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

Hassin Refael
Tel-Aviv University
hassin@post.tau.ac.il

Koshman Alexandra*
Tel-Aviv University
koshmana@post.tau.ac.il

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 the level of congestion. Arriving customers cannot observe the queue length, but they are informed about 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 increase the profit and 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 whether there is a significant difference in profit when rather than being informed about the congestion level, customers only join the system when the server is idle. We also investigate this question considering the classical observable model.

Keywords

Game Theory, Queueing Theory, Dynamic Pricing

1. INTRODUCTION

This article deals with an unobservable single-server model. Consider a system with an unobservable M/M/1 queue. The service rate is μ. Customers are identical, so that they all have the same waiting cost rate C and the same service value R.

Our article deals with a restricted type of dynamic pricing. For a threshold N, if the number of clients in the system, n, is less than N the admission fee is p_L, and it is p_H otherwise. New customers cannot observe the exact queue length, but they know whether the number of customers in the system satisfies $n < N$ or $n \geq N$.

We mention three extreme cases.

(1) $N = \infty$. In this case we have a simple unobservable queueing system.

(2) $N = 0$. In this case we again have a simple unobservable queueing system.

(3) $N = 1$. In this case customers are only informed whether the server is free or not.

Naor (1969) showed that if the overall social welfare is to be maximized then for some threshold value n^*, customers should join the queue only if its length is $\leq n^* - 1$.

The best choice of admission fees would be dynamic pricing, as first observed by Chen and Frank (2001). 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 the arriving customers join according to the threshold n^* and give all of their welfare to the server. This is possible if the server sets dynamic prices, i.e., charging $p(n) = R - C \cdot \frac{n+1}{\mu}$ from a customer who observes $n < n^*$ customers upon arrival, and a higher price otherwise. In this case we attain the socially optimal situation. 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 fees model follows from an article by Hassin (1986), that describes an observable LCFS-PR queueing model. All of the arriving customers join the queue. The last customer decides whether or not to abandon the queue. Since the last customer remains last until he is served or he abandons 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 $n^* + 1$, the socially optimal decision. Note that all arriving customers have the same expected utility which is independent of the queue length. Therefore the server can obtain all of the social welfare by charging the maximal price which 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.

*This research was supported by the Israel Science Foundation grant number 1015/11.

VALUETOOLS 2014, December 09-11, Bratislava, Slovakia

DOI 10.4108/icst.valuetools.2014.258190

A third possibility for achieving S^* follows from work on priority sales 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 the threshold n^* and leaving no customer surplus. An arriving customer buys the lowest priority that has no current customer, and balks if all n^* priorities have customers. To achieve this strategy, we set the price for priority i to be the expected utility of a customer who buys this priority assuming that 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, customers behavior is socially optimal, and the server's profit attains its upper bound. An advantage of this model is that although the outcome is again LCFS among the customers who obtain service, customers may not feel it is unfair because they choose the type of priority to purchase. Also, those who pay eventually obtain service and those who balk do not incur any costs, whereas under the LCFS regime with a single price the waiting costs of reneging customers are not refunded.

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

Our model has many advantages over the other systems. On the one hand, it has a single price and does not have high switching costs. On the other hand, it is fair towards the customers, since it offers FCFS policy, and it has the lowest variance of the number of customers an arriving customer has to wait for, among the presented models.

Several papers consider delay announcements of the type considered here, namely the firm announces whether or not the queue length is above a given threshold. Among such articles are articles written by Altman and Jimenez (2004), Dobson and Pinker (2006), Le Ny and Tuffin (2007) and Allon, Bassamboo and Gurvich (2011),

Other models assume that the service rate changes when the queue length exceeds a threshold. Articles written by Perel and Yechiali (2010), Dimitrakopoulos and Burnetas (2011) and Li and Jiang (2013) discuss such models.

An article which has a special case that resembles 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 in *compartments* of fixed capacity, and before entering the customer is told the compartment in which he will enter and/or the position within the compartment in which he will enter.

Several articles change the queue managing tactics based on the expected waiting time of an arriving customer. Articles written by Kim and Hwang (2009) and Maoui, Ayhan and Foley (2009) are examples of such articles.

We also raise the question of the significance of maintaining a queue and keeping waiting spaces. A similar problem is raised by Masarani and Gokturk (1987), who investigate the profit maximizing size of the waiting room in a queueing system.

2. FINDING THE EQUILIBRIUM λ_L, λ_H AND DEFINING THE PROFIT MAXIMIZATION PROBLEM

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

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 has to decide whether to join the queue or balk. The utility from balking is 0, and the utility from joining the queue is R minus C multiplied by the time spent in the system. The customer cannot wait outside, and if he doesn't join the queue, he leaves and does not come back.

For each admission fee, the arrival process when $n < N$ or $n \geq N$ is Poisson with the 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 values of $\lambda_L, \lambda_H, \mu$ and N, we have a birth and death process such that the birth rate is λ_L if $n < N$, and λ_H otherwise. If $n > 0$, the death rate is equal to μ.

We reduce the number of parameters by introducing the following normalized parameters:

$$\nu = \tfrac{R\mu}{C} \ , \ \rho_L = \tfrac{\lambda_L}{\mu} \ , \ \rho_H = \tfrac{\lambda_L}{\mu} \text{ and } \rho = \tfrac{\Lambda}{\mu}.$$

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

Let π_n be the probability that there are n customers in the system. Then

$$\pi_n = \begin{cases} \rho_L^n \cdot \pi_0, & n \le N \\ \rho_L^N \cdot \rho_H^{n-N} \cdot \pi_0, & n > N. \end{cases} \quad (1)$$

Now we substitute this result into $\sum_{n=0}^{\infty} \pi_n = 1$ and obtain:

$$\pi_0 = \left(\frac{1-\rho_L^N}{1-\rho_L} + \frac{\rho_L^N}{1-\rho_H} \right)^{-1}. \quad (2)$$

Let $W_{<N}(\lambda_L, \lambda_H)$ be the expected waiting time of a customer who joins the system when there are less than N customers in the system. Similarly, denote $W_{\ge N}(\lambda_L, \lambda_H)$. The net utility from joining is $U_L := R - p_L - C \cdot W_{<N}$ in the L state, and $U_H := R - p_H - C \cdot W_{\ge N}$ otherwise.

Without loss of generality, we assume that in equilibrium customers are indifferent between joining and balking (this property will always hold under profit maximization). Thus, in equilibrium the following two conditions hold under the optimal prices p_L and p_H:

$$\begin{cases} C \cdot W_{<N} = R - p_L \\ C \cdot W_{\ge N} = R - p_H. \end{cases} \quad (3)$$

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

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

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

$$W_{\ge N} = \sum_{i=N}^{\infty} \frac{\pi_i}{\sum_{j=N}^{\infty} \pi_j} \cdot W_i = \frac{1}{\mu} \cdot \frac{(N+1) - N \cdot \rho_H}{1-\rho_H}$$

Inserting $W_{<N}$ and $W_{\ge N}$ in (3), we get that:

$$\begin{aligned} p_L &= R - \frac{C}{\mu} \cdot \frac{N \cdot \rho_L^{N+1} - (N+1) \cdot \rho_L^N + 1}{(1-\rho_L^N)(1-\rho_L)} \\ &= \frac{C}{\mu} \cdot \left(\nu - \frac{N \cdot \rho_L^{N+1} - (N+1) \cdot \rho_L^N + 1}{(1-\rho_L^N)(1-\rho_L)} \right) \quad (4) \end{aligned}$$

$$\begin{aligned} p_H &= R - \frac{C}{\mu} \cdot \frac{(N+1) - N\rho_H}{1-\rho_H} \\ &= \frac{C}{\mu} \cdot \left(\nu - \frac{(N+1) - N\rho_H}{1-\rho_H} \right). \quad (5) \end{aligned}$$

The server faces the following optimization problem:

$$\max \tilde{\Pi}(N) = \max_{p_L, p_H} \left(\Pr(n < N) \cdot \ p_L \lambda_L + \Pr(n \ge N) \cdot p_H \lambda_H \right)$$

Define the normalized profit $\Pi(N) = \frac{\tilde{\Pi}(N)}{C}$. Maximizing $\tilde{\Pi}(N)$ is equivalent to solving:

$$\begin{aligned} \Pi(N) &:= \max_{\rho_L, \ \rho_H} \tilde{\Pi}(N) \\ &= \max_{\rho_L, \ \rho_H} \Big\{ \frac{(1-\rho_L^N)(1-\rho_H)}{(1-\rho_L^N)(1-\rho_H) + \rho_L^N(1-\rho_L)} \cdot \rho_L \cdot \\ &\cdot \left(\nu - \frac{N \cdot \rho_L^{N+1} - (N+1)\rho_L^N + 1}{(1-\rho_L^N)(1-\rho_L)} \right) \\ &+ \left(1 - \frac{(1-\rho_L^N)(1-\rho_H)}{(1-\rho_L^N)(1-\rho_H) + \rho_L^N(1-\rho_L)} \right) \cdot \rho_H \\ &\cdot \left(\nu - \frac{(N+1) - N \cdot \rho_H}{1-\rho_H} \right) \Big\} \quad (6) \end{aligned}$$

2.1 Numerical results for $\Lambda = \infty$

We now solve the maximization problem (6) numerically, assuming $\Lambda = \infty$. We have $\nu > 1$ as the input. The graphs in Figure 1 examine the behavior of $\Pi(N)$ for $1 \le N \le 20$ and various values of ν.

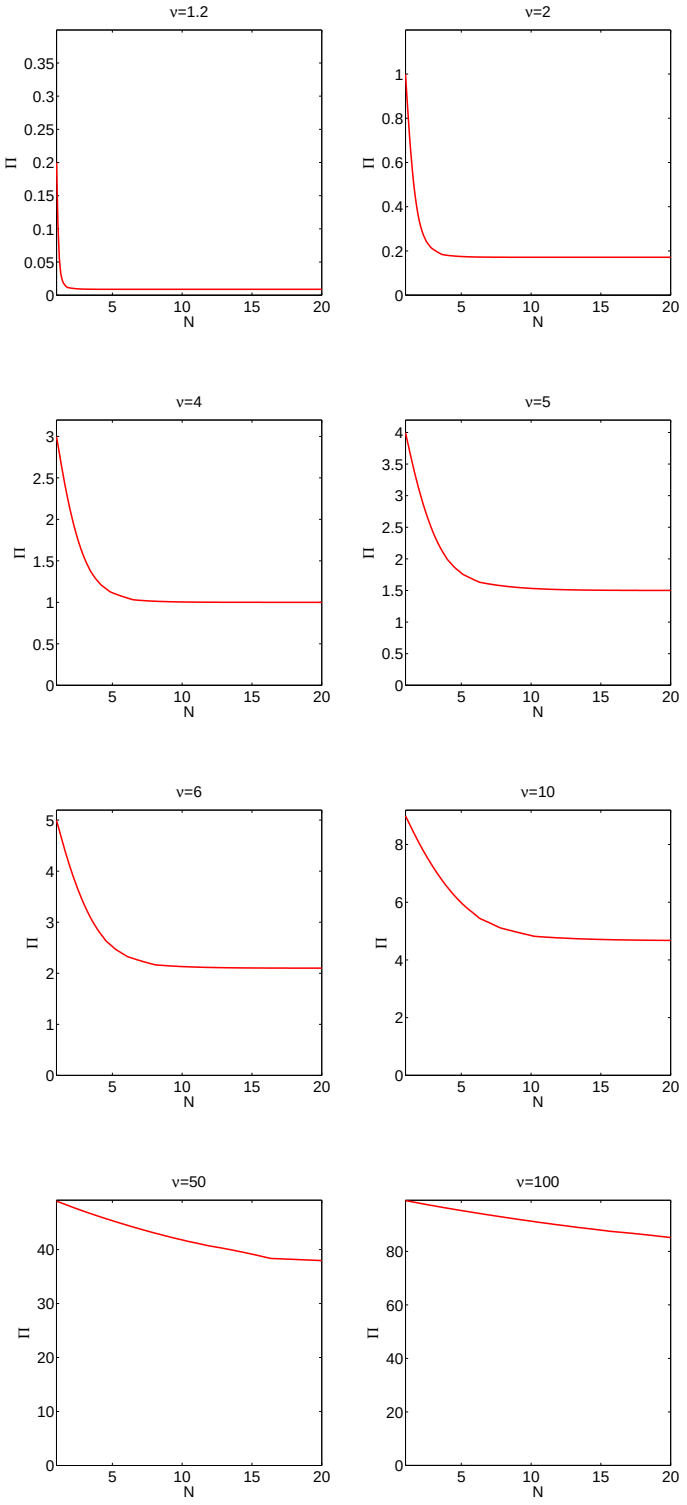

Figure 1: The behavior of Π for different values of ν and $1 \leq N \leq 20$.

As expected, $N = 1$ is optimal for every value of ν. We also conclude that the ratio of the optimal profit value to the profit in the unobservable model satisfies $\frac{\Pi_{N=1}}{\Pi_{N\to\infty}} = \frac{\nu-1}{(\sqrt{\nu}-1)^2} = \frac{\sqrt{\nu}+1}{\sqrt{\nu}-1}$, which grows to ∞ when $\nu \downarrow 1$. On the other hand, when ν is very large, we have a ratio of approximately 1.

2.2 Numerical results for $\Lambda < \infty$

We now solve the problem numerically assuming that $\Lambda < \infty$. It means we have one more parameter, ρ, in addition to ν. The graphs in Figure 2 examine the behavior of $\Pi(N)$ for different values of N, ν and ρ.

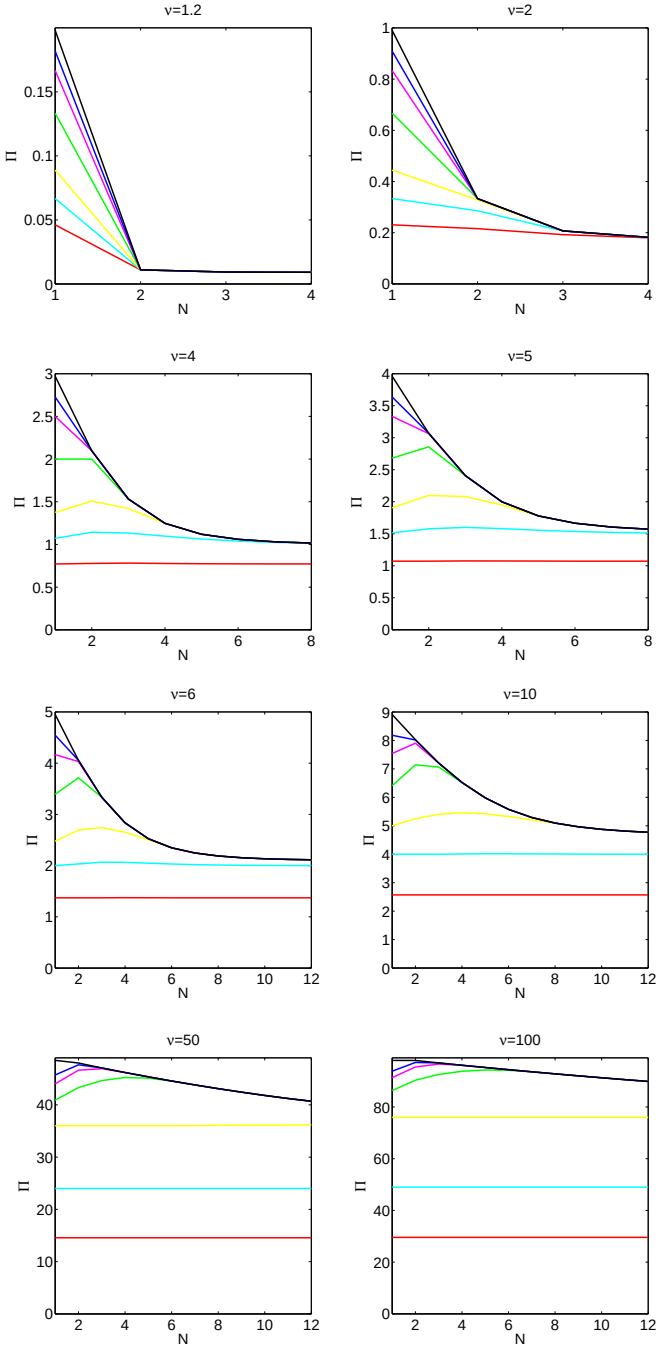

Figure 2: The behavior of Π for different values of ρ and different values of ν. The values of ρ from lower graphs to the higher ones are 0.3, 0.5, 0.8, 2, 5, 10, 100.

When $N \to \infty$, we have the unobservable model, so that $\Pi_{N\to\infty} = \rho\left(\nu - \frac{1}{1-\rho}\right)$.

The flat lines that correspond to small values of ρ can be explained by the fact that when ρ is small, there are almost always less than N in the system. An arriving customer always pays p_L, and thus the system resembles an unobservable queue. The profit rate function is then equal to $\tilde{\Pi} = \lambda \cdot p_L = \lambda\left(R - \frac{C}{\mu}\right)$. Then $\Pi = \frac{\lambda R}{C} - \frac{\lambda}{\mu} = \rho(\nu - 1)$. For example, when $\nu = 100$ and $\rho = 0.3$, we have $\Pi = 0.3 \cdot 100 = 30$.

We also observe that for values of N lower than n^*, customers join the system when the admission fee is p_H, and for values of N equal or greater than n^* they balk when the state is H.

3. THE GAIN FROM MAINTAINING A QUEUE

We notice that the profit maximizing threshold is often equal to 1. We therefore check how much can be gained from choosing $N = n^* > 1$ and gaining $\Pi(n^*) = S^*$, instead of letting the customers join only when the server is idle.

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. Then $\Pi'(1) = \frac{\rho_L}{1+\rho_L} \cdot (\nu - 1)$. The ratio is then equal to:

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

It is straightforward to prove that $\Pi'(1) \geq \frac{1}{2} \cdot S^*$. We also conclude that the largest ratio is attained when $\rho = 1$. This conclusion is illustrated in Figure 3.

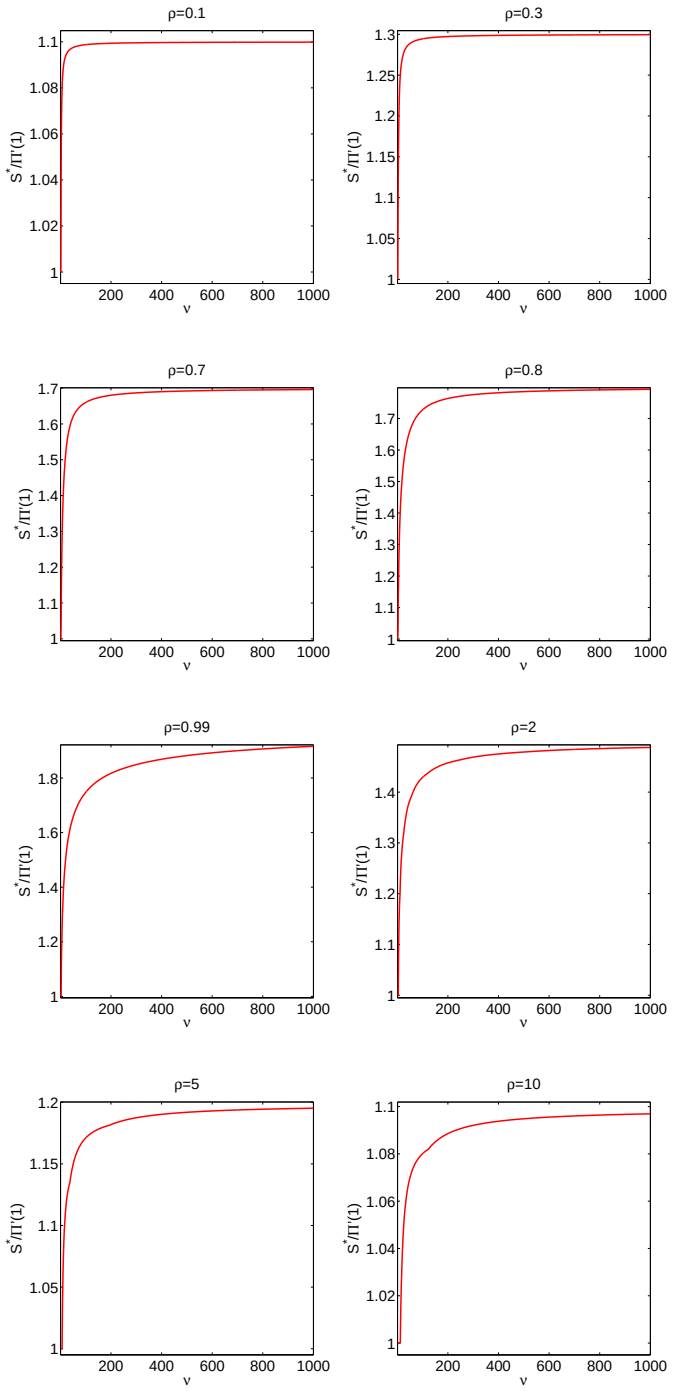

Figure 3: The loss associated with setting a threshold 1 instead of n^*, when n^* is greater than 1.

We also consider the question of the difference between setting a threshold $N = n^* > 2$ instead of letting arriving customers join the queue only if the system is empty or if there is only one customer inside the system, who is being currently served. It can be shown that $\Pi'(2) \geq \frac{3}{4} \cdot S^*$. Moreover, we see that for higher values of n, $\frac{S^*}{\Pi'(n)} \approx 1$. Thus, if waiting space is costly, it would often not be optimal to have more than one waiting space.

4. THE PROFIT ATTAINED BY USING THE OPTIMAL N_M OR 1 IN THE OBSERVABLE MODEL

We observed that in our model we often cannot obtain a much better result by setting N to be larger than 1, and in any case the ratio of profits is bounded by 2. 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 $n = 1$ and price $p = R - \frac{C}{\mu}$.

Naor's profit maximization formula for this model is:

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

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 that:

$$Z_o' = \frac{Z_o}{C} = \rho \cdot \frac{1-\rho^{n_m}}{1-\rho^{n_m+1}} \cdot [\nu - n_m]. \tag{7}$$

Using (7), we see that when $n_m = 1$,

$$Z_o' = \frac{Z_o}{C} = \frac{\rho}{1+\rho} \cdot (\nu - 1). \tag{8}$$

We find that the best we obtain from Naor's formula with n_m yields at most twice the profit attained with $n = 1$. We also find that largest ratio is attained when $\rho = 1$. This conclusion is illustrated in Figure 4.

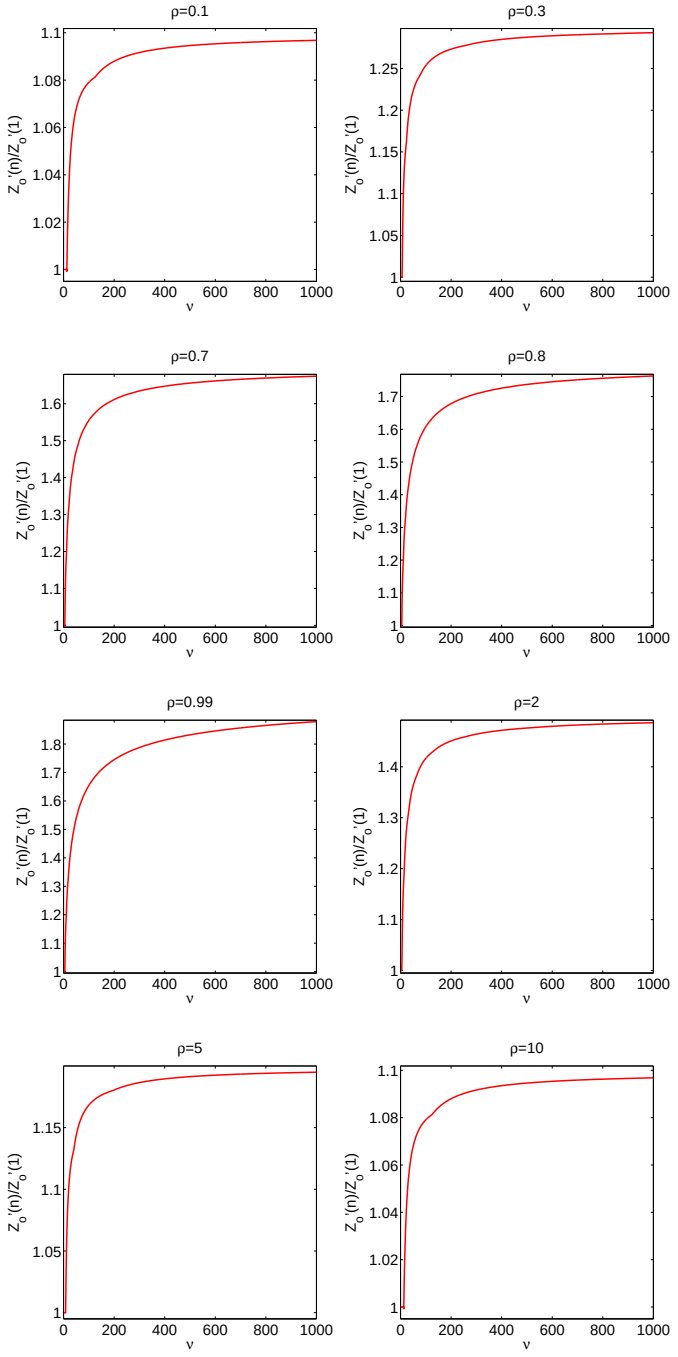

Figure 4: The loss associated with setting a threshold 1 instead of n_m, when n_m is greater than 1.

5. CONCLUSIONS

We consider a mechanism which attains a profit equal to the maximal socially optimal value. Our mechanism is more convenient to implement and does not have the drawbacks of previously presented mechanisms.

In many settings, the optimal threshold N is equal to 1. Generally, we cannot attain more than twice the profit we attain when setting $N = 1$, i.e., not maintaining a queue, and in most cases the attained difference is insignificant. Furthermore, we cannot attain more than $\frac{4}{3}$ the profit we attain when setting $N = 2$, i.e., letting arriving customers join the queue only if there is at most one customer in the system.

Similarly, we find that maintaining a queue in Naor's observable model cannot guarantee a much higher profit. Maintaining a queue can yield no more than twice, and usually much less, than twice the profit that can be attained using $n = 1$.

We conclude that maintaining a queue is often insignificant and in many cases it does not yield a valuable improvement in profit or social welfare.

6. REFERENCES

[1] Adiri, Igal and Uri Yechiali, Optimal priority purchasing and pricing decisions in nonmonpoly and monopoly queues, Operations Research, 22, 1051-1066, 1974.

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

[3] Alperstein Hana, Optimal pricing policy for the service facility offering a set of priority prices, Management Science, 34, 666-671, 1988.

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

[5] Chen, Hong and Murray Frank, State dependent pricing with a queue, IIE Transactions, 33, 847-860, 2001.

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

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

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

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

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

[11] 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

[12] Hassin Refael, Consumer Information in Markets with Random Products Quality: The Case of Queues and Balking , Econometrica, 54, 1185-1195, 1986.

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

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

[15] 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, 1923-1937, 2013.

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

[17] Masarani, F. and S. Sadik Gokturk, Price setting policies for service systems in case of uncertain demand and service time, Zeitschrift Operations Research, 31, B97-B113, 1987.

[18] Naor, P. "The Regulation of Queue Size by Levying Tolls," Econometrica, 37, 15-24, 1969.

[19] Perel, Nir, and Uri Yechiali, Queues with Slow Servers and Impatient Customers, European Journal of Operational Reasearch, 201, 247-258, 2010.

[20] 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 breakdowns, European Journal of Operational Reasearch, 238, 122-129, 2014.