Equilibrium Threshold Strategies: The Case of Queues with Priorities

Refael Hassin[1] and Moshe Haviv[2]

Abstract

Multiplicity of solutions is typical to systems where the individual's tendency to act in a certain way increases when more of the other individuals in the population act in this way. We provide a detailed analysis of a queueing model in which two priority levels can be purchased. In particular, we compute all of the Nash equilibrium strategies (pure and mixed) of the threshold type.

OR/MS classification: Queues: Priority. Games/group decisions: Noncooperative.

1 Introduction

A rational customer who needs the service of a facility has several decisions to take. He chooses the time to arrive, decides whether to join a queue (or *balk*), whether to purchase priority, and after staying for a while in line he may decide to leave (i.e., *renege*) it and give up service. In a system with parallel lines the customer may sometimes choose which line to join.

These decisions depend on the customer's cost and benefit parameters and on the information he possesses at the time each decision is taken. It is often the case that the customer's self-optimizing decisions also depend on the actions of the other customers in the population, actions which at the time of decision are *not* known to him. This puts the question of customers' behavior in a game theoretical framework, rather than in an optimization framework.

For individuals whose objective is to reduce the amount of time they spend in congested systems, a naturally good policy is to try and act differently from others. For example, avoid rush-hour traffic, avoid popular restaurants, etc. In these cases, the individual's tendency to select an action decreases with the tendency of the others in the population to choose it. We refer to this behavior as *avoid the crowd* (ATC). In the opposite case, named *follow the crowd* (FTC), individuals try to imitate others. This behavior is prevalent under numerous

[1] Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel. **email:** hassin@math.tau.ac.il

[2] Department of Statistics, The Hebrew University, Jerusalem 91905, Israel, and Department of Econometrics, The University of Sydney, Sydney, NSW 2006, Australia. **email:** mosheh@sue.econ.su.oz.au

circumstances and for various reasons (see for example, [23], [5], [6] and [27]), but as we will see, this is sometimes the case when avoiding congestion and reducing waiting time are the main objectives.

As a vehicle to convey the concept of FTC and the consequences stemming from it, we analyze a model that has been treated by Adiri and Yechiali [1], in which customers choose between priority levels. We characterize *threshold* type Nash equilibrium strategies, namely strategies in which one purchases priority if and only if upon arrival the queue size is larger than some threshold value. The main reason for the interest in Nash equilibrium strategies is the desire to predict and control customers' behavior. Therefore, we will give a special attention to the existence of multiple solutions which, as we argue in the next section, is typical for the cases in which FTC prevails. We will show that multiplicity of equilibria is a common phenomenon in a class of models including the one considered in this paper.

Weissman [27] obtained a similar result while modeling the demand for priority in electricity supply. In this model, the total supply is a random variable. In the case of a shortage, the supply is first distributed to a group of high priority customers and whatever is left is distributed among the others. Weissman showed that for this model the demand curve for priority may have segments with a positive slope and consequently there may be multiple equilibrium solutions. This phenomenon is related to our concept of FTC since it refers to cases where the more customers buy priority, the higher is the price one is ready to pay for it. Another model of this type is analyzed by Viswanathan and Tse [26]. There, uniqueness of the supply probabilities under pure equilibrium strategies is proved.

The next section contains our framework. In particular, the concepts of ATC and FTC are defined in an environment where it is natural to consider threshold strategies. The rest of the paper deals with the specific model. In Section 3 we present the model of a queueing system with two priority levels. We forbid balking at this stage, and defer the minor changes necessary when balking is allowed to Section 6. In Section 4 we present necessary and sufficient conditions for a given threshold to define a Nash equilibrium. In Section 5 we comment on the number of pure and mixed equilibrium strategies and their structure. In Section 6 we discuss the model with optional balking. We also present in this section a (pathological) example where there exists a pure Nash equilibrium which is *not of the threshold type.* In an appendix we describe a computational procedure for the expected waiting time of a customer who joins the low priority queue when its length is equal to the threshold, n. These values are necessary for computing Nash equilibria. Our algorithm is faster than the one described in [1].

2 The framework

Our setting concerns arrivals of customers to a queueing system characterized by a single non-negative integer-valued state variable (typically, a queue length). Each arriving customer observes the state and chooses between two actions, A_1 and A_2. The customer's objective is to minimize his expected costs. An arriving customer's decision typically depends on actions taken by the other (past or future) arrivals, actions that are unknown to him. A customer's *strategy* is a function $s(L) : I\!N \to [0, 1]$. Its interpretation is that when the observed state is L, the customer chooses A_1 with probability $s(L)$ and A_2 with the complementary probability, $1 - s(L)$.

We consider strategies in which customers choose A_1 if and only if the system's state is less than a given constant called a *threshold*. However, it is often possible to construct instances where, for example, if everyone in the population has a threshold 4 then a deviant whose threshold is 5 has a smaller expected wait, while if everyone in the population adopts a threshold of 5 then deviating to a threshold of 4 reduces the expected wait. (This is the case with the upper function in Figure 1, as will be explained later.) In such cases we may conclude that no strategy of the threshold type defines an equilibrium. Consequently, we extend the definition of a threshold strategy as follows:

A *threshold strategy* with *threshold* $x = n + p$, $n \in I\!N$, $p \in [0, 1)$, has

$$s(L) = \begin{cases} 1 & L = 0, ..., n-1 \\ p & L = n \\ 0 & L = n+1, n+2, ... \end{cases} .$$

In other words, under a strategy with threshold x, a customer always selects A_1 if the state is at most $n - 1$, always selects A_2 if the state is at least $n + 1$, and randomizes between the two with probability p when the state is n. If x is an integer ($p = 0$), the strategy is *pure*. Otherwise, it is *mixed*.

We are interested in models in which the optimal response of an individual who assumes that the others follow the strategy defined by x is of the threshold type: For some integer $k(x)$, if the state is in $\{0, ..., k(x) - 1\}$ choose A_1. Otherwise, choose A_2. This is the case in the model described in the next section as we prove in the next section. The discontinuity jumps of $k(x)$ correspond to threshold strategies x where the individual is indifferent between two consecutive thresholds. A threshold x defines a *Nash equilibrium* if either $k(x) = x$ or x is between $k(x-)$ and $k(x+)$ (it is convenient to view both cases as solutions to the equation $k(x) = x$, that is, as fixed points of $k(x)$). In both cases, if all customers adopt threshold strategy x then this is also an optimal strategy for each of them and none has an incentive

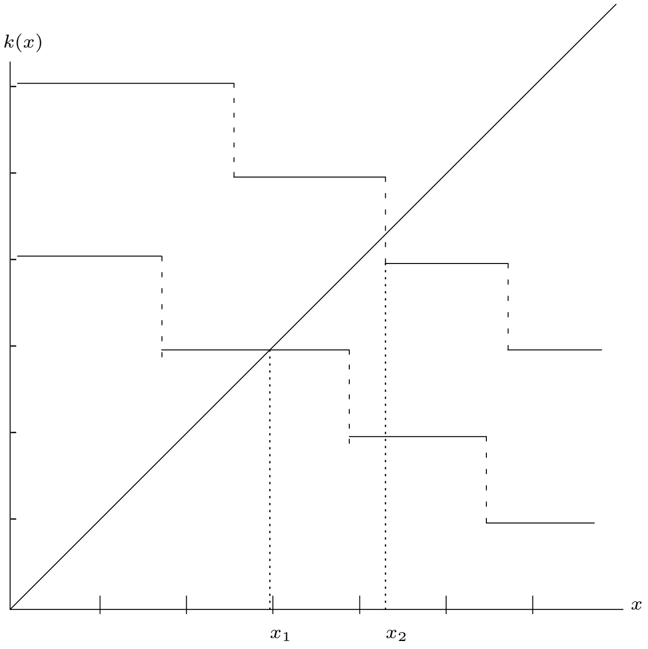

Figure 1: Equilibrium under ATC

to deviate from this strategy. In the case of a mixed strategy with threshold x, then given that the others adopt this x, the individual is indifferent between the two thresholds $k(x-)$ and $k(x+)$.

When the individual's tendency to choose A_1 increases with x, $k(x)$ is monotone non-decreasing. We refer to this type of individual strategy as *follow the crowd*, or in short, FTC. It means that the higher is the threshold adopted by others, the higher is the optimal threshold for a given customer. When the individual's tendency to choose A_1 decreases with x, $k(x)$ is monotone non-increasing. We refer to this type of individual strategy as *avoid the crowd*, or in short, ATC. It means that the higher is the threshold adopted by others, the lower is the optimal threshold for a given customer.

There are important differences between the two cases. In ATC case there is a single fixed point. It may describe a pure strategy or a mixed one. Figure 1 depicts two non-increasing step functions. In one, the equilibrium strategy, obtained at x_1, is pure, in the other, the strategy, obtained at x_2, is mixed. The FTC case is more involved. It may have multiple equilibria, of both pure and mixed type. It can be seen from Figure 2 that there may be numerous fixed points.

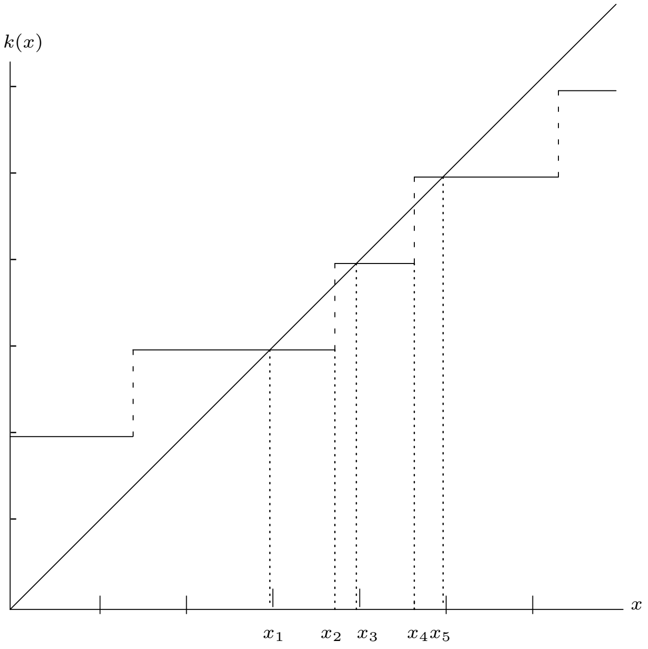

Figure 2: Equilibrium under FTC

3 Queues with priorities

The fundamental model involving customers' decisions in queues with priorities is that of Adiri and Yechiali [1].[3] They made the following assumptions: Two queues are formed in front of a single server, one line for high priority customers and one for low priority customers. A high priority customer who finds upon arrival a low priority customer being served, commences service immediately. The preempted customer resumes service only when no high priority customer is present.[4] The arrival process is Poisson with rate λ per unit of time. Upon arrival, each customer decides whether to purchase high priority, and if so he pays for it an amount θ. Without loss of generality, we assume that low priority comes for free. Customers cannot change priority levels while waiting. The arrivals decide on their priorities after observing the length of the two queues. The cost of waiting is assumed, without loss of generality, to be one per unit of time. The customers' service times are independent and exponential with mean $1/\mu$ regardless of the priority level. Let $\rho = \lambda/\mu$, and to ensure stability assume that $\rho < 1$. Finally, let B be the expected length of a busy period, i.e., $B = 1/(\mu - \lambda)$.

[3] Other models in which customer select priorities are discussed in [4, 8, 12, 14, 20, 22, 25].

[4] Similar results to the ones described here hold in the the case without preemption, known as the *head-of-the-line* priority discipline. These results can be obtained after straightforward modifications.

Denote by (i,j), $i,j \geq 0$, a typical state of the system where i (j) is the number of high (low) priority customers present in the system. An arrival observing a given state has to decide whether to purchase priority or not in order to minimize his expected total cost. Suppose that each arrival has his own policy which tells him for each possible state (i,j) whether to purchase priority or not. The optimal action of a customer depends on the state of the system and the policies adopted by future arrivals. Thus, the solution concept to be adopted here is that of a Nash equilibrium.

Adiry and Yechiali [1] proved that if for some strategy adopted by everybody, it is optimal for an individual to purchase priority when he faces state (i,j), then his *unique* optimal action when facing state (r,j) for $r > i$, is to purchase priority. It follows that for a system that initializes with state $(0,0)$, the state space is 1-dimensional in the sense that the only possible states are $(0,j)$, $j = 0,...,n$ and (i,n), $i = 1,2,...,$. Therefore, we can map the state space into a single dimension: Let i be the *total* number of customers in the system. If $i \leq n$ then the state is $(0,i)$, and if $i > n$ then the state is $(i-n,n)$.

The mixed threshold strategy $n+p$ with $0 < p < 1$, can also prescribe a Nash equilibrium. In this case a customer facing state $(0,n-1)$ is indifferent between purchasing priority and not doing so. The next state then is $(0,n)$ or $(1,n-1)$. However, in both states it is optimal for a future arrival to purchase priority. Specifically, if in state $(0,n-1)$ is was optimal to purchase priority, this is certainly (and uniquely) the case in state $(0,n)$ as one more customer is being overtaken. Likewise, using the argument given above, if in state $(0,n-1)$ it is optimal to purchase priority, then the same (and uniquely) is the case in state $(1,n-1)$. Therefore, the relevant state information is still the total number of customers in the system.

We conclude that the model belongs to the framework of the previous section in the sense that when the queue length is the state variable, threshold strategies can be defined and that they prescribe optimal response for an individual when everybody else uses a common threshold strategy. In the next theorem we prove that the model leads to FTC, the case in which multiple Nash equilibria are common.

Proposition 3.1 *Suppose that the customers in the population, except for a given individual, adopt a common threshold x. Then, $k(x)$, the (integer) optimal threshold for this individual, is non-decreasing in x.*

Proof. Tag a customer and let x be a possible threshold value used by all others. Adiri and Yechiali show that it is optimal for the tagged customer to purchase priority if and only if the number of customers in the system is larger than $k(x)$.

For the tagged customer there are two types of gain from buying priority. The first is in overtaking the low priority customers who are at the system at the time of his arrival. The second is in avoiding being overtaken by future high priority customers who arrive while he is still in the system. Since in the model considered here, an arrival sees the queue lengths upon arrival, the first type of gain depends only on how many low priority customers he finds upon arrival and is not a function of the threshold used by others. The second type of gain depends also on the behavior of future customers. Now, let $x_1 < x_2$ and assume that n, the number of customers in the system is larger than $k(x_2)$. The second type of gain associated with purchasing priority is larger when the other customers use x_1 as their threshold strategy, as compared to x_2. Hence, since under the given queue size, $n > k(x_2)$, it is optimal to purchase priority when x_2 is used by all others, it is certainly optimal to do so under x_1. In particular, $k(x_1) < k(x_2)$. $\Box$

4 Conditions for equilibrium

Suppose the threshold x is adopted by the customers. Note that in this case the maximum possible length of the low priority queue is $\lceil x \rceil$ (the lowest integer that is greater than or equal to x). We are interested in the (future) expected waiting time (queueing plus service) of a customer at the $\lceil x \rceil$-th position of the low priority queue while the high priority queue is empty. We denote this value by $E(x)$. Clearly, $E(x)$ is defined only for $x > 0$. This section concerns necessary and sufficient conditions on $E(x)$ to warrant that x corresponds to a Nash equilibrium. In the Appendix we show how to compute $E(x)$ for any given x.

Proposition 4.1 *The integer threshold $n \geq 1$ specifies a Nash equilibrium if and only if*

$$\frac{1}{\mu} + \theta - B \leq E(n) \leq \frac{1}{\mu} + \theta \quad .$$

The threshold $n = 0$ specifies a Nash equilibrium if and only if $\theta + 1/\mu \leq B$.

Proof. Assume that the entire population uses the integer threshold $n \geq 1$. In order for n to describe an optimal strategy for an individual given that all follow this strategy, two conditions are necessary: First, if upon arrival he sees state $(0, n-1)$ his optimal action is not to buy priority, so that

$$E(n) \leq \frac{1}{\mu} + \theta \quad .$$

Second, if he sees state $(0, n)$, his optimal action is to buy priority, so that

$$B + E(n) \geq \frac{1}{\mu} + \theta \quad .$$

Moreover, these conditions are also sufficient: If it is optimal to buy priority at state $(0, n)$, it is also optimal to do so in states (i, n) for $i \geq 1$. Likewise, if it is optimal not to buy priority at $(0, n-1)$, it is also optimal not to do so at $(0, j)$ when $j \leq n-2$. Finally, the fact that $n = 0$ prescribes Nash equilibrium if and only if $\theta + 1/\mu \leq B$ is straightforward. □

Under an equilibrium strategy with threshold $x = n + p$, $0 < p < 1$, an individual who sees state $(0, n)$ may, with a positive probability, buy priority or he may choose not to buy it. This means that such an individual is indifferent between the two options of purchasing priority or not:

Proposition 4.2 *The threshold $x = n + p$, $0 < p < 1$ specifies a Nash equilibrium if and only if*

$$E(x) = \theta + \frac{1}{\mu} \quad .$$

Recall that $E(x) = B$ for all $0 < x \leq 1$. Indeed, for a customer in the low priority queue the exact value of x is unimportant as long as it is in $(0, 1]$. Combining this observation with the second part of Proposition 4.1 we have the following corollary:

Corollary 4.3 *If $\theta + 1/\mu = B$ then all of the values $0 \leq x \leq 1$ define Nash equilibria.*

Remark 4.4 The strategy $n + p$, when $0 < p < 1$ can be looked at as a mixing of the two pure strategies n and $n + 1$. Mixing among more than two pure strategies cannot lead to Nash equilibrium. It requires that an individual be indifferent between buying priority and not buying when he observes two different low priority queue sizes. However, this cannot be the case since the cost associated with buying priority is independent of the low priority queue length, while that associated with joining that queue increases with its length.

Figure 3 illustrates the function $E(x)$. It represents the actual function computed with $\lambda = 0.8$ and $\mu = 1$ so that $B = 5$. There are several observations which can be made. First, the function is continuous from the left and it has discontinuity points at the integers. Second, for $0 < x \leq 1$ it is equal to B, since the customer in question is the first in the low priority queue and until his departure all of the arrivals will join the high priority queue. Third, for a fixed value of $\lceil x \rceil$, $E(x)$ is monotone decreasing in x. This is the case since when x increases customers are less likely to purchase priority, thereby reducing the future waiting time of the customer currently in position $\lceil x \rceil$ of the low priority queue. Fourth, the size of the jump $E(n+) - E(n)$ is exactly of size B. This is the case since $E(n+)$ equals the expected length of a busy period B (the time it takes him to reach position n from

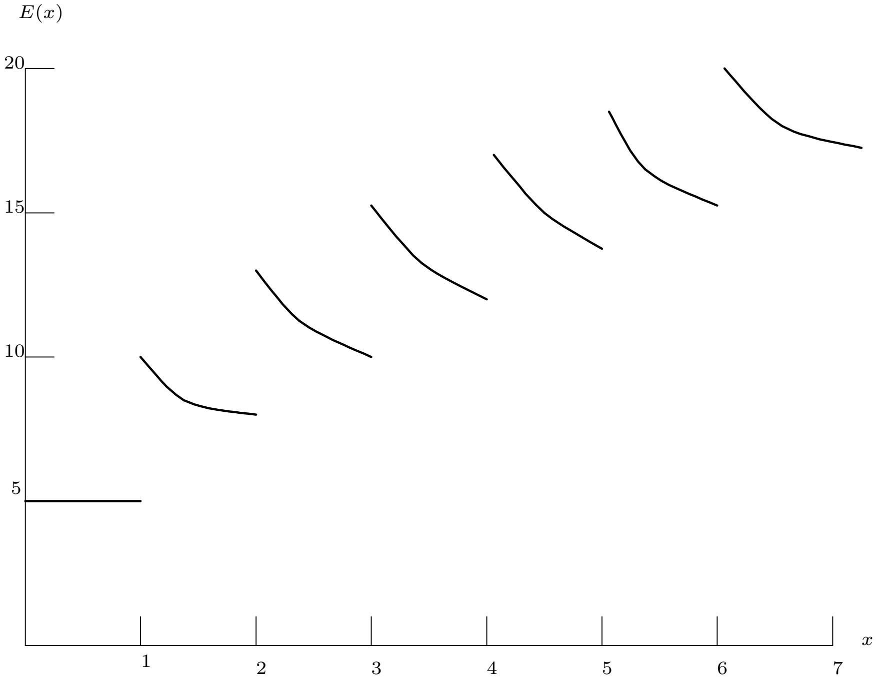

Figure 3: The function $E(x)$ for $\lambda/\mu = 0.8$

position $n+1$) plus the expected waiting time from position n in a system where customers almost surely purchase priority when the low priority queue size is n, i.e., $E(n)$. Finally, the decrease in the function between consecutive integers, $E(n+) - E(n+1)$, is less than or equal to $B - 1/\mu$. This is the case since

$$E(n+) - E(n+1) \leq E(n+) - E(n) - \frac{1}{\mu} = B - \frac{1}{\mu}$$

where the inequality here follows from the forthcoming Proposition 5.1 and the equality is our fourth observation here.

5 Multiple Nash equilibria

We have seen, in Proposition 3.1, that we are in an FTC situation and multiplicity of equilibrium threshold strategies is possible. In this section, we determine an upper bound on the number of such equilibria and demonstrate that it is achieved when θ is large.

Proposition 5.1 *For $n \geq 1$,*

$$E(n+1) - E(n) \geq \frac{1}{\mu} \quad .$$

Hence, the number of pure threshold Nash equilibria is at most $\lfloor 1/(1-\rho) \rfloor$.[5] *Also,*

$$\lim_{n\to\infty} [E(n+1) - E(n)] = \frac{1}{\mu}.$$

Hence, there may exist as many as $\lfloor 1/(1-\rho) \rfloor$ pure Nash equilibria.

Proof. The proof will be based on comparison between two sample paths, one which initializes with state $(0, n)$ and one with $(0, n+1)$ but are otherwise subject to the same (random) events, and where the threshold strategies n and $n+1$ are applied, respectively. Also, a subscript n (resp., $n+1$) corresponds to the former (resp., the latter) system.

Specifically, consider a system which initializes with $(0, n)$ and uses the pure threshold strategy n. Call the last low priority customer in this system at time zero, customer C_n. Notice that when customer C_n clears the system, customer C_{n+1}, is at the head of the queue and at least a service period is ahead of him. In particular, $E(n+1) - E(n) \geq \frac{1}{\mu}$. The upper bound on the number of pure Nash equilibrium threshold strategies is now immediate by Proposition 4.1 and the fact that $B = 1/(\mu(1-\rho))$.

[5] $\lfloor x \rfloor$ denotes the largest integer which is smaller than or equal to x.

Next we prove the second part of the proposition. Let $t_1, t_2, \ldots$ be i.i.d. exponential random variables with mean $1/\lambda$ representing inter-arrival times and let $s_1, s_2, \ldots$ be i.i.d. exponential random variables with mean $1/\mu$ representing service requirements. Let $\tau_n = \sum_{m=1}^{n} t_m$, $n = 1, 2, \ldots$ and let $\sigma_n = \sum_{m=1}^{n} s_m$. Also, let T_n be the service completion time of C_n. Of course, $T_n \geq \sigma_n$ and $\sigma_n/n - 1/\mu = e_n$ where $e_n \to 0$ with probability one. Let A_n be the number of customers who arrive during the time interval $[0, T_n]$. Then,

$$\frac{T_n}{A_n} \geq \frac{\tau_{A_n}}{A_n} = \frac{1}{\lambda} + \nu_n \ ,$$

where $\nu_n \to 0$ with probability one. Let M_n be the number of customers served during the time interval $[0, T_n]$ (a period during which the server was never idle). Then,

$$\frac{T_n}{M_n} = \frac{\sigma_{M_n}}{M_n} = \frac{1}{\mu} + \xi_n \ ,$$

where $\xi_n \to 0$ with probability one. Hence,

$$\begin{aligned} M_n - A_n &\geq T_n \left[\frac{1}{1/\mu + \xi_n} - \frac{1}{1/\lambda + \nu_n} \right] \\ &\geq n(1/\mu + e_n) \left[\frac{1}{1/\mu + \xi_n} - \frac{1}{1/\lambda + \nu_n} \right] \\ &= n(1-\rho)(1+\delta_n), \end{aligned}$$

for some $\delta_n \to 0$ with probability one. In particular, let L_n be the number of customers in the system at time T_n. Then, since $L_n = n - (M_n - A_n)$, we get that $L_n \leq n[\rho(1+\delta_n) - \delta_n]$. Of course, C_{n+1} will conclude service later than C_n by exactly his service requirement unless high priority customer(s) arrive during the period in which he is in service. Let b_n be the number of customers arriving during the service period of C_n and let $r \in (\rho, 1)$. Then,

$$\begin{aligned} P(T_{n+1} > T_n + s_{M_n+1}) &= P(b_{n+1} + L_n > n+1) \\ &\leq P(b_{n+1} \geq n(1-r) \text{ or } L_n \geq nr) \\ &\leq P(b_{n+1} \geq n(1-r)) + P(L_n \geq nr) \\ &\leq P(b_{n+1} \geq n(1-r)) + P(|\delta_n| \geq (r-\rho)/(1-\rho)) \\ &\to 0 \end{aligned}$$

since b_n has a distribution which is not a function of n and $\delta_n \to 0$ in probability. Hence, $T_{n+1} - T_n$ converges in law to s_{M_n+1}, which has an exponential distribution with mean $1/\mu$. Since $T_{n+1} - T_n \geq 0$ is dominated by a busy period, we obtain, by the dominant convergence theorem, that $E(T_{n+1} - T_n) \to 1/\mu$. [6] Finally, by Proposition 4.1 if θ is large enough, there may exist as many as $\lfloor 1/(1-\rho) \rfloor$ pure Nash equilibria. □

[6] The thoerem says that in series of nonnegative random variables which are bounded by an integrable random variable, convergence in law implies convergence in expected value. See e.g, [7], p.289.

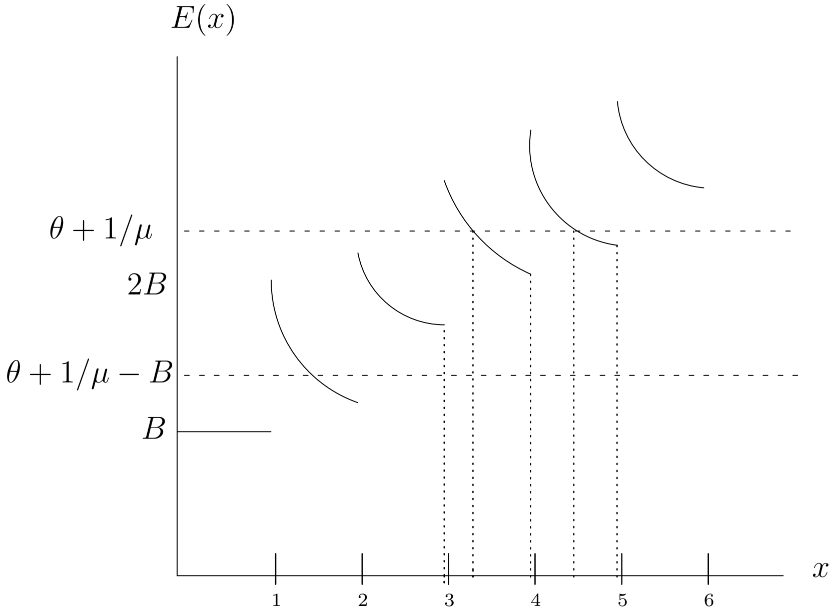

Figure 4: Equilibrium in queues with priorities

Figure 4 illustrates the function $E(x)$. A value of $x = n$ defines a pure equilibrium strategy if $E(n)$ is in $[\theta + (1/\mu) - B, \theta + (1/\mu)]$. A value $x = n + p$, $0 < p < 1$ defines a mixed equilibrium strategy if $E(x)$ equals $\theta + (1/\mu)$. In Figure 4 there are three pure equilibrium threshold strategies at $x = 3, 4, 5$, and two mixed equilibrium threshold strategies, one with $3 < x < 4$ and the other with $4 < x < 5$.

It becomes clear from the figure, that the sequence of equilibrium strategies alternates between pure and mixed strategies. The exception is when θ is such that $E(n) = \theta + (1/\mu)$. In such a case consecutive pure equilibrium thresholds exist. As we have already mentioned, the size of the jump $E(n+) - E(n)$ is exactly B, while the decrease in the function between consecutive integers, $E(n+) - E(n+1)$, is less than B. It comes out therefore that if the equilibrium condition for a mixed strategy at x is satisfied, then both $\lfloor x \rfloor$ and $\lceil x \rceil$ are pure equilibria. In particular, both the smallest and the largest x values that correspond to equilibrium strategies are integers. Moreover, the set of pure equilibria thresholds consists of a set of consecutive integers. These results conform with, and are stronger than, our observations in Section 2 derived from Figure 2.

6 Balking and non-threshold equilibria

Suppose that customers value the service by R and that upon arrival they have the option of balking, that is, of not joining at all.[7] Let $m = \lfloor (R-\theta)\mu \rfloor$. Then, the high priority queue will never be longer than m since an arrival who sees m high priority customers will not purchase priority because in this case his cost $(m+1)/\mu + \theta$ exceeds R. If $m \geq 1$ then under a Nash equilibrium strategy such an arrival will balk. If $m = 0$, then customers will always join the low priority queue as long as its length is smaller than $\lfloor R\mu \rfloor$.

Let B_m be the expected length of a busy period for an $M/M/1/m$ queue, namely a memoryless queue with a buffer of length m. Note that this number includes the one in service. It is easy to prove that for any $\rho > 0$, $B_m = (\sum_{i=0}^{m-1} \rho^i)/\mu$. [8] The analysis for the case with balking follows almost verbatim the analysis for the case without balking. The only difference is that now B_m replaces B.

Corollary 4.3 (with B_m replacing B) has interesting implications. Alperstein [2] proved that in a multi-priority system, a profit maximizer operator will assign a toll such that the thresholds chosen by the customers will be equal to one for all but the highest priority level.[9] Alperstein [2] computed then a set of profit maximizing tolls. However, for two priority levels, for example, these tolls are such that the difference between the high and the low priority tolls are exactly $B_m - 1/\mu$. Therefore, from Corollary 4.3, not only $x = 1$ specifies an equilibrium (as desired) but also every $0 \leq x < 1$ does so. In order to guarantee that the solution at $x = 1$ be unique the toll difference must be increased by some small amount.

We next present an example in which there exists an equilibrium strategy that is not of the threshold type. The example was constructed so that in several cases the arrival is indifferent among options. Any deviation from the given data leads to a unique equilibrium of the threshold type.

Example 6.1 Let $\theta = 3,\ \lambda = \mu = 1$, and $R = 5$. Consider the following strategy: At states $(0,0)$ and $(1,1)$ join the low priority queue, at states $(0,1)$ and $(0,2)$ join the high priority

[7] A low priority customer who sees many high priority customers who joined after his arrival may wish to renege, i.e., leave the system without being served. Following [1], we exclude such a possibility.

[8] Proof: The case $m = 1$ is trivial. Then, an inductive argument coupled with the recursive relationship $(\lambda + \mu)B_m = 1 + \lambda(B_{m-1} + B_m)$ complete the proof.

[9] Alperstein [2] also proved that revenue increases with the number of priority levels. Therefore, when the number of priority levels is not restricted revenue maximization leads to a LIFO regime. She also showed that in this case customers surplus is zero. Therefore, this regime also maximizes social welfare. This gives an alternative proof to a result of Hassin [13].

queue. Otherwise, balk. If this strategy is adopted by all the customers and if the the initial state is $(0,0)$ then only five states will be observed. These are the states $(0,0)$, $(0,1)$, $(0,2)$, $(1,1)$ and $(1,2)$. To prove that the suggested strategy is a Nash equilibrium, we will show that it prescribes for an individual an optimal action for each of these states, given that all the others follow this strategy:

1. State $(0,0)$. If a customer joins the low priority queue his total expected cost is $1/\mu + \lambda/\mu^2 = 2$. The first term is his service time. For the second term, note that λ/μ is the expected number of high priority customers that will arrive (and join the queue) while he is in service. (Note that there is at most one high priority customer in the system, hence our customer will only have to wait for those who arrive during his own service.) Each of them causes a delay of expected value of $1/\mu$. Since 2 is smaller than the service value ($R = 5$) and smaller than the cost associated with high priority purchasing ($\theta + (1/\mu) = 3 + 1 = 4$), joining the low priority is optimal.

2. State $(0,1)$. If the customer buys priority, his expected cost is $1/\mu + \theta = 4$. Thus, balking is not a best response here. If he joins the low priority queue his expected cost is twice the corresponding value for state $(0,0)$, that is, 4. Hence, purchasing priority is an optimal action here (though not the only one).

3. State $(0,2)$. Purchasing priority has an expected cost of $1/\mu + \theta = 4 < R$. Joining the low priority queue leads to an expected cost of 6. Thus, buying priority is optimal.

4. State $(1,1)$. In this case, one is indifferent among three options: Buying priority leads to an expected cost of $\theta + y = 3 + 2 = 5$. Likewise, joining the low priority queue leads to an expected cost of $3/\mu + 2\lambda/\mu^2 = 5$. Thus, joining the low priority queue is optimal.

5. State $(1,2)$. Here one is indifferent between balking and buying priority. These two actions are superior to joining the low priority queue with an expected cost of $4/\mu + 3\lambda/\mu^2 = 7$.

We note that in the above example there are three other possible equilibria. Also, the example can be extended to equilibria where for some integer k the strategy is to join the low priority queue in states $(0,i)$ $i = 0, ..., k-2$, buy priority at $(0, k-1)$, join the low priority queue at $(1, k-1)$, and balk at $(1, k)$.

Acknowledgement. The authors thank Y. Ritov for his very helpful comments.

References

[1] Adiri, I. and U. Yechiali (1974), "Optimal priority purchasing and pricing decisions in nonmonopoly and monopoly queues," *Operations Research*, 22, 1051–1066.

[2] Alperstein, H. (1988), "Optimal pricing policy for the service facility offering a set of priority prices," *Management Science*, 34, 666–671.

[3] Assaf, D. and M. Haviv (1990), "Reneging from processor sharing and random queues," *Mathematics of Operations Research*, 15, 129-138.

[4] Balachandran, K.R., (1972), "Purchasing priorities in queues," *Management Science*, 18, 319–326.

[5] Banerjee, A.V. (1992), "A simple model of herd behavior," *The Quarterly Journal of Economics*, 107, 797-817.

[6] Bikhchandani, S., D. Hirshleifer and I. Welch (1992), "A theory of fads, fashion, custom, and cultural changes as information cascades," *Journal of Political Economy*, 100, 992-1026.

[7] Billingsley, P. (1986) *Probability and Measure, 2nd Ed.*, John Wiley & Sons, New York.

[8] Dewan S. and H. Mendelson (1990), "User delay costs and internal pricing for a service facility", *Management Science* 36, 1502-1517.

[9] Dolan R.J. (1978), "Incentive mechanisms for priority queueing problems", *Bell Journal of Economics* 9, 421–436.

[10] Edelson, N.M. and K. Hildebrand (1975) , "Congestion tolls for Poisson queueing processes", *Econometrica* 43, 81–92.

[11] E.J. Friedman and A.S. Landsberg (1993), "Short run dynamics of multi-class queues", *Operations Research Letters* 14, 221–229.

[12] Glazer, A. and R. Hassin (1986), "Stable priority purchasing in queues," *Operations Research Letters*, 4, 285–288.

[13] Hassin, R. (1985), "On the optimality of first come last served queues", *Econometrica*, 53, 201–202.

[14] Hassin, R. (1995), “Decentralized regulation of a queue”,*Management Science* 41, 163–173.

[15] Hassin, R. and M. Haviv (1994), “Equilibrium strategies and the value of information in a two line queue with threshold jockeying,” *Stochastic Models*, 10, 415-436.

[16] Hassin, R. and M. Haviv (1995), “Equilibrium strategies for impatient customers,”, *Operations Research Letters*, 17, 41–45.

[17] Haviv, M. (1991), “Stable strategies for processor sharing systems,” *European Journal of Operational Research*, 13, 103–106.

[18] Levhari, D. and I. Luski (1978), “Duopoly pricing and waiting lines”, *European Economic Review* 11, 17-35.

[19] Lu F.V. and R.F. Serfozo (1984), “M/M/1 queueing decision processes with monotone hysteretic optimal policies”, *Operations Research* 32, 1116-1132.

[20] Lui, F.T. (1985) “An equilibrium model of bribery”, *Journal of Political Economy* 93, 760-781.

[21] Luski, I. (1976) “On partial equilibrium in a queueing system with two servers”, *The Review of Economic Studies* 43, 519-525.

[22] Mendelson, H. and S. Whang (1990) “Optimal incentive-compatible priority pricing for the $M/M/1$ queue”, *Operations Research* 38, 870-883.

[23] Scharfstein, D.S. and J.C. Stein (1990), “Herd behavior and investment”, *The American Economic Review* 80, 465-479.

[24] Stidham, S. (1992), “Pricing and capacity decisions for a service facility: Stability and multiple local optima”, *Management Science* 38, 1121-1139.

[25] Tilt, B. and K.R. Balachandran (1979), “Stable and superstable customer policies with balking and priority options,” *European Journal of Operational Research*, 3, 485–498.

[26] Viswanathan, N. and T.S. Tse (1989) “Monopolistic provision of congested service with incentive-based allocation of priorities”, *International Economic Review* 30, 153-174.

[27] Weissman, M. (1994), “Upward-sloping segments in the demand curve for priority,” Faculty of Management, Tel Aviv University.

Appendix

We compute $E(x)$ in $O(x^2)$ time for any fixed x, an improvement over the $O(x^3)$ algorithm of [1]. The (future) waiting time of a low priority customer depends on his position in this queue and also on the number of low priority customers behind him. Thus, following [1], let $H_{i,k}(x)$ be the (future) expected waiting time for a low priority customer at position i in his queue (including the one in service), with k low priority customers behind him and when the high priority queue is empty. This value is defined when the threshold x is used by all. Of course, $i \geq 1$, $k \geq 0$ and $i+k \leq \lceil x \rceil$. We will use the H values to compute $E(x) = H_{\lceil x \rceil,0}(x)$.

Assume a threshold $x = n + p$, $0 < p \leq 1$. The maximum possible length of the low priority queue (including the customer in service) is $\lceil x \rceil = n + 1$:[10]

$$\begin{aligned}
H_{1,n}(x) &= B \ , \\
H_{1,n-1}(x) &= 1 + \lambda p H_{1,n}(x) + \lambda(1-p)(B + H_{1,n-1}(x)) \ , \\
H_{1,k}(x) &= 1 + \lambda H_{1,k+1}(x) \quad k = 0, 1, \ldots, n-2 \ , \\
H_{i,k}(x) &= 1 + \lambda H_{i,k+1}(x) + \mu H_{i-1,k}(x) \ , \quad i = 2, \ldots, n-1 \quad k = 0, 1, \ldots, n-i-1 \ , \\
H_{i,n-i}(x) &= 1 + \lambda p H_{i,n-i+1}(x) + \lambda(1-p)(B + H_{i,n-i}(x)) + \mu H_{i-1,n-i}(x) \ , \quad i = 2, \ldots, n \\
H_{i,n-i+1}(x) &= 1 + \lambda(B + H_{i,n-i+1}(x)) + \mu H_{i-1,n-i+1}(x) \ , \quad i = 2, \ldots, n+1 \ .
\end{aligned}$$

These equations lead to a recursive structure.[11] An algorithm directly follows:

Initialization.

```
If $p \neq 0$, let $H_{1,n}(x) = B$;
Let $H_{1,n-1}(x) = (1 + \lambda B)/(1 - \lambda(1-p))$;
For $k = n-2$ until $k = 0$ do:
$H_{1,k}(x) = 1 + \lambda H_{1,k+1}(x)$.
```

The recursion.

```
For $i = 2$ until $i = n$ do:
$H_{i,n-i+1}(x) = (1 + \lambda B + \mu H_{i-1,n-i+1}(x))/\mu$;
$H_{i,n-i}(x) = (1 + \lambda p H_{i,n-i+1}(x) + \mu H_{i-1,n-i}(x) + \lambda(1-p)B)/(1 - \lambda(1-p))$.
If $i \neq n$, for $k = n-i-1$ until $k = 0$ do:
$H_{i,k}(x) = 1 + \lambda H_{i,k+1}(x) + \mu H_{i-1,k}(x)$.

If $p \neq 0$, let $H_{n+1,0}(x) = (1 + \lambda B + \mu H_{n,0}(x))/\mu$.
```

[10] We select the time units so that $\lambda + \mu = 1$.

[11] When with $p = 0$ the first and the last equations are not relevant. This doesn't affect the computational procedure.