# Strategic customer behavior in a single server queue

Moshe Haviv
Department of Statistics
The Hebrew University of Jerusalem
91905 Jerusalem
Israel

July 20, 2009

**Abstract**

The purpose of this chapter is to introduce the reader to the topic of strategic behavior of customers in queues. We assume that customers individually decide on issues such as whether or not to join the queue, whether or not to renege from a queue after waiting for a while, or whether or not to pay some fee in order to belong to a higher priority class. For some examples, we define the appropriate non-cooperative games and look for their Nash equilibria. We restrict ourselves to M/M/1 queues with homogeneous customers (with respect to their cost/reward parameters).

## 1 Introduction

Usually when one considers a queueing model, such as an M/M/1 queue, both arrival and service rates are given. Sometimes an optimization question regarding these parameters is also posed. For example, denote the arrival and service rates by $\lambda$ and $\mu$, respectively. Suppose the service rate of $\mu$ comes with a cost of $c(\mu)$ per unit of time whenever the server is available (and not only while he/she is actually utilized) and it costs $C$ per customer per unit of time in the system. Assume $\lambda$ to be fixed. Then, one looks for the value of $\mu > \lambda$ which minimizes $c(\mu) + \lambda C/(\mu - \lambda)$, the total social cost per unit of time. In this chapter we move one step further: It is the individual customers, rather than a central planner, who decide on their actions. Moreover, they do that by optimizing their selfish needs, while ignoring the effect of their moves on others, known as *externalities.* For example, for a given $\mu$, they may decide whether or not to join the queue. It is only their aggregate behavior which leads to the completion of the model via determining the value of $\lambda$. Commencing with the seminal paper [23], this branch of research got much attention in the literature. A state of the art summary (until 2003) on it is given in [15]. Below we describe some of the main results and ideas in this field while considering a few key examples. We limit ourselves to the M/M/1 model with homogeneous (with respect to their cost and rewards parameters) customers. In fact, in this case strategic behavior is more subtle than in the heterogeneous case.

Customers who consider joining a queue or who are already waiting, may face various types of dilemmas: To join or not to join a queue, to pay or not to pay a fee in order to belong to a high priority class, or whether or not to renege after waiting for a while. We assume that they are selfish and hence maximize their own utility, not minding the externalities that their moves impose on others. While deliberating, they are aware of the fact that other customers may face the same type of dilemma. An individual customer wants to select his/her optimal action, but which action is best depends on what others are doing. For example, if all do not join the queue, one better join as this will come with no need to wait, while if they all do join, one may face too much waiting and might do best by not joining. Such decision making problems are best described as non-cooperative games. We do not attempt to review thoroughly this model. The interested reader is referred to any text in game theory, such as [9] or [22].

In non-cooperative games one's utility for any given strategy, is a function of the strategies selected by others. If one's best strategy is not a function of what others do, one's strategy is called *a dominant strategy.* However, in many cases such strategies do not exist and we look for a *pure Nash equilibrium* strategy profile. By that we mean a set of strategies, one for each individual, such that each one of them is the best for the corresponding individual, given that all others follow their part in the profile. Yet, the problem may persist: there may not be such a profile.[1] Enlarging the strategy

[1] As we see below, there is another possible problem: Multiple pure equilibria profile

space by allowing mixing (namely, players select a lottery or a probability distribution over their set of pure strategies and assess their utility given their and others' behavior by the corresponding expected value, when all lotteries are carried out independently), may guarantee the existence of a Nash equilibrium. Of course, a pure equilibrium is a special case of a mixed equilibrium.

In all models dealt with below, we assume a symmetric game. By that we mean that all customers share the same set of strategies and the same utility function. In particular, customers are homogeneous with respect to their cost/reward parameters. Under a symmetric strategy profile all use the same strategy. Next, for any given such strategy, all assume steady-state conditions under this behavior and all assess their utilities accordingly. Then, we look for a *symmetric Nash equilibrium*, namely an equilibrium profile under which all use the same strategy. More formally, let $S$ be the common set of mixed strategies and let $u(x, y)$ be one's utility if one selects strategy $x \in S$ given that all others select strategy $y \in S$.[2] Thus, $x_e$ is defined as a symmetric Nash equilibrium if

$$x_e \in \arg\max_{y \in S} u(y, x_e).$$

From now on when we refer to an equilibrium, we mean a symmetric Nash equilibrium.

Let $v(x)$ be the social aggregate utility when all use strategy $x$. Typically, $v(x)$ is defined as the sum (or integral) over the individual utilities. A social optimizer looks for an $x^*$ such that

$$x^* \in \arg\max_{x \in S} v(x).$$

Usually, $x_e$ and $x^*$ do not coincide. A mechanism design question is how to manipulate the system so as the resulting new $x_e$ in the manipulated system coincides with the old $x^*$. This can sometimes be achieved by charging prices on various behaviors. In other cases, more complicated schemes are called for.

Our basic model is that of an M/M/1 queue, namely customers arrive in accordance to a Poisson process with rate $\Lambda$ and their service times are

may exist.

[2] By utility here and throughout the rest of the paper we refer to the expected utility when the two lotteries are carried out independently.

independent and exponentially distributed with mean $\mu^{-1}$. Unless stated otherwise, service is granted on a first-come first-served (FCFS) basis. Customers suffer a cost of $C$ per unit of time in the system (inclusive of service) and are rewarded by $R$ due to service completion. In order to avoid trivialities, assume that $R > C/\mu$. Without loss of generality, assume that not joining comes with zero utility.

We are not claiming to supply a comprehensive survey on the field of customers' behavior in queues. Rather, we would like to give the flavor of the issues dealt with, while considering some of the most important queueing decision models. Note that we restrict ourselves to the case where customers are homogeneous with respect to their parameters $C$ and $R$. Many models where this assumption is removed are also surveyed in [15]. We point out that as opposed to strict optimization problems, the homogeneous case is not a special case of the non-homogeneous case. In fact, it is usually harder for analysis as it calls for the use of mixed strategies under equilibrium, while in the fully heterogeneous case (where no two customers share the same parameter values such as $R$ and/or $C$), the equilibrium is usually pure under which all use a parameter dependent pure strategy.

# 2 To queue or not to queue

The most basic question customers ask is whether or not to join the queue. We next deal with two different scenarios. In the first, customers do not observe the queue while taking their decisions and once they join, they cannot change their mind if upon arrival they observe a longer than expected queue. In the second, when they observe the queue length upon their arrival and based on this information they decide whether or not to join. In both cases we also look at the socially optimal behavior and look for ways to regulate the system by levying appropriate tolls so that the resulting equilibrium behavior coincides with the socially optimal behavior. Finally, we discuss the 'avoid the crowd' phenomenon. The results on the unobservable model appear originally in [6] while those on the observable case are from [23]. More comprehensive summaries appear, respectively, in Chapter 3 and Chapter 2 of [15].

## 2.1 The unobservable case

### 2.1.1 Equilibrium behavior

If $\Lambda < \mu$ and all join, then everybody's utility from joining is $C/(\mu - \Lambda) - R$. Thus, if this utility is positive, then joining is a dominant strategy. Indeed, no matter which fraction $p$ of customers join, resulting in a Poisson arrival process with a rate of $\Lambda p$ and with a mean waiting time of $1/(\mu - p\Lambda)$, one's best action is to join. This is not the case where $\Lambda \geq \mu$ or where $\Lambda < \mu$ but $R < C/(\mu - \Lambda)$. Here a dominant strategy does not exist: if all join, one better not join, and if none of them join, one better join since, as assumed, $R > C/\mu$. The equilibrium strategy is to join with probability $p$ for $p$ obeying $R - C/(\mu - p\Lambda) = 0$. Indeed, under such a $p$, $0 < p < 1$, if used by all, one is indifferent between joining and not joining, and hence mixing between joining and not joining with probabilities $p$ and $1 - p$, respectively, is a best response for an individual. Such a $p$ is unique. Denoting it by $p_e$, then

$$p_e = \frac{\mu - \frac{C}{R}}{\Lambda}.$$

Also note that when $p_e < 1$ the utility of all customers, those who join and those who do not, equals zero. No other symmetric equilibrium exists.

### 2.1.2 To avoid or to follow the crowd

Many decision problems stemming from queueing models possess exactly one of the following two phenomena: One is when an individual should not follow the crowd while in the other, one should. Specifically, suppose the set of strategies can be presented by an interval, namely each strategy corresponds to a number in this interval. This is certainly the case when only two pure strategies are available and the probability with which one of the option is selected, describes the mixed strategy completely. In particular, the unit interval can be identified with the set of strategies. Let $u(y, x)$ be as defined in the introduction but now $x$ and $y$ are having a numerical value. Define $y^*(x) \in \arg\max_y u(y, x)$. In other words, $y^*(x)$ is one's *best response* (ties are possible) when all use strategy $x$. We say that the phenomenon of avoid (follow, respectively) the crowd prevails, when $y^*(x)$ is monotone non-increasing (non-decreasing, respectively). We use the acronym ATC and FTC, respectively.

The situation we face in the 'to queue or not to queue' decision problem is that of 'avoid the crowd'. Specifically, suppose all join with probability $p$. Then $u(q,p)$ is one's utility when one joins with probability $q$. Note that

$$u(q,p) = \begin{cases} q\left(R - \frac{C}{\mu - p\Lambda}\right) & p\Lambda < \mu \\ -\infty & p\Lambda \geq \mu \end{cases}$$

It is easy to see that $q^*(p) = 0$ when $p > p_e$ and $q^*(p) = 1$ when $p < p_e$.[3] When $p = p_e < 1$, all strategies are best responses against $p_e$. As it turns out, symmetric equilibria are the points where the best response function $q^*(p)$ intersects the line $q = p$. By definition, in the ATC case, the best response function is not increasing. The function $q = p$ is of course increasing with $p$. Hence, the two continuous functions, $q^*(p)$ and $p$, intersect at exactly one point. In other words, a unique symmetric equilibrium exists in the ATC case. See Figure 1 below. Things are more involved in the FTC cases as we exemplify later.

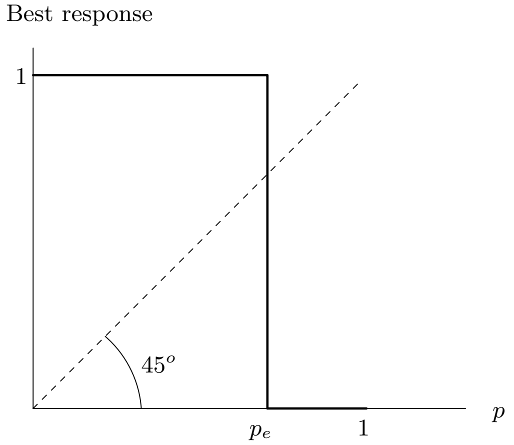

**Figure 1**: Avoid The Crowd

### 2.1.3 Socially optimal behavior

In case of a joining probability of $p$ and assuming that $p\Lambda < \mu$, the social gain per unit of time equals

$$p\Lambda(R - \frac{C}{\mu - p\Lambda}). \tag{1}$$

[3] In the case where $p = 1$ is dominant, $p_e = 1$.

Thus, a social optimizer looks for the value of $p$, $0 \leq p \leq 1$, which maximizes (1). Denote this value by $p_s$. It is easy to see that

$$p_s = \begin{cases} 1 & \Lambda < \mu - \sqrt{\frac{C\mu}{R}} \\ \frac{\mu - \sqrt{\frac{C\mu}{R}}}{\Lambda} & \text{otherwise} \end{cases}$$

In the case where $p_s = 1$, the social gain is as stated in (1) with $p = 1$. When $p_s < 1$, it equals

$$(\sqrt{R\mu} - \sqrt{C})^2. \tag{2}$$

It is also possible to see that unless $p_s = 1$, $p_s < p_e$. In other words, left for themselves customers tend to join with a rate which is higher than is socially desired. This is the case since joining customers ignore the *negative externalities*, in terms of added waiting time to future arrivals, that those of them who join inflict on each other.

In the case where $p_s < p_e$, by artificially reducing $R$ or increasing $C$, one can make the new $p_e$ and the old $p_s$ coincide. It is a simple exercise to see that the amount to reduce $R$ with is $\sqrt{CR/\mu}$, while the amount to increase $C$ with is $\sqrt{RC\mu} - C$. The former is achieved by imposing an entry fee of $\sqrt{CR/\mu}$ while the latter is achieved by charging customers an extra $\sqrt{RC\mu} - C$ per unit of time in the system, making customers suffer in fact a cost of $\sqrt{RC\mu}$ (instead of $C$) per unit of waiting time. The collected fees under both charging schemes go to the society's coffer. In both cases individual customers end up with zero utility. Under this optimal entry toll, the social planner collects $(\sqrt{R\mu} - \sqrt{C})^2$ per unit of time. In other words, all consumer surplus as appears in (2), is transferred completely from the individual customers to the social purse. Such complete transfer of consumer surplus can be achieved when individuals are not more informed than the central planner is.

## 2.2 The observable case

The model dealt with here is as in the previous subsection but with one key distinction: customers observe the queue length upon their arrival and decide accordingly whether or not to join. Thus, a pure strategy in this case means a recipe which for any possible queue length prescribes joining or not joining. Assuming that nobody leaves the system after joining it in the first place, and recalling that the service distribution is memoryless, a sensible symmetric strategy will be of the threshold type. By that we mean that there exists some integer value $n \geq 1$, such that all join if and only if the queue length they observe upon arrival is less than $n$.[4] As a consequence of the FCCS discipline in this observable model, individual's utility is not a function of unknown actions taken (or to be taken) by others, a dominant strategy exists.

### 2.2.1 Equilibrium and socially optimal behavior

First, the threshold equilibrium behavior is trivial: Join if and only if the number in the system is smaller than $n_e$ where[5]

$$n_e = \left\lfloor \frac{R\mu}{C} \right\rfloor . \tag{3}$$

In fact, this strategy is dominant. As opposed to the unobservable case (when $\Lambda$ is large enough), those who join end up with a positive utility. Second, finding the socially optimal threshold $n_s$ is more cumbersome but this was done in paper [23].[6] In particular, $n_s$ is the unique integer value of $n \geq 1$ obeying $g(n-1) \leq \frac{R\mu}{C} \leq g(n)$ where $g(n) = \frac{n(1-\rho)-\rho(1-\rho^n)}{(1-\rho)^2}$. It was shown in [23] that $n_s \leq n_e$ with equality if and only if $n_e = 1$. By definition, under social optimization the consumer surplus is on average larger than it is under the equilibrium behavior.

As in the unobservable case, individual and social behavior will be aligned by imposing an entry fee or by charging a fee per unit time in the system so that the new $n_e$ coincides with the original $n_s$. From (3) we learn that any entry fee $T$ obeying

$$n_s = \left\lfloor \frac{(R+T)\mu}{C} \right\rfloor$$

is suitable for this purpose. In other words,

$$\frac{Cn_s}{\mu} - R \leq T < \frac{C(n_s+1)}{\mu} - R.$$

Under any socially optimal entry fee some of the consumer surplus stays in the hands of the individuals. This is the case since those who join, even after

[4] Yet, things might be a bit more involved. See [14] for a formal treatment for this issue.

[5] In the case where $R\mu/C$ is an integer, both $n_e$ and $n_e - 1$m as as any mixing between them, define an equilibrium profile.

[6] See [15], pp.27-29, for an alternative proof.

paying the entry fee, end up with a positive utility: The shorter the queue they join, the larger is their surplus. This is the case since customers can adjust their behavior to the actual queue length, while the central planner is limited to a single, one for all price. Put differently, customers are more informed than the social planner is. On the other hand, if a queue-length dependent entry fee is imposed, then all of the consumer surplus can be transferred to the common coffer. Indeed, charging (a bit less than) $p_n$ for one who joins the system when $n$ customers are present, with

$$p_n = \begin{cases} R - \frac{C(n+1)}{\mu} & 0 \leq n \leq n_s - 1 \\ \infty & \text{otherwise} \end{cases}$$

makes customers behave socially optimal while leaving no surplus in their hands. For more on this pricing see [4].

Another interesting way to elicit the socially optimal behavior was suggested in [10]. Instead of a FCFS policy, a joining customer is placed at any position but the last one. A possible name for this queueing regime is *not-FCFS*. This allows preempting the one who is in service. In fact, in case where only one customer is in the system, preempting him/her is the only option. Now all join but they have the option to renege (i.e., to abandon) as they observe how many are in front of them. An equilibrium strategy is to renege from the rear of the queue as soon as the total number in the system reaches $n_s + 1$. This leaves the number in the system as it is socially optimal. The reason for this alignment between individual and social behaviors is the fact that the one in the rear of the queue (the one who in fact needs to take decisions) inflicts no externalities on the others under the not-FCFS service policy. Hence, his/her interests and those of the society coincide.

## 2.3 The case of processor sharing or random queues

In the case where service is granted so as the next to commence or to conclude service is decided randomly among all those in line was dealt with in [2] and in [3]. Specifically, a random queue is one where the FCFS discipline is replaced by letting the next to commence service being decided randomly among all those present in line at the epoch of service completion. The processor scheme scheme is such that the service splits in capacity evenly among all those who are present at the system. Yet, in the case of exponential service times, this is equivalent to performing service but deciding to whom it is granted at random among all those present only in its completion. It is clear that one should better not stay in line if it is too long. In [3] it is assumed that the one to renege is the last to arrive (who in fact balks upon arrival). Thus, the question here is in fact to queue or not to queue. As expected, an equilibrium policy prescribes joining if and only if the queue length is smaller than some threshold. Yet, since waiting times are also functions of later to arrive customers, true interaction between customers' policies exists here. Specifically, an ATC prevails here and the unique equilibrium policy can be pure or mixed. In the latter case, mixing is among two consecutive threshold queue lengths.

In [2] it is assumed that as soon as one arrives, he/she is considered as all others. Thus, in case that one needs to leave, all those in line should be treated as equals. It is shown there that in this model no equilibrium exists. Yet, an $\epsilon$-equilibrium for any $\epsilon > 0$ exists.[7] Under this profile, all stay as long as the total number in queue is smaller than some threshold, renege with some rate when at the threshold (meaning reneging after exponentially distributed time if no change in the number is queue took place in the meanwhile), and when the threshold is crossed, to renege at a very high rate (equivalently, in the limit, to performing a lottery which decides who among those in queue should leave immediately). This is a two parameter policy (threshold queue-length and reneging rate). In [2] it is shown how to find it.

# 3 To upgrade or not to upgrade

We go on dealing with the previous model but with two key changes. First, customers have the option of belonging to a premium class. Those who belong to this class have preemptive priority over the others, called regular customers. Among customers belonging to the same class, FCFS is assumed. Yet, belonging to the premium class comes with a cost of $\theta$. No payment is required for those elected to be regular customers. Second, customers do not have the option of not joining. Thus, from the decision problem point of view of paying or not for priority, the value of service is irrelevant. For stability we need to assume that $\Lambda < \mu$. To keep our notation standard with the literature, we replace here $\Lambda$ with $\lambda$.

[7] An $\epsilon$-equilibrium is a strategy profile such that if followed by all, one who deviates from it can gain at most $\epsilon$.

The trade off is between paying for priority and having to wait longer in case of not paying. As all other customers face the same decision problem and as one's waiting time in both classes is a function of the decisions taken by others, the decision model is that of a non-cooperative game. In the next subsection we deal with the unobservable version of the model. Here the strategy space is simple: how to mix between paying and not paying for priority. In the subsequent subsection we deal with the observable version. There, a pure strategy is more involved as for any number of customers from the two groups one may face upon arrival, one needs to decide if to purchase priority or not.[8] In the final subsection of this section we deal again with the unobservable case but now customers can pay as much as they wish while the more one pays, the higher is one's priority level.

## 3.1 The unobservable case

The following appears in [15]. Assume the symmetric mixed strategy of purchasing priority with probability $p$ is used by all. Then, by standard results from priority queues, the mean time in the system of premium customers is $1/(\mu-\lambda p)$ and that of regular customers is $\mu/[(\mu-\lambda)(\mu-\lambda p)]$. Hence, in this case the added value (which can be negative) from belonging to the premium class in comparison with belonging to the regular class is

$$f(p) \equiv \frac{\rho C}{\mu(1-\rho)(1-\lambda p)} - \theta.$$

It is easy to see that $f(p)$ is monotone increasing with $p$. In other words, the more customers purchase priority, the more priority is valuable to an individual. This is an example of the 'follow the crowd' (FTC) phenomenon.[9]

The next question is what are the equilibrium values for $p$. There are three mutually exclusive and exhaustive possibilities:

1. **Case:1** $f(0) \geq 0$**.** Since $f(p)$ is monotone increasing, one better purchase priority no matter how many others do so. In other words, pur-

[8] True, the number of premium customers present in line is irrelevant.

[9] As oppose to 'to queue or not to queue' problem where it is was a-priori clear that the ATC phenomenon prevails, the situation here is more subtle. Indeed, the more purchase priority, the less is the number of regular customers to be overtaken by a premium customer (hinting towards ATC). Yet, the more purchase priority, the more one needs priority for oneself in order not to be overtaken by later to arrive premium customers (hinting towards FTC). The latter effect seems to 'win' among these two antagonistic forces.

chasing priority is a dominant strategy. In particular, $p_e = 1$. See the left graph in Figure 2 for the best response function.

2. **Case 2:** $f(1) \leq 0$**.** Due to a similar argument to the one stated in the previous item, not purchasing priority is a dominant strategy. In particular, $p_e = 0$. See the middle graph in Figure 2 for the best response function.

3. **Case 3:** $f(0) < 0 < f(1)$**.** The first thing to observe is that the two pure strategies define equilibria. Indeed, when all or none purchase priority, it is optimal for an individual to follow suit. But there is a third equilibrium. Let $p_e$ be the unique value such that $f(p_e) = 0$. Of course, such a value exists under the assumption of this case. If all use this strategy, one is indifferent between purchasing priority and not, and hence purchasing priority with probability $p_e$ is one's best response. No other equilibrium exists.[10] The best response function is depicted in the right graph of Figure 2. Specifically, when $p < p_e$ the unique best response is 0. It is 1 when $p > p_e$, and any mixing is a best response against all using $p_e$. Thus, the best response function is non-decreasing. It can hence intersect the 45 degree line more than once. Indeed, this occurs three times: at 0, $p_e$ and 1. In other words, there are three equilibria as stated in the third item above. Indeed, multiple equilibria are the rule in FTC situations.

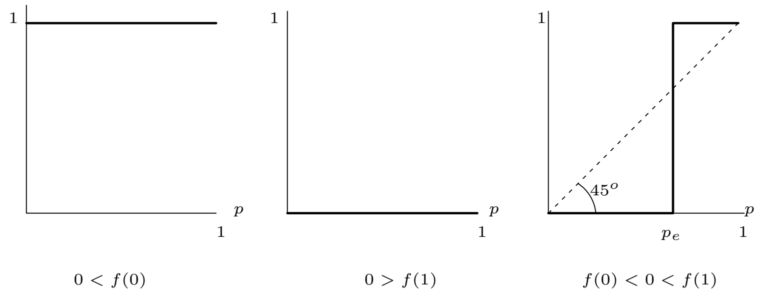

**Figure 2**: Follow The Crowd

[10] The interested reader is referred to [15], p.84, where it is shown that the equilibrium under $p_e$ suffers from lack of stability (defined there) while the two pure equilibria are stable.

## 3.2 The observable case

The following model was originated in [1] where the analysis below appears in [13]. Suppose an arrival to the system observes state $(i, j)$, namely $i \geq 0$ regular customers and $j \geq 0$ premium customers are present. He/she considers whether or not to purchase priority. The value of $j$ does not effect his/her decision: They inflict on him/her a mean queueing time of $j/\mu$, regardless if he/she purchases priority or not. Thus, without loss of generality assume that $j = 0$ and remove further reference to that value. It is clear that the larger $i$ is, the larger is the value of priority. But priority is valuable not only in order to overtake regular customers, but also in order not to be overtaken by future to arrive customers who elected to purchase priority. Thus, one minds here also the purchasing strategies used by others, making this decision model a non-cooperative game.

Suppose all use the pure threshold strategy of $n$, namely they purchase priority if and only if they see $n$ or more regular customers upon arrival. Clearly, when all behave like that, $n$ is the largest possible length of the regular queue. Let $W(n)$ be the mean queueing time (exclusive of service) ahead of a customer who is currently in position $n$ in the regular customer queue where the premium queue is empty. Note that $W(n) + 1/\mu$ is his/her mean total time in the queue. Let $B$ be the mean length of a busy period in an M/M/1 queue. Of course, $B = 1/(\mu - \lambda)$.

**Theorem 1** *The pure threshold strategy $n$, $n \geq 1$, is an equilibrium if and only if*

$$\theta - CB \leq CW(n) \leq \theta. \tag{4}$$

**Proof.** Two conditions need to be met in order for $n$ to prescribe an equilibrium. First, if an arrival faces $n - 1$ regular customers, not purchasing priority, and hence having to be in the system for time whose mean equals $W(n) + 1/\mu$ is better than purchasing priority and being there for a time whose mean is $1/\mu$. For that it is required that $CW(n) \leq \theta$. Second, if an arrival faces $n$ regular customers, purchasing priority and hence having to be in the system for time whose mean is $1/\mu$ is better than not and then having to be for time whose mean is $B + W(n) + 1/\mu$. Note that the term $B$ here is due to the fact that when all adopt the threshold strategy $n$, it requires a busy period to move from position $n + 1$ to position $n$ in the regular queue. For this to be an optimal move, it is required that $\theta \leq C(B + W(n))$. •

In [13] is it shown that there exists at least one value for $n$ which meets inequality (4). As importantly, it is possible that there are a few consecutive values for such $n$, leading to multiple equilibria. As said, this is typical for FTC situations as exemplified already in the unobservable version of this problem. Indeed, the larger is the threshold used by others (namely, the less aggressive they are in terms of purchasing priority), the larger (i.e., less aggressive) should be the optimal threshold used by an individual since the risk of being overtaken is now smaller. This intuitive argument is proved in [11] to be correct. It is also shown there that if both $n$ and $n+1$ prescribe an equilibrium, then there exists another equilibrium, one which mixes between these two thresholds. Under such mixing the mean queueing time of the one in the worst possible position in the regular queue (under this policy) equals $\theta/C$. [11] Finally, in [11] an efficient algorithm for computing the required mean queueing times in developed.

## 3.3 How much to pay for an upgrade

We now look at a different scenario as suggested in [8]. We assume an unobservable M/M/1 queue where Customers are allowed to pay any amount they wish. The more they pay, the higher is their priority level, regardless on how much more they have paid or when they actually have arrived. We assume no preemption and hence the next to commence service among those in line is the one who paid the most. As in the previous subsection, the option of not joining does not exist.[12] Ties in payments are broken on a FCFS basis. Looking for a Nash equilibrium, it is clear that no pure one exists. Indeed, if all pay the same amount, we end up with a FCFS system. Yet, given this behavior, an individual better pay a bit more and overtake all others. By the same token, the mixed equilibrium should be without atoms. Again, had an atom existed, say at $x$, one who pays a little bit more than $x$, would overtake the quantum who pays $x$. Finally, the support of the values paid is a continuous interval. This is the case since had their been a gap, say in $[a, b]$ with values no one selects, it would have been better paying $a$ than a bit more than $b$ since it comes with a clear saving without actually losing any priority value. This contradicts the fact that all pure strategies in the

[11] Yet, as opposed to the pure equilibria, the mixed ones are not stable.

[12] For the case where this option exists and all customers share the same value for service $R$, see [21].

support come with the same gain. In fact, the support should be $[0, u]$ for some value for $u$.

A customer who uniquely gets top priority queues up for time whose mean equals $\rho/\mu$. On the other hand, the corresponding value for a customer who is uniquely most inferior equals $\rho/[\mu(1-\rho)^2]$.[13] Hence,

$$u' \equiv \frac{\rho C}{\mu(1-\rho)^2} - \frac{\rho C}{\mu}$$

is an upper bound on the value of priority. Thus, $u \leq u'$. In fact, we next argue that $u = u'$: The one who pays zero (and is uniquely inferior to all) and the one who pays $u$ (and is uniquely prior to all) share the same utility. Had $u < u'$, it would have been better for the former to pay a bit more than $u$ rather than pay zero. Hence, $u = u'$.

The question left now is what is the distribution over payments on $[0, u]$ defining the equilibrium strategy. Its cumulative distribution function was derived in [8]. Specifically, let $F(x)$ be the probability of paying up to $x$. Then, for any $x \in [0, u]$,

$$F(x) = \frac{1}{\rho}\left[\left(\frac{1}{(1-\rho)^2} - \frac{x\mu}{C\rho}\right)^{-\frac{1}{2}} - (1-\rho)\right].$$

In the above model it is assumed that if one pays more than another, regardless by how much more, one has priority over the other. An alternative model, with a relative priority flavor, is suggested in [19]. Specifically, if upon service completion there are $n$ customers in queue and customer $i$ paid $x_i \geq 0$ towards priority, then customer $i$ is the next to commence service with probability $x_i/\Sigma_{j=1}^n x_j$, $1 \leq i \leq n$. It is shown there that a unique equilibrium exists. Moreover, it is pure and all pay

$$\frac{\rho^2 C}{\mu(1-\rho)(2-\rho)}.$$

# 4 To renege or not to renege

The phenomenon of customers who renege (abandon) the queue some time after joining, is common. The terminology of reneging hints to the possibility

[13] Proof: A busy period comes with a mean value of $1/[\mu(1-\rho)]$. Upon arrival, the expected number of customers in the system is $\rho/(1-\rho)$. The product of these two values is the mean time of what this inferior customer has to spend in the system.

that customers ‘walked away from a commitment’ or that they ‘change their mind’. Standard modeling does not allow for this possibility. Instead, we assume that customers join the queue while planning to leave it (abandon) later in case that something occurs or some relevant information is revealed to them.

Before giving further details on the model we pose, we make the following observation. In an unobservable M/M/1 FCFS queue, reneging is not due to learning while waiting that the queue length is in fact longer than anticipated. It was shown in [11], that no matter what is the reneging policy employed by the others, for an individual the longer is his/her past waiting time, the shorter is his/her expected time ahead. Thus, if one did not renege until now, it will be inconsistent from one’s side to do it now (or later). Thus, reneging needs to be attributed to increasing waiting costs per unit of time (which can be interpreted as loss of patience) or due to a deterioration in the value of service. In fact, these two factors are two sides of the same coin: The cost/reward ratio should go up while waiting in order to observe reneging.

We next describe two such situations in the unobservable M/M/1 FCFS queue. In the first, when the cost/reward is constant for some time and then it jumps abruptly so as to enforce immediate reneging. Interestingly, no reneging takes place ahead of this time. Yet, one may elect not to join at all. In the second, we let the cost/reward ratio go up smoothly, leading to a mixed equilibrium strategy regarding when to abandon the queue. Thus, reneging can take place at some continuous interval of time after waiting.

## 4.1 The bang bang case

The model we deal with here is as our standard unobservable M/M/1 FCFS model with one key distinction: For anyone who stayed already $T$ units of time in the system without completing service (regardless if by then it was commenced or not), the value of service drops to zero. Assuming that customers recall the length of their past queueing time (i.e., each one of them holds a watch), a pure strategy here tells whether or not to join, and in the latter case, if and when to renege. As always, mixing is possible. In [11] it is shown that under any behavior of the others, if a best response prescribes joining (not necessarily uniquely), then reneging after waiting for less than $T$ units of time, is never part of the best response. This is due to the fact that the probability of completing service in the next instant of time, goes up with the length of the past waiting time. Thus, in our quest for an equilibrium strategy we need to look for an equilibrium joining probability, while it is clear that one who joins reneges after $T$ units of time if service is not completed by then. In summary, the problem we face here is of finding the equilibrium joining probability.

Let $f(p)$ be the expected utility of one who joins and plans to leave after $T$ units of time if one's service does not end by then, given that all others do the same with probability $p$ (and do not join at all with probability $1-p$). In [11] it is shown that

$$f(p) = \frac{R(\mu - \mu e^{-(\mu-\lambda p)T}) + \frac{\lambda p C T(\mu-\lambda p)e^{-(\mu-\lambda p)T} - C\mu(1-e^{-(\mu-\lambda p)T})}{\mu-\lambda p}}{\mu - \lambda p e^{-(\mu-\lambda p)T}}.$$

Since $f(p)$ is monotone decreasing with $p$, if $f(1) \geq 0$, then joining (and then reneging at $T$) is a dominant strategy. Otherwise, there exists a unique value for $p$, $0 < p < 1$, with $f(p) = 0$. Denote this value by $p_e$. Then joining with probability $p_e$ (and in case of joining, reneging after $T$ units of time) is the unique equilibrium strategy.

## 4.2 Smooth deterioration in waiting conditions

We now relax the assumption that the cost of waiting per unit of time is constant. So assume that $c(t)$ is the cost of waiting per unit of time for one who waits already $t$ units of time. In other words, one who has already waited for $t$ units of time suffered so far a loss of $\int_{\tau=0}^{t} c(\tau)\,d\tau$. The situation here is of course more complicated then the one described in the previous subsection. We describe one possibility as reported in [18].

**Theorem 2** *Assume the following:*

1. $c(0) = 0$ *and* $c(t)$ *is monotone increasing and concave. Also,* $\lim_{t\to\infty} c(t) > R\mu$*. Let* $T_2$ *be with* $c(T_2) = \mu R$*.*[14]

2. $\beta(t) \equiv c(t)/R - c'(t)/c(t)$ *is monotone increasing.*

3. *Let* $T_1$ *be with* $\beta(T_1) = \mu - \lambda$ *and assume that* $T_1 < T_2$*.*

[14] It is clear that all renege not latter than $T_2$ units of time after joining.

*Define the cumulative distribution function $G(t)$ by*[15]

$$G(t) = \begin{cases} 0 & 0 \leq t \leq T_1 \\ 1 - \frac{\mu - \beta(t)}{\lambda} & T_1 \leq t < T_2 \\ 1 & t \geq T_2 \end{cases}$$

*Then, the unique equilibrium strategy is to join and then to renege after time which follows the distribution defined by the function $G(t)$.*

In [18] it is also shown that if a pure equilibrium reneging strategy exists (of course, under different assumptions then those postulated Theorem 2), it prescribes reneging at time $T_2$. Sufficient conditions for this to be the case are stated there.

# 5 Concluding remarks

A number of decision problems faced by customers in a memoryless single server queue were surveyed here as non-cooperative games. This survey is by no means exhaustive. Other problems were dealt with in the literature. For example, in an M/M/1 when customers never learn the queue situation, they can try again and again, hoping to find the server available. These retrials comes with a cost. The question is then when to retry. The case where customers do not recall their past experience and do not even possess a watch, leading to identically, exponentially and independently distributed intervals between consecutive retrials, was solved in [12]. The cases of full or partial information regarding their own whereabouts (such as when they have retried last) are still open. Another question is when to arrive (given a finite period and a distribution on the number of arrivals during that period). See [7] for further details.

The assumption of stationarity, or steady-state conditions assumed throughout this survey, is removed in [16]. Here customers know their time of arrival and need to decide if to try to get service or to leave for good upon arrival to an M/M/N/N queue (when trials are costly and hence finding all servers busy after trying ends up with a loss). Assuming an empty queue at time $t = 0$, equilibrium strategy prescribes trying if the time of arrival is smaller than some threshold value $t_e$. Afterwards, one tries with some time-dependent

[15] Note that since $c(T_2) = \mu R$, there is an atom of size $c'(T_2)/(R\lambda\mu)$ at $T_2$.

probability $p_e(t)$. Interestingly, $p_e(t_e+) < 1$. Treating similar models, for example the simplest M/M/1 to queue or not to queue model dealt with throughout this survey, is still open. Clearly it is a technically challenging problem.

For more problems dealing also with general service distributions, heterogeneous customers and with multi-server queues see [15]. In particular, see [17] and [20] for the observable version of the 'to queue or not to queue' problem in the M/G/1 case.[16]

Finally, many models where *informational externalities* are involved can be designed. One example is dealt with in [24]. There, a customer who joins one out of two queues sends, just by his/her own action, a message to future arrivals that this is the better queue to join according to his/her belief. Hence, it sometimes makes sense to join the longer queue. Another example appears in [5]. Here there is only one server whose utility is not known. Customers get private signals and after inspecting the queue length, decide if they prefer joining over balking. Here too anyone who joins may affect the posterior regarding the value of service for future to arrive customers. Again, longer queues can sometimes be more appealing than short one.

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

[3] Altman, E. and N. Shimkin (1998), "Individual equilibrium and learning in processor sharing system," *Operations Research*, 46, 776-784.

[4] Chen, F. and M. Frank (2001), "State dependent pricing with a queue," *IIE Transactions*, 33, 847-860.

[5] Debo, L.G., C.A. Parlour and U. Rajan (2008), "Inferring quality from a queue," Chicago Booth School of Business Working Paper 09-26. Also

[16] The unobservable version is a straightforward generalization of what was presented here in Section 2.1.

in http://faculty.chicagobooth.edu/laurens.debo
/Research/Papers/debo_herding_single_queue/pdf.

[6] Edelson, N.M. and K. Hildebrand (1975), “Congestion tolls for Poisson queueing processes,” *Econometrica*, 43, 81-92.

[7] Glazer, A. and R. Hassin (1983), “?/M/1: On the equilibrium distribution of customer arrival,” *European Journal of Operational Research*, 13, 146-150.

[8] Glazer, A. and R. Hassin (1986), “Stable priority purchasing in queues,” *Operations Research Letters*, 4, 285-288.

[9] Fudenberg, D. and J.Tirole (1991), *Game Theory*, The MIT Press, Cambridge, MA.

[10] Hassin, R. (1985), “On the optimality of first come last served queues,” *Econometrica*, 53, 201-202.

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

[12] Hassin, R. and M. Haviv (1996), “Optimal and equilibrium retrial rates in a busy system,” *Probability in the Engineering and Informational Science,* 10, 223-227.

[13] Hassin, R. and M. Haviv (1997), “Equilibrium threshold strategies: The case of queues with priorities,” *Operations Research*, 45, 966-973.

[14] Hassin, R. and M. Haviv (2002), “Nash equilibrium and subgame perfection,” *Annals of Operations Research*, 113, 15-26.

[15] Hassin, R. and M. Haviv (2003), *To queue or not queue: Equilibrium behavior in question systems*, Kluwer’s International Series, Boston. Also in http://www.math.tau.ac.il/∼hassin/book.html.

[16] Haviv, M., O. Kella and Y. Kerner (2009), “Equilibrium strategies based on time or index of arrival,” *Probability in the Engineering and the Informational Sciences*, (to appear).

[17] Haviv, M. and Y. Kerner (2007), “On balking from an empty queue,” *Queueing Systems: Theory and Applications*, 55, 239-249.

[18] Haviv, M. and Y. Ritov (2001), “Homogeneous customers renege from invisible queues after random times under deteriorating waiting conditions,” *Queueing Systems: Theory and Applications*, 38, 495-508.

[19] Haviv, M. and J. van der Wal (1997), “Equilibrium strategies for processor sharing and queues with relative priorities,” *Probability in the Engineering and the Informational Sciences*, 11, 403-412.

[20] Kerner, Y. (2009), “Equilibrium joining probabilities for an M/G/1 queue,” EURANDOM Working Paper 2009-020.

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

[22] Osborne, M. and A. Rubinstein (1994), *A Course in Game Theory*, The MIT Press, Cambridge, MA.

[23] Naor, P. (1969), “The regulation of queue size by levying tolls,” *Econometrica*, 37, 15-24.

[24] Veeraraghavan, S. and L.G. Debo (2009), “Joining longer queues: informational externalities in queue choice,” *Manufacturing & Service Operations Management*, (to appear).