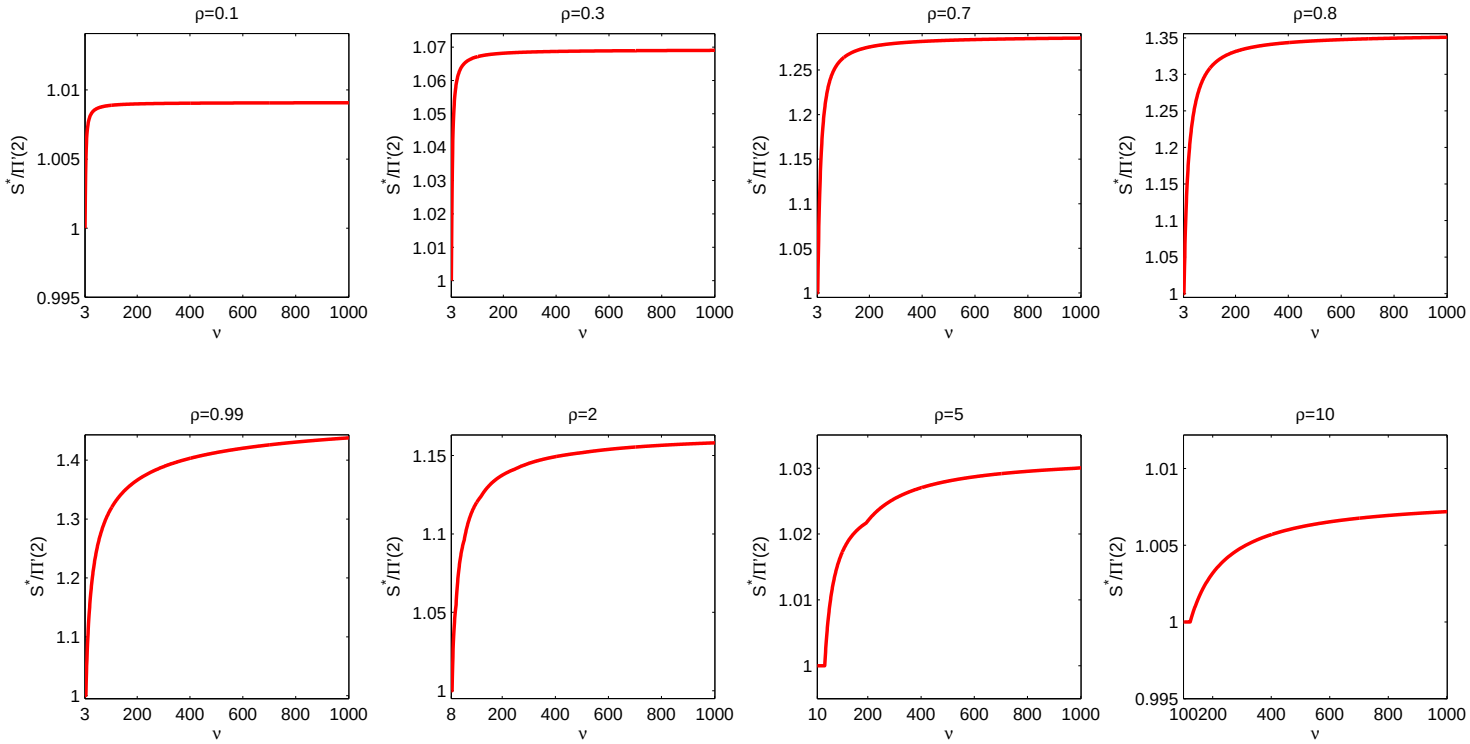

Figure 7: The gain associated with setting a threshold of 2 instead of n^*, when n^* is greater than 2.

We now find the maximal value of $N > 1$ that can give a significant improvement, if chosen, instead of $N = 1$.

The ratio $\frac{\Pi'(N+1)}{\Pi'(N)}$ is given by

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

In Figure 8 we illustrate the optimal value of this ratio for $1 \leq N \leq 15$. For each value of N, we consider all possible values of ν and ρ and find the highest achievable value of $\frac{\Pi'(N+1)}{\Pi'(N)}$.

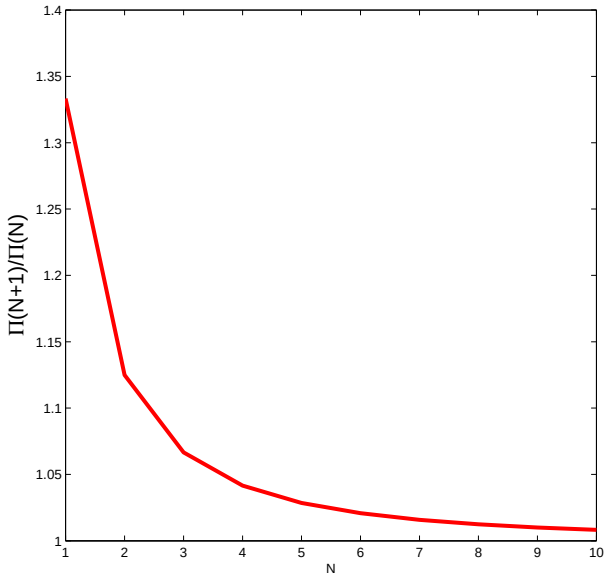

Figure 8: The behavior of the optimal $\frac{\Pi'(N+1)}{\Pi'(N)}$ as a function of N for $1 \leq N \leq 15$.

We conclude from Figure 8 that $\frac{\Pi'(2)}{\Pi'(1)} \leq \frac{4}{3}$. We now prove this numerical result.

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

Proof. We first prove that $\frac{\Pi'(N)}{\Pi'(1)}$ is a monotone increasing function of ν.

We can rewrite $\frac{\Pi'(N)}{\Pi'(1)}$ as :

$$\frac{\Pi'(N)}{\Pi'(1)} = \frac{\nu(1+\rho_L)(1-\rho_L^N)(1-\rho_L)-(1+\rho_L)(N\cdot\rho_L^{N+1}-(N+1)\rho_L^N+1)}{\nu(1-\rho_L^{N+1})(1-\rho_L)-(1-\rho_L^{N+1})(1-\rho_L)} := \frac{A\cdot\nu-B}{C\cdot\nu-D}.$$

In order for $\frac{\Pi'(N)}{\Pi'(1)}$ to be a monotone increasing function of ν, $\frac{\partial}{\partial\nu}\left(\frac{\Pi'(N)}{\Pi'(1)}\right)$ must be ≥ 0. It means 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 that

$$(1+\rho_L)(N\cdot\rho_L^{N+1}-(N+1)\rho_L^N+1)\cdot(1-\rho_L^{N+1})(1-\rho_L) \geq (1+\rho_L)(1-\rho_L^N)(1-\rho_L)\cdot(1-\rho_L^{N+1})(1-\rho_L),$$

or equivalently,

$$(N \cdot \rho_L^{N+1} - (N+1)\rho_L^N + 1) \cdot (1-\rho_L^{N+1})(1-\rho_L) \geq (1-\rho_L^N)(1-\rho_L) \cdot (1-\rho_L^{N+1})(1-\rho_L).$$

We can rewrite this inequality as

$$(1-\rho_L^{N+1})(1-\rho_L) \cdot [(N \cdot \rho_L^{N+1} - (N+1)\rho_L^N + 1) - (1-\rho_L^N)(1-\rho_L)] \geq 0,$$

which is equivalent to

$$(1-\rho_L^{N+1})(1-\rho_L) \cdot \rho_L[(N-1) \cdot \rho_L^N - N \cdot \rho_L^{N-1} + 1] \geq 0.$$

First of all, for any ρ_L, $(1-\rho_L^{N+1})(1-\rho_L)\cdot\rho_L \geq 0$. We want to show that also $(N-1)\cdot\rho_L^N - N\cdot\rho_L^{N-1}+1 \geq 0$ for any value of ρ_L.

The derivative is equal to

$$N(N-1)\rho_L^{N-1} - N(N-1)\rho_L^{N-2} = N(N-1)(\rho_L^{N-1} - \rho_L^{N-2}) \quad .$$

It is positive for $\rho_L > 1$ and negative for $\rho_L < 1$, which means that $(N-1) \cdot \rho_L^N - N \cdot \rho_L^{N-1} + 1$ is a unimodal function with a unique minimum at $\rho_L = 1$.

Thus we conclude that for any ρ_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:

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

The derivative is then equal to:

$$\frac{d}{d\rho_L}\frac{\Pi'(N)}{\Pi'(1)} = \frac{N\rho_L^{N+1} - \rho_L^{2N} - N\rho_L^{N-1}+1}{\rho_L^{2N+2} - 2\rho_L^{N+1}+1} \quad .$$

We then get that:

$$\lim_{\rho_L \to 1} \frac{N\rho_L^{N+1} - \rho_L^{2N} - N\rho_L^{N-1}+1}{\rho_L^{2N+2} - 2\rho_L^{N+1}+1} = 0 \ .$$

In order to prove that $\rho_L = 1$ maximizes $\frac{\Pi'(N)}{\Pi'(1)}$, we consider the second derivative.

$\frac{d^2}{d\rho_L^2}\frac{\Pi'(N)}{\Pi'(1)} = \frac{\rho_L^{N-2}(N^2(1-\rho_L^2)(\rho_L^{N+1}+1)-N(-5\rho_L^{N+1}+\rho_L^{N+3}+3\rho_L^2+1)+2\rho_L^2(\rho_L^{2N}-1))}{(\rho_L^{N+1}-1)^3}$.

Finally, we see that

$$\lim_{\rho_L\to 1}\frac{d^2}{d\rho_L^2}\frac{\Pi'(N)}{\Pi'(1)} = -\frac{N(N-1)}{3(N+1)} < 0\ ,$$

which means that $\rho_L = 1$ maximizes $\frac{\Pi'(N)}{\Pi'(1)}$ when $\nu = \infty$.

□

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

Proof. We can express $\frac{\Pi'(2)}{\Pi'(1)}$ as:

$$\begin{aligned}\frac{\Pi'(2)}{\Pi'(1)} &= \frac{(1+\rho_L)\cdot[\nu(1-\rho_L^2)(1-\rho_L)-(2\rho_L^3-3\rho_L^2+1)]}{(1-\rho_L^3)(1-\rho_L)(\nu-1)}\\ &= \frac{(1+\rho_L)\cdot[\nu(1-\rho_L^2)(1-\rho_L)-(1-\rho_L)(1-2\rho_L^2+\rho_L)]}{(1-\rho_L^3)(1-\rho_L)(\nu-1)}\\ &= \frac{(1+\rho_L)\cdot[\nu(1-\rho_L^2)-(1+\rho_L-2\rho_L^2)]}{(1-\rho_L^3)(\nu-1)}.\end{aligned}$$

We now 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 $\lim_{\nu\to\infty}\frac{\Pi'(2)}{\Pi'(1)} = \frac{4}{3}$.

□

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

7 Comparison to the observable model

This section considers the question of how much we can improve profit by using our method and attaining S^*, instead of using Naor's optimal solution.

Consider the observable model. Naor's profit maximization formula for this model is:

$$Z_o = \lambda \cdot \frac{1-\rho^{n_m}}{1-\rho^{n_m+1}} \cdot \left[R - \frac{n_m C}{\mu}\right],$$

where $n_m = \lfloor \nu_m \rfloor$, and ν_m is the solution to $\frac{R\mu}{C} = \nu_m + \frac{(1-\rho^{\nu_m-1})(1-\rho^{\nu_m+1})}{\rho^{\nu_m-1}(1-\rho)^2}$.

We can translate the formula of Z_o to the terms of ρ and ν and get:

$$Z'_o = \frac{Z_o}{C} = \rho \cdot \frac{1-\rho^{n_m}}{1-\rho^{n_m+1}} \cdot [\nu - n_m] \ . \tag{7.1}$$

Using this formula, we present in Figure 9 plots of profit as a function of ν.

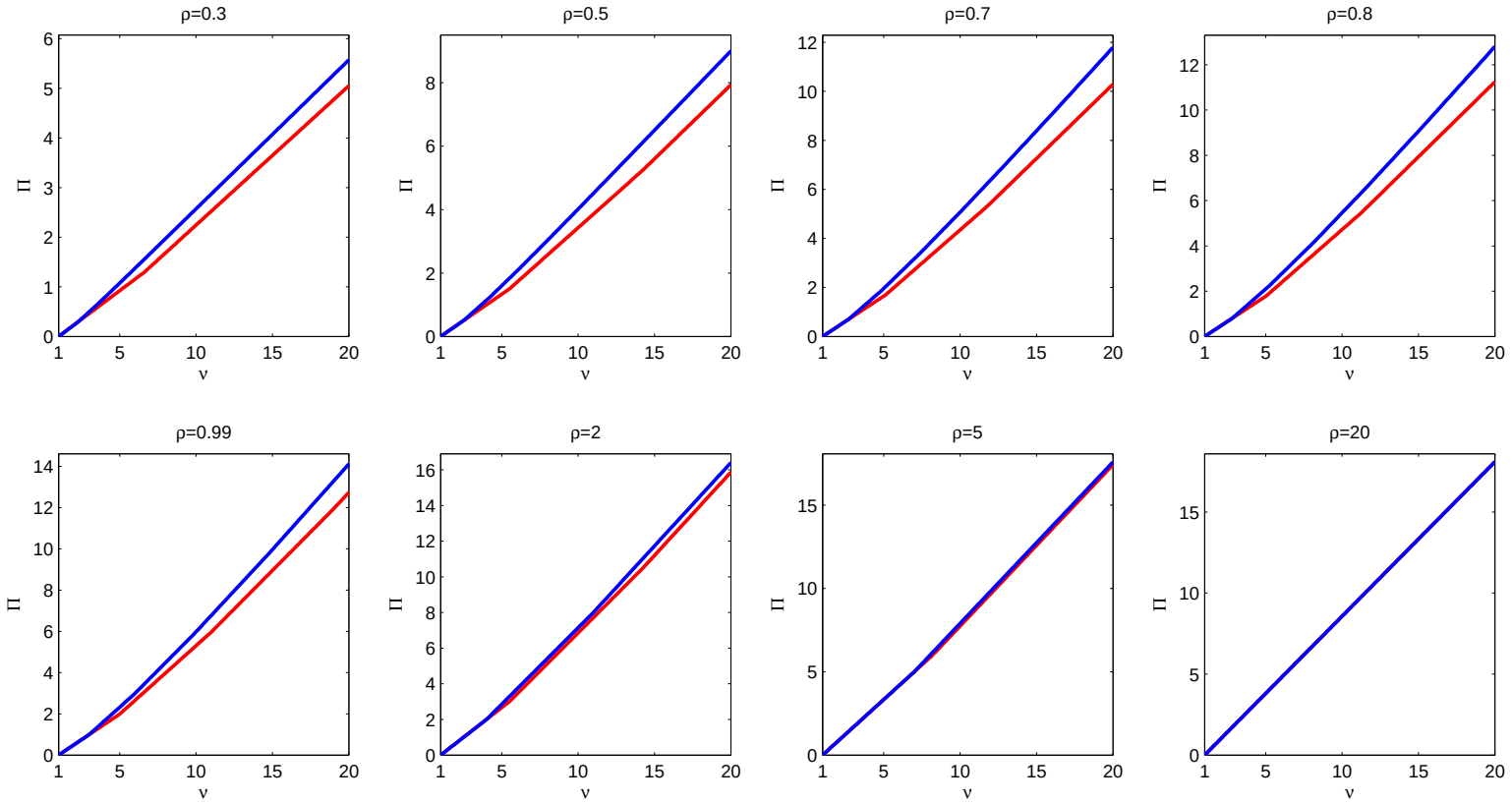

Figure 9: Comparison of the profit in the observable model and S^*. The lower graph is profit from the observable model, and the upper graph is S^*.

As expected, we attain higher values using our model. For example, when $\rho = 0.9$, and $\nu = 10$, we get almost 6 in our model, in comparison to 5 in the observable model.

We present in Figure 10 the ratio between S^* and profit attained using the observable model.

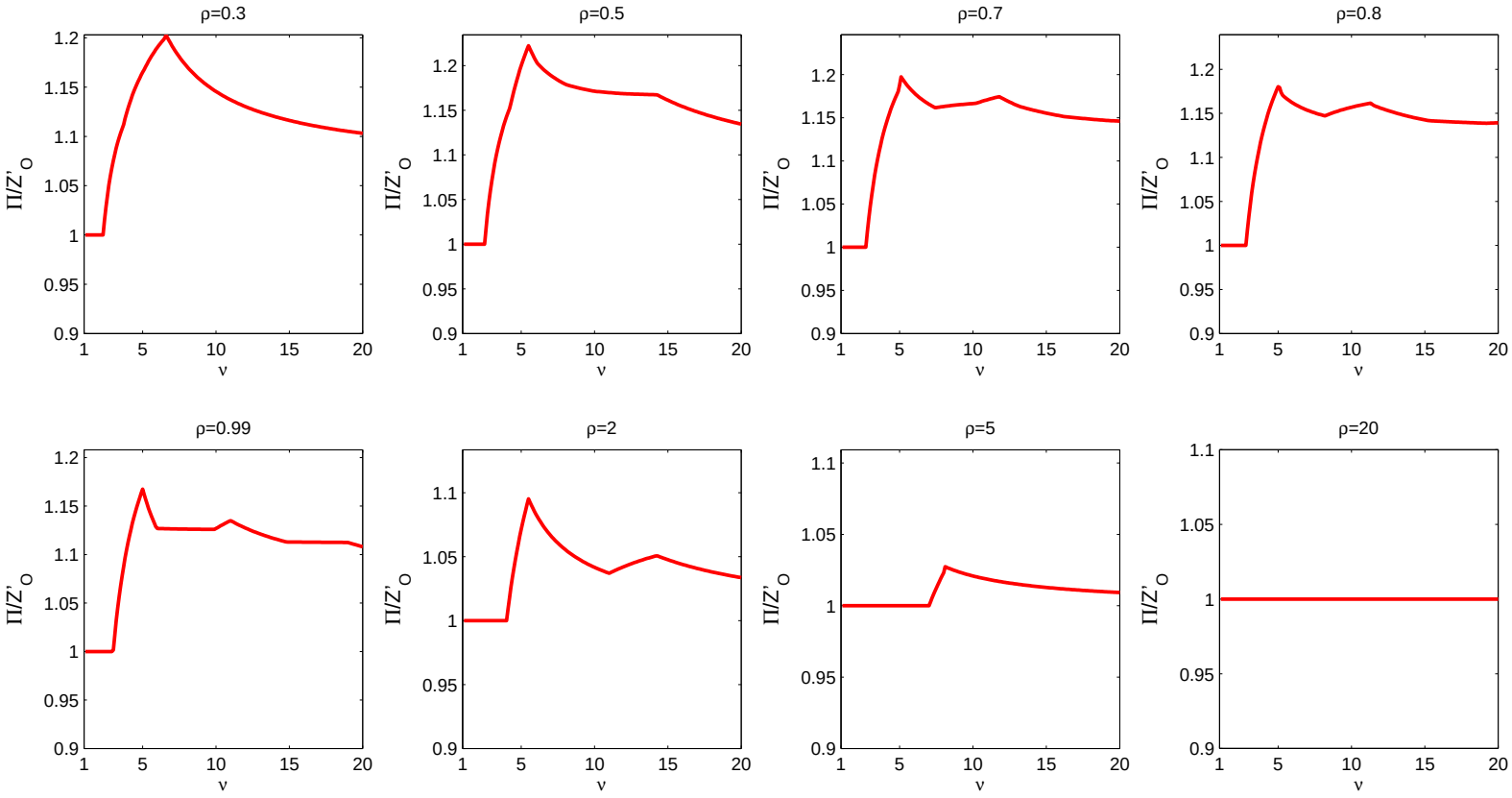

Figure 10: The ratio between our model and the observable model.

Our numerical studies reveal that the maximal ratio is equal to 1.2246 and is attained when $\nu = 5.54$ and $\rho = 0.47$. For large values of ρ, the ratio is equal to 1 since $n^* = 1$.

8 Comparison of the profit maximization attained by using the optimal n_m and 1 in the observable model

We observe that in our high-low delay announcements model we cannot obtain a much better result by setting N larger than 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 $p = R - \frac{C}{\mu}$ in the formula of the observable model's profit.

Using (7.1), we see that when $n_m = 1$,

$$Z_o' = \frac{Z_o}{C} = \frac{\rho}{1+\rho} \cdot (\nu - 1). \tag{8.1}$$

Consider the graphs in Figure 11, of profit obtained from using the optimal formula and profit obtained from using the threshold of 1.

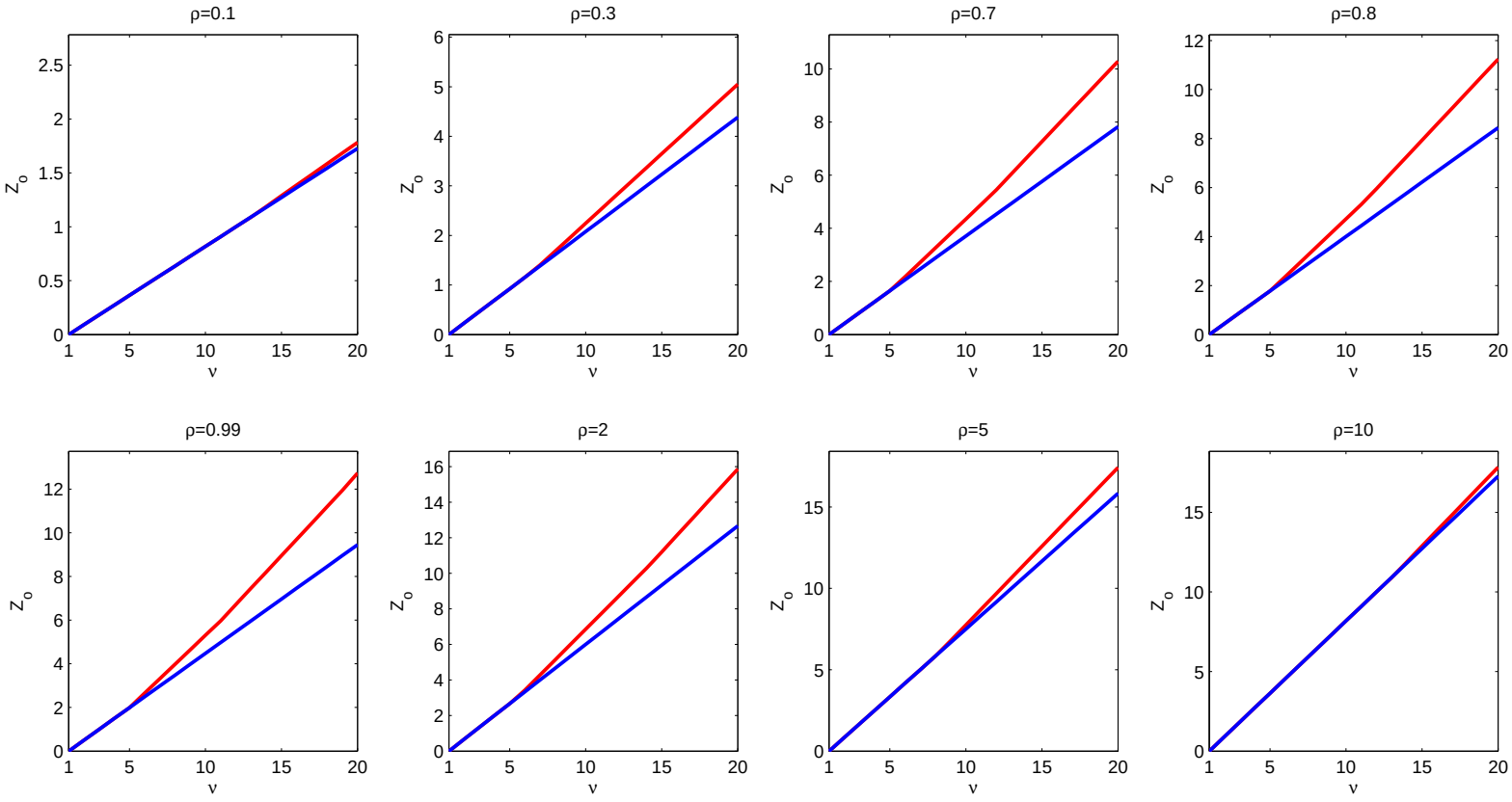

Figure 11: Comparison of profit attained using the threshold n_m and profit attained when $n = 1$ The upper graph is obtained with n_m, while the lower graph is obtained with $n = 1$ and $p = R - \frac{C}{\mu}$.

Figure 12 shows the graphs of the ratio between the profits.

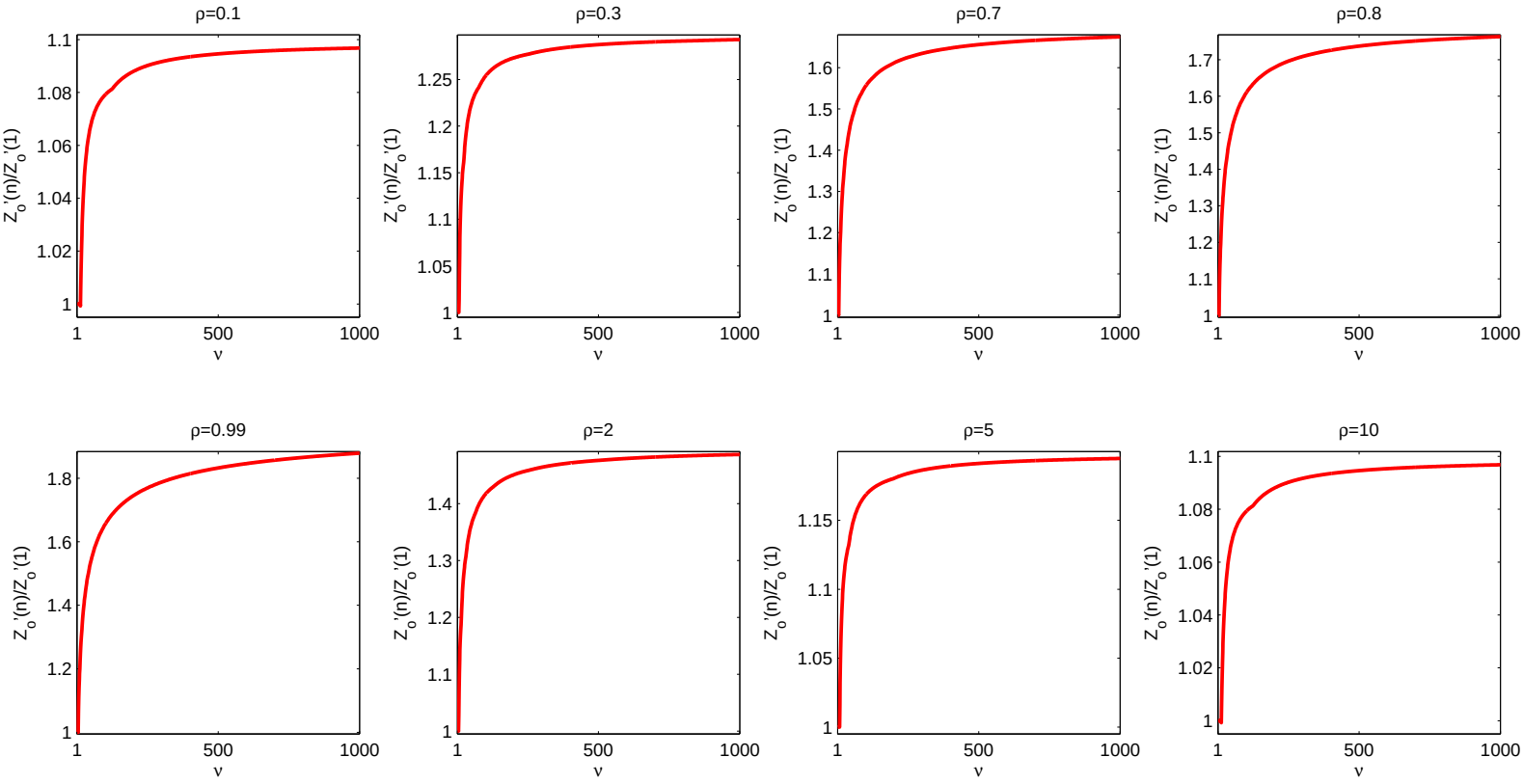

Figure 12: The ratio between optimal profit and when $n = 1$.

We can see that when $\rho = 1$, we get the highest ratio when $\nu \to \infty$. When $\rho = 1$, the ratio is equal to 2. We see this in Figure 13:

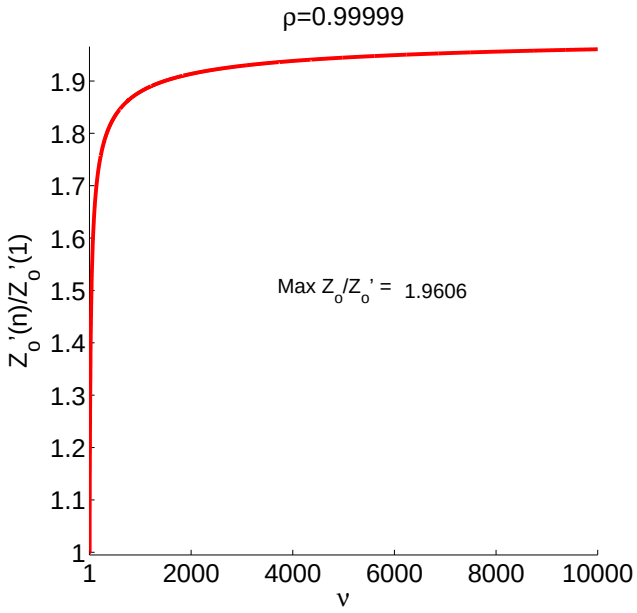

Figure 13: The ratio between optimal profit and when $n = 1$ for $\rho \to 1$.

We conclude that the best we can get from Naor's formula with n_m 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 ρ is close to 1 and ν has a large value, it is worth maintaining a queue.

Also, when $\rho > 1$, we see from the graphs that the ratio is getting smaller, and approaches 1 for higher values of ρ. Particularly, we see in Figure 14 that for ρ slightly larger than 1, the ratio is less than 2.

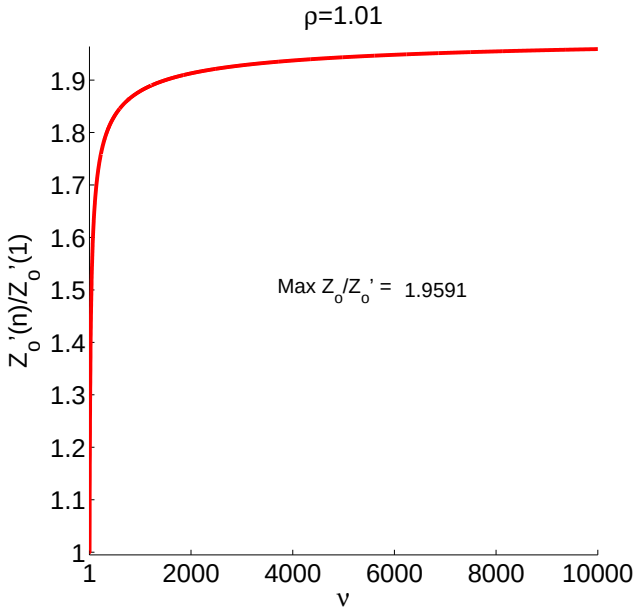

Figure 14: The ratio between optimal profit and when $n = 1$ for $\rho = 1.01$.

9 Conclusions

We consider a model in which arriving customers are being informed whether the queue length is below (at least) a given threshold, and pay p_l (p_h) if they decide to join the queue. Such model attains a profit equal to the socially optimal value. Our model is more convenient to implement and does not have some of the drawbacks of previously presented related models. We conduct analytical and numerical studies on our queueing system and conclude that in order to attain optimality, customers shouldn't enter when the admission fee is p_H, and all of the customers should enter when the admission fee is p_L.

Moreover, almost always the optimal N is equal to 1. In some cases the gain from $N = 2$ is significant, but never more than $\frac{1}{3}$. Generally, we cannot attain more than twice the profit we attain when setting $N = 1$, but in most cases the attained difference is insignificant. Since we have concluded that customers shouldn't enter when the admission fee is p_H, $N = 1$ means that customers only enter if the system is empty, so no queue is maintained.

We compare our queueing system to the classical observable and unobservable queueing systems. Our model yields better results than the unobservable model, because it is a special case of our model, when $N = \infty$ or $N = 0$. The optimal profit we attain is always higher than the profit attained in the unobservable case, and the ratio of the optimal profits is unbounded.

The fact that maintaining a queue usually doesn't give a significant improvement, raised an interest to check 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 the same results and conclusions with regard to social welfare.

Thus we conclude that maintaining a queue is often insignificant and in most cases it yields a limited improvement in profit or social welfare.

References

[1] Adiri, Igal and Uri Yechiali, Optimal Priority Purchasing and Pricing Decisions in Nonmonpoly and Monopoly Queues, *Operations Research*, 22 (1974), 1051-1066.

[2] Al-Athari, Faris Muslim, Estimation on the Mean of Truncated Exponential Distribution, *Journal of Mathematics and Statistics*, 4 (2008), 284-288.

[3] Allon, Gad, Achal Bassamboo and Itay Gurvich, "We Will be Right with You": Managing Customers with Vague Promises and Cheap Talk, *Operations Research*, 59 (2011), 1382-1394.

[4] Alperstein Hanna, Optimal Pricing Policy for the Service Facility Offering a Set of Priority Prices, *Management Science*, 34 (1988), 666-671.

[5] Altman, Eitan and Tania Jimenez, Admission Control to an M/M/1 Queue with Partial Information, *Lecture Notes in Computer Science*, 7984 (2013), 12-21.

[6] Chen, Hong and Murray Frank, State Dependent Pricing With a Queue, *IIE Transactions*, 33 (2001), 847-860.

[7] Dimitrakopoulos, Yiannis and Apostolos Burnetas, Customer Equilibrium and Optimal Strategies in an M/M/1 Queue with Dynamic Service Control, 2011.

[8] Dimitrakopoulos, Yiannis and Apostolos Burnetas, The Value of Service Rate Flexibility in an M/M/1 Queue with Admission Control, 2012.

[9] Dobson, Gregory and Edieal J. Pinker, The Value of Sharing Lead Time Information ,*IIE Transactions*, 38 (2006), 171-183.

[10] Economou, Antonis and Spyridoula Kanta, Optimal Balking Strategies and Pricing for the Single Server Markovian Queue with Compartmented Waiting Space , *Queueing Systems*, 59 (2008), 237-269.

[11] Edelson, N.M. and D.K. Hildebrand ,Congestion Tolls for Poisson Queueing Processes, *Econometrica*, 43 (1975), 81-92.

[12] Erlichman, Jenny and Refael Hassin, Strategic Overtaking in a Monopolistic Observable M/M/1 Queue, *Transactions on Automatic Control*.

[13] Hall, Joseph M., Praveen K. Kopalle, and David F. Pyke, Static and Dynamic Pricing of Excess Capacity in a Make-to-Order Environment, *Production and Operations Management*, 4 (1975),411-425.

[14] Hassin Refael and Moshe Haviv, *To Queue or not to Queue: Equilibrium Behavior in Queueing Systems*, Kluwer Academic Publishers, 2003.
Also http://www.math.tau.ac.il/∼hassin/book.html

[15] Hassin Refael, On the Optimality of First Come Last Served queues, *Econometrica*, 53 (1985), 201-202.

[16] Kim Young, Joo and Hark Hwang, Incremental Discount Policy of Cell Phone Carrier with Connection Success Rate Constant, *European Journal of Operational Research*, 196 (2009), 682-687.

[17] Le Ny, Louis-Marie and Bruno Tuffin, Pricing a Threshold-Queue with Hysteresis, 2007.

[18] Li, Na and Jiang Zhibin, Modeling and Optimization of a Product-Service System with Additional Service Capacity and Impatient Customers, *Computers and Operations Research*, 40 (2013), 1923-1937.

[19] Maoui, Idriss, Hayriye Ayhan and Robert D. Foley, Optimal static pricing for a service facility with holding costs, *European Journal of Operational Reasearch*, 197 (2009), 912-923.

[20] Masarani, F. and S. Sadik Gokturk, Price Setting Policies for Service Systems in Case of Uncertain Demand and Service Time, *Zeitschrift Operations Research*, 31 (1987), B97-B113.

[21] Naor, P. The Regulation of Queue Size by Levying Tolls, *Econometrica*, 37 (1969), 15-24.

[22] Perel, Nir, and Uri Yechiali, Queues with Slow Servers and Impatient Customers, *European Journal of Operational Research*, 201 (2010), 247-258.

[23] Shi, Xiutian, Houcai Shen, Ting Wu and T.C.E. Cheng, Production Planning and Pricing Policy in a Make-to-Stock System With Uncertain Demand Subject to Machine Mreakdowns, *European Journal of Operational Research*, 238 (2014), 122-129.

[24] Wang, Eric T.G., and Terry Barron, Computing Services Supply Management: Incentives, Information, and Communication, *Decision Support System*, 19 (1997), 123-148.