A model for rational abandonments from invisible queues

Avishai Mandelbaum [a] and Nahum Shimkin [b]

[a] *Department of Industrial Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel*
E-mail: avim@tx.technion.ac.il

[b] *Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel*
E-mail: shimkin@ee.technion.ac.il

Received 3 August 1999; revised 1 December 1999

We propose a model for abandonments from a queue, due to excessive wait, assuming that waiting customers act rationally but without being able to observe the queue length. Customers are allowed to be heterogeneous in their preferences and consequent behavior. Our goal is to characterize customers' patience via more basic primitives, specifically waiting costs and service benefits: these two are optimally balanced by waiting customers, based on their individual cost parameters and anticipated waiting time. The waiting time distribution and patience profile then emerge as an equilibrium point of the system. The problem formulation is motivated by teleservices, prevalently telephone- and Internet-based. In such services, customers and servers are remote and queues are typically associated with the servers, hence queues are invisible to waiting customers. Our base model is the M/M/m queue, where it is shown that a unique equilibrium exists, in which rational abandonments can occur only upon arrival (zero or infinite patience for each customer). As such a behavior fails to capture the essence of abandonments, the base model is modified to account for unusual congestion or failure conditions. This indeed facilitates abandonments in finite time, leading to a nontrivial, customer dependent patience profile. Our analysis shows, quite surprisingly, that the equilibrium is unique in this case as well, and amenable to explicit calculation.

Keywords: multiserver exponential queues, abandonments, Nash equilibrium, call centers

1. Introduction

The problem of customer abandonments from a queue, due to excessive waiting times, is of considerable importance and concern in various applications. Traditional queueing theory has dealt successfully with the analysis of queues under the assumption of a given patience distribution (*patience* is the time a customer is willing to wait in queue). It is, however, also of obvious importance to consider the factors which affect this distribution, such as individual preferences and system performance. In this paper we take a decision-theoretic viewpoint towards understanding the abandonment phenomena: the abandonment time for each customer is based on an individual utility optimization, which balances perceived waiting costs against the benefits of service, and from which the patience distribution emerges as an equilibrium point.

1.1. Background and motivation

On the application side, our study is motivated by the fast-expanding area of teleservices, which prominently include telephone call centers and the emerging Internet-based market. Our model assumptions, therefore, are geared towards such systems where customers and service providers are remote from each other. There is little need to elaborate here on the significance of Internet-based services. As for call centers, these currently constitute a multibillion dollar industry which is rapidly expanding. (Some estimate the 1998 yearly revenues of the U.S. market alone at about $5 billion, growing at a rate of over 27% annually.)

Customers of call centers increasingly demand quick and efficient service, otherwise abandonments of waiting customers become prevalent and of major concern. Indeed, AT&T studies [3] indicate that a 15 s wait to an operator response caused 44% of the callers to abandon the call; for a 30 s wait that figure increased to 69%. The Help Desk Institute, in its annual report [11], specifies that about 43% of call centers have a target for the abandon-rate, and about 40% of the call centers experience call abandon-rates over 10%. It should be observed that in toll-free services such as 1–800, holding times of customers (including ones that eventually abandon) are paid by service providers. With the explosive growth of toll-free services, these costs have become a major economic driver. Abandonments may also have a significant effect on system's performance [7], leading to an improved service level for the remaining customers. With these observations, it is clear that the phenomena of abandonments must play a central role in any definition of teleservice quality and call-center efficiency, hence, it should be well understood and quantified.

1.2. Assumptions and results

As we wait in queue for service, our willingness to wait further may well be influenced by our assessments concerning the remaining time to service. This effect is explicitly captured in our model, through a cost function that weights anticipated waiting costs against service utility. A basic ingredient of this model is customers' expectations regarding their waiting times, which each customer summarizes as a distribution function. We shall employ here a consistency assumption (section 2.3), namely that these expectations, formed for example through experience, coincide with the actual waiting time distribution in the queue. Since the latter depends, in turn, on customer abandonment decisions, the system behavior then emerges as a Nash equilibrium point, namely, a fixed point of the map induced by the individual decision model and consistency assumption.

Another factor that may have considerable influence on customer patience is the on-line information available regarding the current system state or position in queue; see, e.g., [12]. In the present paper we assume that such information is not available to the customer, which is a realistic assumption in current call center applications. Increasingly though, state information is purposely provided by call centers, and the integration of such information into our model is an important topic for further research.

Our decision model assumes service utilities that are time-invariant and waiting costs that are linear in waiting times; these parameters may vary, however, across individual customers. As a base model, we consider the $\mathrm{M}/\mathrm{M}/\mathrm{m}$ queue. We show (theorem 7) that in this case the rational (individually optimal) decision for each customer is either to abandon the queue immediately upon arrival, or else to stay in the queue until served. Such a simple behavior is a consequence of the property that the hazard rate function for the virtual waiting time in the queue is increasing (IHR), for any $\mathrm{M}/\mathrm{M}/\mathrm{m}$ queue with general abandonments ($\mathrm{M}/\mathrm{M}/\mathrm{m}+G$ in the notation of [2]). While this leads to a complete and relatively easy characterization of a unique equilibrium, it is obviously quite unsatisfactory from a descriptive point of view, since finite abandonment times prevail in practice.

Several options are available to address this deficiency, as elaborated in section 6. Here we focus on the IHR issue. In reality, customers who are left waiting for a long time are expected to start loosing confidence and, if anything, will assess their likelihood of obtaining service in the near future as declining or even diminishing. To accommodate such a tendency, we consider an extended model which includes a fault option. According to this model, denoted $\mathrm{M}/\mathrm{M}/\mathrm{m}(q)$, each arriving customer joins the regular queue with probability q, but with probability $(1-q)$ will be placed at a fault position where service will never be provided, without being notified of this situation. The modified model can be considered both as addressing individual faults, where indeed individual customers are occasionally 'forgotten' by the system; or system-scale faults, where occasionally the system is malfunctioning and all arriving customers are subject to slow service. This model also provides a proxy for other causes of congestion which are not captured by the standard $\mathrm{M}/\mathrm{M}/\mathrm{m}$ queue, such as varying number of servers, time-dependent arrival rates, service priorities, etc.

It turns out (proposition 4) that the $\mathrm{M}/\mathrm{M}/\mathrm{m}(q)$ system has an eventually-decreasing (and, in fact, unimodal) hazard rate function, which makes finite abandonment times feasible as rational choices (proposition 2). Naturally, this additional option both enriches the space of potential equilibria and complicates the analysis. Still, by exploiting the very special structure of the $\mathrm{M}/\mathrm{M}/\mathrm{m}+G$ queue and some explicit expressions for its performance (section 3.2), it will be established that the $\mathrm{M}/\mathrm{M}/\mathrm{m}(q)$ model gives rise to a unique equilibrium point. Formulas which allow to compute the equilibrium distribution of the abandonment times are also obtained (theorem 8).

Regarding the latter uniqueness result, it should be mentioned that Nash equilibrium solutions are typically non-unique in an essential way, and multiple unconnected equilibria may exist in general. In view of the heterogeneity in user behavior, range of possible decisions and the complexity of a stochastic model, it is hardly apparent that the equilibrium should be unique in the present case. Some general methods have been suggested in the literature to establish uniqueness of the Nash equilibrium in nonzero sum games, exploiting such properties as convexity [18] and contraction [15]; however, none of these has been found applicable to our problem. As it stands now,

the uniqueness result rests on specific and explicit analysis, which in turn relies on the special structure of the $\mathrm{M}/\mathrm{M}/\mathrm{m}+G$ queue.

1.3. Related research

Concerning previous literature, most related to the present study is the work by Hassin and Haviv [9]. This paper considers a similar rational model in an $\mathrm{M}/\mathrm{M}/1$ queue, but assuming that all customers have an identical cost function. It is further assumed that the service utility vanishes once service is not completed within a fixed time beyond arrival, that abandonments are possible during service as well, and that customers may decide to reneg (not join the queue at all). A unique equilibrium is shown to exist in which each customer joins the queue with a fixed probability, and then stays until his service time expires. While differences in details exist, this result is also a consequence of the IHR property of the relevant queue and is closely related to the findings regarding our base $\mathrm{M}/\mathrm{M}/\mathrm{m}$ model. A recent paper of Haviv and Ritov [10] considers again the homogeneous customers case, but under a convex waiting cost, and shows under certain conditions the existence of a unique equilibrium which induces a continuous distribution of abandonment times.

A different temporal equilibrium problem is treated in [6,20], where motorists optimize their arrival time at a congested bottleneck road, and a deterministic fluid traffic model is used. Additional work on individual equilibrium in queues includes [1,5,8,16,19].

It is both highly relevant and of historical interest to mention the classical work by Palm [17], who develops methods for estimating the *inconvenience* experienced by customers due to delayed telephone connection. Palm proposed a simple parametric model for the inconvenience, as a function of experienced waiting time, and proceeded to estimate its parameters by linking inconvenience to the abandonment rate and measuring the latter. The link is provided by an $\mathrm{M}/\mathrm{M}/\mathrm{m}+G$ model, after postulating that the hazard rate of customers' patience is directly proportional to the marginal inconvenience (*irritation* in terms of [17]). It is interesting to note that the empirical data used in [17] were collected in certain exchanges at the Stockholm area, where "relatively often, ... through errors in dialing, ... (subscribers) would not receive any ringing tone, so that they were presented with a delay time of unlimited duration." An $\mathrm{M}/\mathrm{M}/\mathrm{m}(q)$ system indeed!

1.4. Contents

Our paper is structured as follows. The next section presents the model description, including the individual decision model and the definition of equilibrium. Section 3 develops some preliminary results concerning rational decisions. In particular, we explore the relation that exists between these decisions and the monotonicity properties of the service hazard rate function, we recall the waiting time distribution for the $\mathrm{M}/\mathrm{M}/\mathrm{m}$ queue with general patience distribution G, and establish the monotonicity properties of the hazard rate function in this and the $\mathrm{M}/\mathrm{M}/\mathrm{m}(q)$ model. Section 4

contains the main results regarding the uniqueness and structure of the equilibrium, while the proofs of the relevant results for the M/M/m(q) model are deferred to section 5. Finally, sections 6 and 7 offer some concluding remarks, with a discussion of modeling choices and possible extensions.

2. Model formulation

This section presents the rational equilibrium model that is the subject of this paper. We start by briefly introducing the queueing system, followed by a definition of the individual decision model and the utility function employed by each customer. We then consider the system as a whole and discuss the equilibrium concept that results by reconciling customer expectations with actual system performance.

2.1. The basic queue

Our base model is the M/M/m queue, with Poisson arrivals at rate λ, i.i.d. exponentially distributed service times with expected duration $1/\mu$, and m servers that cater to customers in order of arrival (FCFS). The queue capacity is assumed infinite. We shall also consider an extension of this model, where each arriving customer enters the main queue with probability q, but has a probability $(1-q)$ of being placed in a fault position and never obtaining service. This model will be denoted by M/M/m(q), with $0 < q < 1$, and will be further elaborated on in section 3.2. We assume throughout that the queue is in steady state.

During their waiting period in the queue, customers may decide to abandon the queue and give up the offered service. Abandonments do not occur after service commences. Abandonment times are chosen individually by each customer, based on a decision model which we now specify.

2.2. Individual utility and rational decisions

After joining the system, a customer may abandon the queue at any time $T \geqslant 0$ before admitted to service. It is assumed that no information is conveyed to customers during this period regarding the status of the queue or their standing in it. Thus, an abandonment policy for each customer is simply the time T she is willing to wait in the queue before abandoning it. (See section 6 for some comments on the equivalent sequential, or "real-time", formulation of abandonment choices.)

Observe that a decision to abandon at $T = 0$ is different than not joining the system at all, since in the former case the customer enjoys the opportunity of obtaining service immediately upon arrival. Such a decision corresponds to the widely observed phenomenon of customers who abandon immediately upon recognizing a delay.

We now define an individual utility function for the customers over their set of choices. We consider a heterogeneous customer population, and customers will be categorized into different types according to their decision model parameters. Let $z \in Z$ denote the type, with Z the set of possible types.

A customer of type z will be characterized by the following elements:

(i) r_z, the service utility, assumed to be positive.

(ii) c_z, the marginal cost of waiting, or simply the cost coefficient, also assumed positive. The waiting cost (or disutility) is assumed linear in the waiting time, and given by $c_z w$, where w is the time until the customer abandons or is admitted to service.

(iii) $F_z(\cdot)$, a probability distribution function which reflects the customer's belief about her virtual waiting time V, namely, the time from her arrival until she enters service, provided that she does not abandon the queue. Denote $\overline{F}_z = 1 - F_z$.

Observe that F_z as used here is a subjective quantity, which is required in order to define the customer's expected utility. (We shall later impose the *consistency condition* that the subjective distributions F_z all coincide with the distribution function of the true virtual waiting time.)

Define the cost-benefit ratio $\gamma_z := c_z/r_z$. This parameter will play a central role in our analysis.

Consider a customer that decides to abandon the queue after $T \geqslant 0$ time units if not admitted to service by then. The actual waiting time will be $W = \min\{V, T\}$, where abandonment occurs if $T < V$, and otherwise the customer enters service. The expected utility for such a customer will be

$$\begin{aligned} U_z(T) &= E_z\big(r_z \mathbb{1}\{T \geqslant V\} - c_z \min\{V, T\}\big) \\ &= \int_{0-}^{T} [r_z - c_z v]\, \mathrm{d}F_z(v) - c_z T \overline{F}_z(T), \end{aligned} \tag{1}$$

where E_z stands for the expectation with respect to the subjective probability F_z. Note that F_z conceivably includes a point mass at $T = 0$, representing the probability of finding a free server immediately upon arrival, and the integral is taken to include this point; thus, $U_z(0) = r_z F_z(0)$. Observe also that we do not explicitly account for the expected time-in-service in this utility function; however, this may be easily incorporated in the service utility r_z.

An optimal decision for a type-z customer is a time $T_z \geqslant 0$ that maximizes the expected utility U_z. For concreteness, in case that the utility function attains it maximum in more than one point we shall choose the *later* time. (Any other choice may be made without affecting the results; indeed, in equilibrium it will turn out that non-unique optimal choices may occur only for one specific customer type, which has a zero measure according to the regularity assumption imposed below.) This definition fixes T_z as a deterministic quantity for each customer of a given type z.

As already observed, with our utility function it is assumed implicitly that waiting customers do not obtain information regarding the current state of the queue or their standing in it. This justifies the convenient viewpoint that abandonment times are chosen once upon arrival to the queue.

To complete the system description, we require an additional quantity:

(iv) P_Z, a probability distribution over the set of customer types Z. The type z of each customer is randomly chosen according to P_Z, independently across customers. We shall assume for simplicity that the cost–benefit ratio γ_z, considered as a random variable with distribution induced by P_Z, has a density on the (positive) real line.

Some comments regarding the customer type and its associated distribution P_Z are in order here. As defined, the type variable z parameterizes the distribution F_z, along with the cost coefficients r_z and c_z. However, under the consistent equilibrium condition considered in the sequel all distributions F_z must coincide with the actual one, so that customer types differ only in the coefficients c_z and r_z. Thus, P_Z can then be interpreted as a probability distribution over these coefficients. Moreover, it will be seen that the maximizer of the utility function U_z depends only on the cost–benefit ratio $\gamma_z = c_z/r_z$, so that a customer type may be identified with this ratio. The assumption that γ_z has a density under P_Z is not crucial, but quite conveniently alleviates the need to consider randomized decisions, which would otherwise be essential for the existence of an equilibrium (see [9]). Apart from that, the distribution of γ_z is general.

Suppose now that we are given the type distribution P_Z, as well as the customer parameters c_z, r_z and F_z for each type $z \in Z$. Assuming that customers behave according to the decision model described above, and that the optimal decisions T_z are well-defined, this induces a distribution on the abandonment times of each customer, namely, a patience distribution G, which is i.i.d. across customers. The model is then completely specified as an $\mathrm{M}/\mathrm{M}/\mathrm{m} + G$ queue, and its performance can be analyzed using, e.g., the results of [2].

Our point of departure from $\mathrm{M}/\mathrm{M}/\mathrm{m} + G$ scenario concerns the assumption that F_z is given a priori, without regard to actual system performance. This will be replaced by a consistency assumption, which we consider next.

2.3. *Equilibrium*

As noted, given a patience distribution G, one can compute the system statistics and, in particular, the "true" (or *objective*) distribution of the virtual waiting time, denoted F. Our basic assumption here is the consistency requirement, that the subjective distributions held by all customers coincide with the true one, namely $F_z = F$ for all $z \in Z$. This leads to the following definition of system equilibrium, which is just the Nash equilibrium under the consistency assumption:

Definition 1. The system is in a *consistent equilibrium* (or just equilibrium) if the following hold:

(i) Individual rationality: Each customer of type z is using an individually optimal abandonment time T_z, as defined above; recall that this choice is based on a utility function which involves the subjective virtual waiting time distribution F_z.

(ii) Consistency: The subjective and objective virtual waiting time distributions coincide: $F_z = F$, for every customer type z.

We then refer to the set $\{T_z,\ z \in Z\}$ as an *equilibrium profile*, and to F as the *equilibrium distribution*.

The consistency requirement implies that customers have complete knowledge regarding the statistics of the waiting time in the system. In practice, such knowledge may grow out of previous visits to the system.

It is evident that the definition of equilibrium is not explicit, but rather specifies F as a fixed point of an appropriate map, which may be summarized in the following two steps:

- $F \to G$. Given F, the consistency assumption $F_z \equiv F$ together with other customer utility and type characteristics (c_z, r_z and P_Z) determine the patience distribution G.
- $G \to F$. Given the patience distribution G, the virtual waiting time distribution F is determined through the queue dynamics.

Note that this map is on a space of probability distribution functions, so that we obtain a *functional* fixed point condition. Another option is to consider the fixed point of the map between *decision profiles* $\{T_z\}$, as follows:

- $\{T_z\} \to F$. $\{T_z\}$ together with P_Z determine the patience distribution G, which in turn defines the queue statistics and the virtual waiting time distribution F.
- $F \to \{T_z\}$. The consistency assumption $F_z \equiv F$, together with the utility parameters c_z and r_z, determine the optimal individual decisions $\{T_z\}$.

Since the support of Z is in general of infinite cardinality, decision profiles belong to an infinite-dimensional function space, and again we obtain a functional fixed-point condition. We shall find this formulation more convenient for analysis than the previous one.

The prominent questions regarding the equilibrium point include existence, uniqueness, structural properties, and computation. These are all addressed in the sequel.

3. Preliminary analysis

3.1. Individual optimization and the hazard rate

In this subsection we examine some properties of the optimal abandonment times and, in particular, their relation with the hazard rate function H_z associated with the virtual waiting time distribution F_z. We consider here a fixed customer type z, with a *given* subjective distribution F_z. We will show that monotonicity properties of the hazard rate function lead to interesting structural properties of the optimal abandonment time, which will be instrumental in the equilibrium analysis to follow.

Assume throughout that $F_z(t)$ is continuously differentiable for $t > 0$ (that is, it has a continuous density F_z', except possibly for a point mass at $t = 0$), and that F_z'

has a right limit at 0. This smoothness property is indeed enjoyed by all distribution functions that arise in later sections. Differentiating the utility function (1) with respect to $T > 0$ gives

$$U_z'(T) = [r_z - c_z T]F_z'(T) - c_z\overline{F}_z(T) + c_z T F_z'(T) = r_z F_z'(T) - c_z\overline{F}_z(T). \quad (2)$$

Since $r_z > 0$ by assumption, when $\overline{F}_z(T) > 0$ this may be written in the following way:

$$U_z'(T) = r_z\overline{F}_z(T)\left[\frac{F_z'(T)}{\overline{F}_z(T)} - \gamma_z\right] = r_z\overline{F}_z(T)\big[H_z(T) - \gamma_z\big], \quad (3)$$

where $\gamma_z = c_z/r_z$ is the cost–benefit ratio, and H_z is the hazard rate function associated with the virtual waiting time distribution F_z, namely,

$$H_z(t) := \frac{F_z'(t)}{\overline{F}_z(t)}, \quad t > 0.$$

We shall also define $H_z(0) = H_z(0+)$. The first order condition for a local optimum at $T > 0$, namely, $U_z'(T) = 0$, can now be simply stated as

$$H_z(T) = \gamma_z.$$

Thus, an abandonment can take place only when the hazard rate crosses a specific level, which is just the cost–benefit ratio.

We proceed to characterize the form of the optimal solution under certain monotonicity assumptions on the hazard rate function. We shall consider the following cases:

(a) Increasing hazard rate (IHR): H_z is monotone increasing.

(b) Decreasing hazard rate (DHR): H_z is monotone decreasing.

(c) Increasing–decreasing hazard rate (IDHR): H_z is unimodal, initially increasing and then decreasing.

In all cases we consider $H_z(t)$ for $t \geqslant 0$. Note that monotonicity is not required to be strict, so that the IDHR class includes the other two as special cases.

Consider the IHR case first. It is easily seen from (3) that in this case $U_z(T)$ is either increasing, decreasing, or decreasing–increasing over $[0, \infty)$ and, therefore, will be globally maximized at one of the edges, namely, $T = 0$ or ∞. The optimal decision is, therefore, one of the following:

(1) $T = 0$: abandon immediately if not admitted to service upon arrival.

(2) $T = \infty$: never abandon.

The implication is that it is never optimal to abandon after a finite (nonzero) amount of time.

The DHR case is considered next. Here H_z decreases from $H_z(0)$ to $H_z(\infty)$. In one extreme case the graph of H_z may lie entirely below the level γ_z, implying

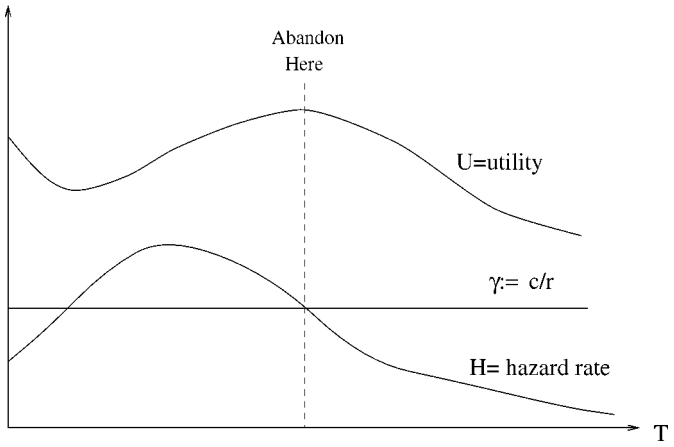

Figure 1. An increasing–decreasing hazard rate. The utility function U_z is maximized either at $T = 0$, or at the intersection of γ_z with the decreasing part of H.

U_z' negative and an optimal decision at $T = 0$. In the other extreme, the graph of H_z lies entirely above γ_z, implying U_z' positive and an optimal decision at $T = \infty$. The interesting case is the intermediate one: when H_z intersects γ_z, the intersection point T_z is easily seen from (3) to correspond to a global maximum of the utility and, hence, is the optimal decision.

We, finally, consider the IDHR case. Here γ_z can intersect H_z at two points at the most (see figure 1), the first at the increasing part of H_z and the second at its decreasing part. The first intersection corresponds to a local minimum, hence is of no consequence. If the second intersection does not exist, then the situation is similar to the IHR case, namely, an optimal decision at $T = 0$ or $T = \infty$. If the second intersection does exist, then similarly to the DHR case it corresponds to a local maximum of H_z, which in fact is the unique local maximum over $T > 0$. In this case the global maximum can be either at that intersection, or at $T = 0$.

We summarize these findings in the following proposition:

Proposition 2. Given F_z and $\gamma_z := c_z/r_z$, let T_z be the optimal abandonment time with respect to the utility function (1). Then:

(i) In the IHR case, $T_z = 0$ or $T_z = \infty$.

(ii) In the DHR case, if γ_z intersects $H_z(T)$ then T_z is that intersection point. Otherwise, either γ_z is above $H_z(\cdot)$ (i.e., $H_z(T) < \gamma_z$ for every T) and $T_z = 0$, or γ_z is below $H_z(\cdot)$ and $T_z = \infty$.

(iii) In the IDHR case, if γ_z intersects the decreasing part of $H_z(T)$ then either T_z is that intersection point or $T_z = 0$. Otherwise, $T_z = 0$ or ∞, with $T_z = 0$ if γ_z is above $H_z(\cdot)$ and $T_z = \infty$ if γ_z is below $H_z(\cdot)$.

Discussion. Let us briefly discuss the three possible monotonicity assumptions that were considered above. As we shall see in the next section, the actual hazard rate in $\mathrm{M}/\mathrm{M}/\mathrm{m}+G$ queues is increasing for any patience distribution G, which motivates the IHR case. However, from the subjective point of view of a waiting customer this assumption does not seem to be realistic, since it implies that customers who wait for a long time become more and more optimistic about the opportunity of obtaining service in the near future, while in a typical scenario we expect that customers eventually become pessimistic about obtaining service speedily. The DHR case, on the other hand, excludes those cases where the virtual waiting time is characterized by some typical value, and the hazard rate will be increasing at least initially up to this value. The IDHR form is the simplest one that accommodates both these tendencies. Moreover, while there is no special reason to maintain *a priori* that the actual hazard rate in a system will be unimodal, from the subjective point of view customers are hardly likely to adopt a more complicated form for their estimate of the hazard rate. Thus, the IDHR case presents a very reasonable balance between simplicity and the ability to capture the essential ingredients of the problem. We are thus led to consider models where this form of the hazard rate arises naturally.

3.2. *The* $\mathrm{M}/\mathrm{M}/\mathrm{m}(q)+G$ *queue*

We now consider some characteristics of the queueing models that are treated in this paper, which are valid for any distribution G of the customer patience. Of special interest are the distribution function F of the virtual waiting time, and the associated hazard rate function $H = F'/\overline{F}$. We start by recalling some explicit expressions for F in the $\mathrm{M}/\mathrm{M}/\mathrm{m}+G$ queue, which play a central role in our analysis. We then observe that the hazard rate function H is increasing (IHR case) in this model. We shall then consider the $\mathrm{M}/\mathrm{M}/\mathrm{m}(q)+G$ queue, where with probability $(1-q)$ customers are subjected to a fault state with infinite waiting time. In this case the hazard rate function turns out to be increasing–decreasing (IDHR).

Consider first the $\mathrm{M}/\mathrm{M}/\mathrm{m}+G$ queue, with patience distribution G. Each customer is characterized by a patience T which is stochastically chosen according to G, independently of other arrival and service primitives. A customer abandons the system if not admitted to service within T time units of arrival. The distribution G may be defective, i.e., $G(\infty) < 1$, so that some customers may have infinite patience. The stability condition $\lambda[1 - G(\infty)] < m\mu$ is assumed to hold [2]. Denote $\overline{G} = 1 - G$, and let

$$I(t) = m\mu - \lambda\overline{G}(t).$$

Let F denote the distribution function of the virtual waiting time V in steady state. Then, from [2] we have

$$F'(t) = \lambda\pi_{m-1}\exp\left(-\int_0^t I(s)\,\mathrm{d}s\right), \quad t \geqslant 0, \tag{4}$$

where π_{m-1} is the stationary probability of having exactly $m-1$ servers occupied. This probability is determined through the normalization condition

$$\sum_{j=0}^{m-1} \pi_j + \int_0^\infty F'(t)\,\mathrm{d}t = 1, \tag{5}$$

and $\pi_j = (1/j!)(\lambda/\mu)^j \pi_0$ for $j = 0, \ldots, m-1$.

We note that F' is well defined and $\overline{F}(t) \neq 0$, as was assumed in section 3.1. Moreover, the second derivative F'' also exists for all $t > 0$ (except possibly at jump points of G, where we may simply define $F'' = -I \cdot F'$). Note also that $F(0-) = 0$, $F(0) = \sum_{j=0}^{m-1} \pi_j$ (an atom at $t = 0$), and $F(\infty) = 1$.

It is well known that for an exponential queue without abandonments, the virtual waiting time V, given that $V > 0$, is exponentially distributed, hence, gives rise to a constant failure rate. When abandonments are present, the following property holds.

Proposition 3. The virtual waiting time distribution F in an $\mathrm{M/M/m} + G$ queue is IHR. Furthermore, the hazard rate $H(t)$ is strictly increasing up to the first point t (possibly infinite) where $G(t) = G(\infty)$, and is constant thereafter.

Proof. Differentiating $H = F'/\overline{F}$ and using $F'' = -I \cdot F'$ gives

$$H'(t) = \frac{F''\overline{F} + (F')^2}{(\overline{F})^2}(t) = \frac{F'(t)}{\overline{F}(t)^2}\left[-I(t)\int_t^\infty F'(s)\,\mathrm{d}s + F'(t)\right].$$

Let $K(t) > 0$ denote the positive term that precedes the square brackets. Since I is increasing, $F' > 0$, and $F'' = -I \cdot F'$, we obtain

$$H'(t) \geqslant K(t)\left[-\int_t^\infty I(s)F'(s)\,\mathrm{d}s + F'(t)\right] = K(t)\left[\int_t^\infty F''(s)\,\mathrm{d}s + F'(t)\right] = 0,$$

where $F'(\infty) = 0$ was used for the last equality. It may also be seen that the above inequality is strict unless I (equivalently G) is constant beyond t, which establishes the claim.

We note that the IHR property may also be established by showing that F''/F' is decreasing, and the latter equals $-I(t)$ which is decreasing since $\overline{G}$ is decreasing. However, the direct calculation used is more instructive. □

Consider next the $\mathrm{M/M/m}(q) + G$ model. This model modifies the standard $\mathrm{M/M/m} + G$ queue, by assuming that an arriving customer has a probability $(1-q)$ of being positioned in a fault state, where he is neglected and never admitted to service. It is important to note that the customer does not know whether he is in a fault state or not.

It is evident that the active part of this queue, of customers that are not in the fault state, is just a standard $\mathrm{M/M/m} + G$ queue with a modified arrival rate $\lambda_q = q\lambda$. Let F denote the virtual waiting time distribution in that queue, given by the

expressions above with λ replaced by λ_q. Let F_q denote the corresponding quantity in the complete system. (Note that F does depend on q through λ_q, however, this dependence is suppressed for notational convenience.)

An arriving customer joins with probability q the main queue, where her virtual waiting time V is distributed according to F; and with probability $(1-q)$ is placed in a fault position, where $V = \infty$ by definition. It follows that

$$F_q(t) := P(V \leqslant t) = qF(t), \quad t \geqslant 0,$$

which is the basic relation for this model. Note also that $\overline{F}_q := 1 - F_q = 1 - qF = q\overline{F} + (1-q)$, and the corresponding hazard rate function can be expressed as

$$H_q = \frac{F_q'}{\overline{F}_q} = \frac{qF'}{1-qF} = \frac{F'}{\overline{F}+g}, \quad t > 0, \tag{6}$$

where $g = (1-q)/q$. It is not hard to verify that for $0 < q < 1$ the hazard rate function H_q will be eventually decreasing, in contrast to the standard case of $q = 1$ as discussed above. Indeed, observe that for large t, $\overline{F}_q$ in the denominator of H_q converges to $(1-q)$, while the numerator decays exponentially as $\exp(-I(\infty)t)$, where $I(\infty) = m\mu - \lambda_q\overline{G}(\infty) > 0$.

It is also easily seen that for q close enough to 1 (so that g is small enough), H_q will inherit the increasing property from $H := F'/\overline{F}$ near $t = 0$, i.e., it will be initially increasing. Therefore, the simplest class to which H_q might generally belong in terms of its monotonicity properties is the IDHR class, defined in section 3.1 – provided that H_q is unimodal. This is verified in the following:

Proposition 4. The virtual waiting time in the $\mathrm{M}/\mathrm{M}/\mathrm{m}(q) + G$ model, with $q < 1$, has the IDHR property; that is, the hazard rate function H_q is unimodal and eventually decreasing. Moreover, it is strictly decreasing with a strictly negative first derivative beyond its maximal point.

Proof. For unimodality it suffices to verify that H_q' can have at most one sign change, from positive to negative. Differentiating H_q and noting that $F'' = -IF'$ by (4) gives

$$H_q' = \frac{F''(\overline{F}+g) + (F')^2}{(\overline{F}+g)^2} = H_q(H_q - I). \tag{7}$$

Noting that $F' > 0$, hence, $H_q > 0$, it follows that H_q' is sign-equivalent to $H_q - I$. But since I is a nondecreasing function of t by its definition, it immediately follows that once H_q' becomes (strictly) negative it will stay that way. This verifies that H_q is unimodal and, moreover, strictly decreasing beyond it maximum. □

Remark. In the definition of the $\mathrm{M}/\mathrm{M}/\mathrm{m}(q)$ system we have assumed that the fault state is an individual state to which each customer is subjected independently of the others. Another important interpretation may be given in terms of a system fault.

Assume that the whole system is in a fault state a fraction $(1-q)$ of the time, during which all arriving customers are subjected to the individual fault state as defined before. Then, provided transients between the operating and fault states of the system can be neglected, as arriving customer will enter a standard M/M/m queue with probability q, and the fault state otherwise; thus, from the customer point of view the situation is equivalent in the two cases. Note also that the system fault interpretation is close in spirit to a server vacation model.

3.3. Some properties of a consistent equilibrium

The consistency assumption implies, in particular, that the subjective distributions F_z all coincide: $F_z \equiv F$. We now develop some consequences of this equality. These properties are not restricted to the M/M/m queue.

We first establish the reassuring property that the rational abandonment times are decreasing in the cost–benefit ratio.

Proposition 5. Let z and y be two customer types, with $F_z = F_y := F$ and $\gamma_z < \gamma_y$. Then the respective individually optimal abandonment times satisfy $T_z \geqslant T_y$. Furthermore, the strict inequality $T_z > T_y$ holds provided that: $0 < T_z < \infty$, F' is continuous at T_z, and $\overline{F}(T_z) > 0$.

Proof. We first observe that if $\overline{F}(T_z) = 0$, meaning that customers who wait in the queue more than T_z will never obtain service, then waiting more than T_z cannot be optimal for any customer; thus $T_z \geqslant T_y$ in this case. Assume henceforth that $\overline{F}(T_z) > 0$.

The optimal decisions are obviously unaffected if we normalize each utility function U_z by $1/r_z$, that is, replace U_z by $W_z = r_z^{-1} U_z$. From (2), the derivative of this normalized utility is

$$W_z' = F' - \gamma_z \overline{F}.$$

Since $\overline{F}$ is non-negative, this derivative is decreasing in γ_z, that is, $W_z' \geqslant W_y'$ at every point t, with strict inequality if $\overline{F}(t) > 0$. This implies that $W_z(t_2) - W_z(t_1) \geqslant W_y(t_2) - W_y(t_1)$ for any pair of points $t_2 > t_1 \geqslant 0$, with strict inequality if $\overline{F}(t_1) > 0$. Now, if $T_z < T_y$, we can identify T_z with t_1 and T_y with t_2, and obtain

$$W_z(T_y) - W_z(T_z) > W_y(T_y) - W_y(T_z).$$

However, this contradicts the assumptions that T_z is z-optimal (hence, $W_z(T_y) - W_z(T_z) \leqslant 0$) and T_y is y-optimal (hence, $W_y(T_y) - W_y(T_z) \geqslant 0$). It follows that $T_z < T_y$ is false, thus $T_z \geqslant T_y$.

To establish the strict inequality under the stated assumptions, note the the continuity of F' at T_z implies continuity of the utility function derivative W_z' at that point, so that the first-order optimality condition $W_z'(T_z) = 0$ must hold. But as observed above, due to $\overline{F}(T_z) > 0$ the strict inequality $W_y' < W_z'$ holds at T_z, so that $W_y'(T_z) < 0$, which implies that T_z is not optimal in W_y. □

The following lemma establishes a useful continuity property of optimal abandonment times, which is valid when the hazard rate function is IDHR.

Lemma 6. Suppose that $F_z = F$ for all $z \in Z$, and assume that the hazard rate function $H = F'/\overline{F}$ is increasing–decreasing, and in fact strictly decreasing beyond its maximum. Then the optimal abandonment times T_z are a continuous function of γ_z, for $\gamma_z \in (0, \infty)$, except for one possible jump from $T_z = 0$ to a positive value.

Proof. As established in proposition 2, an optimal decision T_z in the increasing–decreasing case is either 0 or at the intersection of γ_z with the decreasing part of H. The present assertion is an immediate consequence of that fact. □

4. Existence, uniqueness and structure of the equilibrium

We now turn to the questions of uniqueness, structure and computation of the consistent equilibrium point. We first consider the relatively simple case of the M/M/m queue, in theorem 7, and then extend the results to the M/M/m(q) model, for which the main results are summarized in theorem 8. The detailed derivations and proofs of the latter are deferred to the next section.

The following theorem reveals the special structure of the equilibrium point in the M/M/m model, that is essentially a consequence of the IHR property inherent in the M/M/m + G queue. This structure is employed to establish uniqueness.

Theorem 7. Consider the M/M/m queue with the rational abandonment model. Then there exists a unique consistent equilibrium point, which is of the following form: $T_z = 0$ for $\gamma_z > \theta$, and $T_z = \infty$ for $\gamma_z \leqslant \theta$, where the constant θ is the unique solution of the equation $\theta = I_\theta$, with $I_\theta = m\mu - \lambda P_Z\{z\colon\ \gamma_z < \theta\}$.

Proof. Assume that the system is in consistent equilibrium. From proposition 3 we know that the hazard rate function is monotone increasing. It then follows from part (i) of proposition 2, together with the monotonicity in γ_z of the optimal decisions established in proposition 5, that any equilibrium point must be of the stated form.

Uniqueness now follows using a basic monotonicity argument with respect to the equilibrium parameter θ. Essentially, increasing θ means that more customers remain in the queue, hence, the queue becomes more congested; but then less customers will find it optimal to stay, leading to a unique balance point.

More formally, assume that customers are following the decision rule above with some threshold θ. This leads to a patience distribution G which satisfies $\overline{G}(t) = \overline{G}(0) = P_Z\{z\colon\ \gamma_z < \theta\}$ for $t \geqslant 0$. Substitution in equation (4) yields

$$F'(t) = \lambda \pi_{m-1}\, \mathrm{e}^{-I_\theta t},$$

where $I_\theta = [m\mu - \lambda\overline{G}(0)]$. Consequently, by integration $\overline{F}(t) = I_\theta^{-1} F'(t)$, and $H(t) = F'/\overline{F} = I_\theta$; that is, the hazard rate is constant. Proposition 2 (DHR case) implies then

that the optimal abandonment times are $T_z = 0$ if $\gamma_z > I_\theta$, $T_z = \infty$ if $\gamma_z < I_\theta$, and neutral if $I_\theta - \gamma_z = 0$ (in which case we choose $T_z = \infty$ by convention). For the initially assumed and the latter optimal decision rules to coincide it is required that $\theta = I_\theta$. It remains to verify existence of a unique solution to that equation. By its definition, I_z is decreasing and continuous in z (where the latter follows by our standing assumption that P_Z has a density). Thus, $z - I_z$ is continuous, and strictly increasing from a negative value (at $z = 0$) to $+\infty$, so that $z = I_z$ indeed has a unique solution. □

We remark that if P_Z was allowed to contain point masses, then a similar result could be retained by allowing a probabilistic splitting of customers of identical type (as in [9,10]).

Observe that under the established equilibrium profile, a fraction $\overline{G}(0) = P_Z\{z\colon\ \gamma_z < \theta\}$ of arriving customers have infinite patience and will never abandon the queue, while the remaining customers will abandon immediately if not admitted to service upon arrival. The distribution of nonzero waiting times in this queue (that is the distribution of V conditioned on $V > 0$, which equals $F'(\cdot)/\overline{F}(0)$) coincides with that of a standard M/M/m queue with arrival rate $\lambda\overline{G}(0)$. However, the chance of finding a free server upon arrival will be smaller in the present case due to the effect of the impatient customers.

We have thus established the uniqueness of the consistent equilibrium in the M/M/m queue, and obtained an explicit form for the equilibrium abandonment decisions. The notable property of this equilibrium is that abandonments should occur only immediately upon arrival; as noted, this is a consequence of the IHR property which is inherent in the M/M/m + G queue. Obviously, this structural constraint presents a serious limitation of this model.

We now turn to the M/M/m(q) model. As has already been shown, the introduction of the fault state introduces a decreasing tail in the hazard rate function, and consequently abandonments after a finite wait in the queue become feasible as a rational choice.

As soon as finite abandonment times are introduced, the fixed-point problem becomes multi-dimensional, and a simple monotonicity argument as used in the last proof cannot be applied to establish uniqueness of the equilibrium point. To be specific, consider the case of only two customer types, $z = 1$ and $z = 2$, and assume an equilibrium point with abandonment times T_1 and T_2. It is quite reasonable that another equilibrium point with uniformly larger times ($T_1' > T_1$ and $T_2' > T_2$) cannot exist, since then the system becomes more congested and a rational choice should be to abandon earlier rather than later. However, if T_1 and T_2 are modified in opposite directions (say, $T_1' > T_1$ but $T_2' < T_2$), it is not clear what would be the overall effect on the system, and whether these new values might constitute an additional equilibrium.

This difficulty will be tackled by first establishing detailed structural properties that must hold in any equilibrium point. For this purpose we exploit the special structure

of the virtual waiting time distribution in the M/M/m(q)+G queue, as inherited from the M/M/m+G queue. In the process we develop some formulas and relations which will enable explicit computation of the equilibrium profile.

The next theorem summarizes our main findings concerning the structure and computation of the equilibrium in the M/M/m(q) model. The following quantities will be required. For $0 < \gamma < \infty$, let

$$I_\gamma = m\mu - \lambda_q P_Z\{z\colon\ \gamma_z < \gamma\}$$

and define $\gamma^{\rm o}$ as the unique solution to $I_\gamma - \gamma = 0$. Further define, for $0 < \gamma \leqslant \gamma^{\rm o}$,

$$J(\gamma) = \exp\left(\int_0^\gamma (I_y - y)^{-1}\,{\rm d}y\right), \qquad L(\gamma) = \left(\frac{\gamma}{\lambda_q B_m} + 1\right) J(\gamma),$$

where B_m is specified in (18).

Theorem 8. Consider the M/M/m(q) model with rational abandonments.

(i) A consistent equilibrium exists and is unique.

(ii) The equilibrium profile has one of the following two alternative forms:

(a) If $L(\gamma^{\rm o}) \geqslant (1-q)^{-1}$: Let θ be the unique solution of $L(\theta) = (1-q)^{-1}$ on $(0, \gamma^{\rm o}]$. Then $T_z = 0$ for $\gamma_z > \theta$, and

$$T_z = \tau(\gamma_z) - \tau(\theta) := \int_{\gamma_z}^{\theta} \frac{y^{-1}}{I_y - y}\,{\rm d}y \quad \text{for } \gamma_z \leqslant \theta.$$

(b) If $L(\gamma^{\rm o}) < (1-q)^{-1}$: $T_z = 0$ for $\gamma_z > \gamma^{\rm o}$, and

$$T_z = T^{\rm o} + \int_{\gamma^z}^{\gamma^{\rm o}} \frac{y^{-1}}{I_y - y}\,{\rm d}y \quad \text{for } \gamma_z \leqslant \gamma^{\rm o},$$

where $T^{\rm o} > 0$ is given by the solution to (23), namely,

$$T^{\rm o} = \frac{1}{\gamma^{\rm o}} \log\left(\frac{(1-q)^{-1}}{J(\gamma^{\rm o})} - \frac{\gamma^{\rm o}}{\lambda_q B_m}\right).$$

(iii) If the probability density of γ_z is bounded in magnitude, then $L(\gamma^{\rm o}) = \infty$ and the equilibrium is necessarily in form (a).

(iv) The equilibrium hazard rate function H_q is non-increasing. In fact, it is strictly decreasing in case (a), while in case (b), $H_q(t) \equiv \gamma^{\rm o}$ for $0 \leqslant t \leqslant T^{\rm o}$, and it is strictly decreasing thereafter.

The proof of this theorem may be found in the next section, together with a more detailed discussion of the equilibrium structure.

The two possible forms of equilibrium are depicted in figure 2. The examples below serve to further illustrate these results.

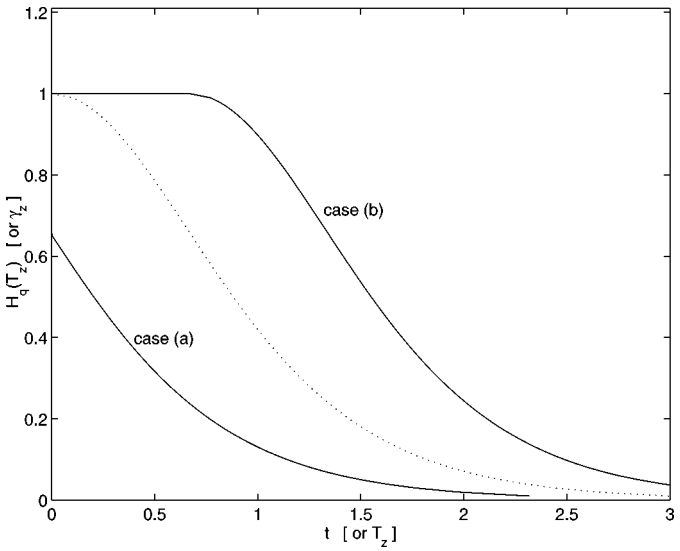

Figure 2. An illustration of the two possible equilibrium forms. The graph depicts the equilibrium hazard rate function $H_q(t)$ as a function of t. Since $H_q(T_z) = \gamma_z$, the inverse function of H_q displays the abandonment times T_z as a function of γ_z. (The illustrated equilibria corresponds to example 1, with $q = 0.5$ for case (a), $q = 0.85$ for case (b), and $q = 0.755$ in between.)

Given (iii) of the last theorem, it is evident that the equilibrium profile will be in form (a) in most cases of interest. In fact, the question may be raised whether form (b) of the equilibrium is obtainable at all. The following example gives the positive answer.

Example 1. Let $\lambda_q = 2$, $m = 1$, $\mu = 2$, and $P_Z(z) := P_Z\{z' \colon z' < z\} = 1 - \sqrt{1-z} - 0.5z$ for $0 \leqslant z \leqslant 1$, and arbitrary for $z > 1$. Note that $P_Z(0) = 0$, $P_Z(1) = 0.5$, and the associated density f_Z equals $(2\sqrt{1-z})^{-1} - 0.5$ on $[0,1]$, hence, is unbounded near 1. To determine the equilibrium form according to proposition 15 we evaluate $J(z^{\mathrm{o}})$. Here $I_z - z = 2\sqrt{1-z}$ on $0 \leqslant z \leqslant 1$, with stationary point $z^{\mathrm{o}} = 1$, and $J(z^{\mathrm{o}}) = \exp(\int_0^{z^{\mathrm{o}}} (2\sqrt{1-z})^{-1}\,\mathrm{d}z) = \exp(1) = \mathrm{e}$. Since $z^{\mathrm{o}}/\lambda_q B_m = 0.5$, the equilibrium will be in form (a) if $1.5\mathrm{e} \geqslant (1-q)^{-1}$ (i.e., $q \leqslant 1-(1.5\mathrm{e})^{-1} \approx 0.755$), but in form (b) otherwise. Figure 2 shows the different equilibrium profiles obtained for several choices of q.

Finally, we show how explicit solutions may be computed when the customer-type distribution is specified.

Example 2. *Uniform type distribution.* To illustrate the computational results, we consider the case of a uniform distribution P_Z, namely, γ_z is distributed uniformly on $[0,1]$. Then

$$I_\gamma = m\mu - \lambda_q P_Z\{z \colon \gamma_z < \gamma\} = m\mu - \lambda_q \min\{\gamma, 1\} \quad \text{for } \gamma \geqslant 0.$$

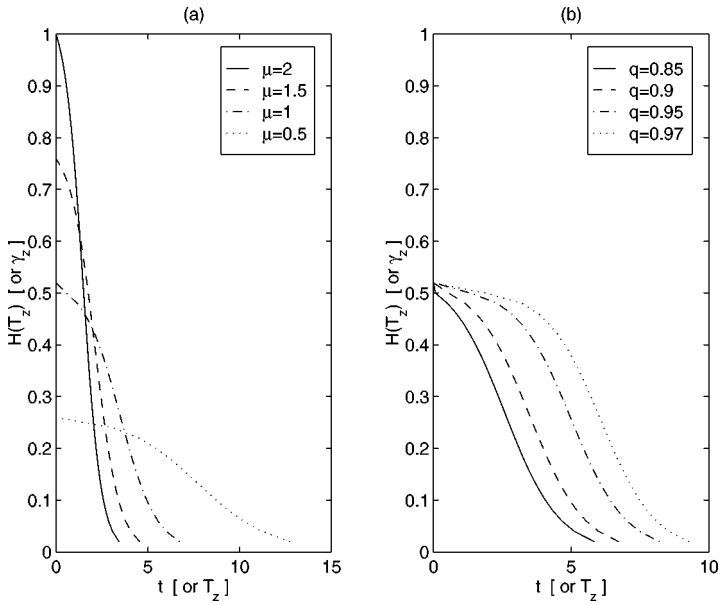

Figure 3. Equilibrium profiles for an M/M/m(q) system with uniformly distributed customer types (example 2). In part (a) $q = 0.9$ while μ is modified, and in part (b) $\mu = 1$ while q is modified.

Assume for simplicity that $m\mu/(\lambda_q + 1) \leqslant 1$ (the computations otherwise are similar but somewhat more cumbersome). Then the solution γ^{o} to $\gamma - I_\gamma = 0$ is simply $\gamma^{\mathrm{o}} = m\mu/(\lambda_q + 1)$. Next,

$$J(\gamma) = \exp\left(\int_0^\gamma \frac{1}{m\mu - (\lambda_q + 1)y}\,\mathrm{d}y\right) = \exp\left(\frac{1}{\lambda_q + 1}\log\frac{\gamma^{\mathrm{o}}}{\gamma^{\mathrm{o}} - \gamma}\right)$$
$$= \left(\frac{\gamma^{\mathrm{o}}}{\gamma^{\mathrm{o}} - \gamma}\right)^{1/(\lambda_q+1)}.$$

It may be seen that $J(\gamma^{\mathrm{o}}) = \infty$, hence, $L(\gamma^{\mathrm{o}}) = \infty$, which implies that the equilibrium profile must be in form (a), as implied by theorem 8(iii). The equilibrium parameter θ is the solution to

$$L(\theta) := \left(\frac{\theta}{\lambda_q B_m} + 1\right) J(\theta) = (1 - q)^{-1},$$

which needs to be evaluated numerically. Finally, $T_z = \tau(\gamma_z) - \tau(\theta)$ for $\gamma_z \leqslant \theta$, with

$$\tau(\gamma) = \int^\gamma \frac{-1}{y(m\mu - (\lambda_q + 1)y)}\,\mathrm{d}y = \frac{1}{m\mu}\log\frac{m\mu - (\lambda_q + 1)\gamma}{\gamma}.$$

Some numerical results for this example are presented in figure 3 for a system with parameters $m = 1$, $\lambda = 1$. Part (a) of this figure presents the equilibrium points obtained with q fixed at 0.9, for several values of the service rate μ. It may be seen that as μ increases, the fraction of customers who will not abandon immediately (given

by θ, in light of the uniform type distribution on $[0, 1]$) also increases, approximately in linear proportion to μ. However, the abandonment times of those customers who choose to stay tend to become shorter. Part (b) depicts the equilibria obtained for the same system, with μ fixed at 1, and several values of the service reliability parameter q. As q increases, the abandonment times of waiting customers become larger. However, the fraction of customers who abandon immediately remains almost constant.

5. Proof of theorem 8

In this section we provide the proofs for the main results in the previous section concerning the M/M/m(q) model, as summarized in theorem 8. The analysis proceeds through several lemmas. We first identify in lemma 9 the general structure of the equilibrium profile, which is a consequence of the IDHR property inherent in the M/M/m(q) + G queue: the abandonment times are zero above some type threshold, and then are positive and increasing as the type decreases below this threshold. The key lemma 11 considers the positive part of the abandonment profile, and derives an explicit function of the customer types which specifies positive abandonment times to within a constant shift. The transition from zero to positive abandonment times is addressed in lemmas 10 and 12, which establish that this transition is either done continuously or at a specific value of the type parameter. These results provide us with a set of candidate equilibrium profiles, specified in proposition 13, which are essentially parameterized by a one-dimensional parameter and strictly dominate each other. Uniqueness will then be established by using the normalization condition (5).

For the purpose of the forthcoming analysis, it will be convenient to use a canonical parameterization of the customer types, namely,

$$z \equiv \gamma_z,$$

which identifies the customer type with the cost-benefit ratio parameter. According to our assumptions on γ_z, z is then distributed on $(0, \infty)$ according to the distribution P_Z which admits a density. Except for replacing γ_z with z, other notations are not affected. This canonical parameterization will be maintained till the end of this section.

We start by pointing to some basic relations that will be used repeatedly in the following. Given a decision profile $\{T_z\}$, the virtual waiting time distribution F in the active (M/M/m) part of the M/M/m(q) system is given by (4), with $I(t) = m\mu - \lambda_q \overline{G}(t)$, and

$$\overline{G}(t) = P_Z\{z\colon\ T_z > t\}.$$

Assume next that $\{T_z\}$ is a consistent equilibrium profile. Then we can deduce the important observation that $\overline{G}(T_z)$ is a fixed quantity for each z. Indeed, monotonicity of T_z in z (proposition 5 with $z \equiv \gamma_z$) implies that

$$\overline{G}(T_z) = P_Z\big\{z'\colon\ T_{z'} > T_z\big\} = P_Z\big\{z'\colon\ z' < z\big\} := P_Z(z).$$

Obviously the latter is a function of z alone and does not depend on the particular equilibrium considered. We thus obtain

$$I(T_z) = m\mu - \lambda_q P_Z(z) := I_z, \tag{8}$$

where I_z again depends only on z.

We further recall that optimality of T_z implies that $H(T_z) = \gamma_z$ whenever $T_z > 0$; hence, $H(T_z) = z$ under parameterization $z = \gamma_z$.

The first lemma concerns the structure of an equilibrium profile, and is a consequence of the IDHR property of the $\mathrm{M}/\mathrm{M}/\mathrm{m}(q) + G$ queue.

Lemma 9. Let $\{T_z\}$ be an equilibrium profile, and H_q the corresponding hazard rate function. Then

(i) $\{T_z\}$ is of the following form, for some constant $\theta > 0$:

  (a) $T_z = 0$ for $z > \theta$.

  (b) $T_z > 0$ for $z < \theta$, and is then specified by the intersection of z with the decreasing part of H_q; in particular, $H_q(T_z) = z$.

  (c) $z = \theta$ is indifferent between $T = 0$ and $T = \lim_{z\uparrow\theta} T_z \geqslant 0$. By convention we define T_θ as the larger value.

(ii) T_z is a strictly decreasing and continuously differentiable function of z on $0 < z \leqslant \theta$, and $H_q(t)$ is strictly decreasing for $t > T_\theta$.

Proof. (i) By proposition 4, H_q is in the IDHR class. The stated form of the equilibrium point then follows from proposition 2(iii) combined with the monotonicity result in proposition 5. The neutrality of $z = \theta$ follows from continuity of the cost function in z. Finally, it is easily argued that for z small enough (diminishing waiting cost) it will be preferable to stay in the queue for some positive time rather then abandon immediately, so that $\theta > 0$.

(ii) By proposition 4, H_q is strictly decreasing beyond its maximum point. But T_θ is already on the decreasing part, so that $H_q(T_z) = z$ implies that T_z is strictly decreasing (and continuous) in z for $z < \theta$.

To establish differentiability, note the F' is continuous by its expression in (4), hence, so is $H_q = F'/(\overline{F} + g)$. Also, for $t > T_\theta$,

$$\overline{G}(t) = P_Z\{z\colon\ T_z > t\} = P_Z\big\{z\colon\ H_q(T_z) < H_q(t)\big\} = P_Z\big\{z\colon\ z < H_q(t)\big\},$$

and since P_Z has a density (i.e., is absolutely continuous) by assumption it follows that $\overline{G}$ is continuous. Revisiting (4), where $I = m\mu - \lambda_q \overline{G}$, it follows that F', hence H_q, is continuously differentiable, and $H_q(T_z) = z$ with H_q strictly decreasing implies the same for T_z. □

Note that the definition of T_z in lemma 9 extends to every $z > 0$, even if z is not in the support of P_Z. This will conveniently enable to consider derivatives with respect to z on the entire positive real line.

The next lemma establishes a basic cutoff value in the type (or cost–benefit ratio) parameter, beyond which customers will necessarily choose to abandon the queue immediately if not admitted to service upon arrival.

Lemma 10. Let $I_z = m\mu - \lambda_q P_Z(z)$, as defined in (8). Then for every $z > 0$, $z - I_z > 0$ implies $T_z = 0$. Equivalently, $T_z = 0$ for $z > z^{\mathrm{o}}$, where z^{o} is the unique solution of $z - I_z = 0$.

Proof. Assume $T_z > 0$. We proceed to show that $z - I_z \leqslant 0$, thereby verifying the first assertion. Differentiating H_q, as in the proof of proposition 4, shows that H_q' is sign-equivalent to $[H_q - I]$; hence, $H_q'(T_z)$ is sign-equivalent to $[H_q(T_z) - I(T_z)]$, while $I(T_z) = I_z$ by (8).

From lemma 9, $H_q' \leqslant 0$ at $t = T_z > 0$, so that $H_q(T_z) - I_z \leqslant 0$. However, the optimality condition for $T_z > 0$ is $H_q(T_z) = z$, so that $z - I_z \leqslant 0$ follows, as we set out to show. Finally, the existence of a unique solution to the equation $z - I_z = 0$ was established in the proof of theorem 7. □

Next, we provide an explicit characterization of the equilibrium profile T_z for positive abandonment times, which specifies these times to within a constant shift. This is done, essentially, by moving backwards on the waiting-time axis, from large to small T, and simultaneously constructing the equilibrium profile and the virtual waiting time distribution F.

Lemma 11.

(i) There exists a function $\tau(z)$, independent of the equilibrium point considered, so that every equilibrium profile satisfies, for some constant C:

$$T_z = \tau(z) + C \quad \text{whenever } T_z > 0.$$

(ii) When $T_z > 0$, both $F'(T_z)$ and $\overline{F}(T_z)$ depend only on z but not on the particular equilibrium point. We denote these values as F_z' and $\overline{F}_z$, respectively.

Proof. (i) Let $\{T_z\}$ be an equilibrium profile, of the form specified in lemma 9. Consider $z < \theta$, where $T_z > 0$ by definition of θ and the optimality condition $H_q(T_z) = z$ holds. Recalling that $H_q = F'/(\overline{F} + g)$, this optimality condition may be written as

$$z^{-1}F'(T_z) - \overline{F}(T_z) = g.$$

Differentiating with respect to z gives

$$-z^{-2}F'(T_z) + z^{-1}F''(T_z)\frac{\mathrm{d}T_z}{\mathrm{d}z} + F'(T_z)\frac{\mathrm{d}T_z}{\mathrm{d}z} = 0,$$

where all derivatives are well defined (cf. lemma 9(ii)).

From (4) we know that $F''(T_z) = -I(T_z)F'(T_z)$, where $I(T_z) = I_z$ as specified in (8). Substituting in the last equation and cancelling $F' > 0$ gives $(-z^{-1}I_z + 1)(\mathrm{d}T_z/\mathrm{d}z) = z^{-2}$, or

$$\frac{\mathrm{d}T_z}{\mathrm{d}z} = -\frac{z^{-1}}{I_z - z}. \tag{9}$$

Since the right-hand side does not depend on the equilibrium point considered, this establishes part (i) of the lemma, with

$$\tau(z) = \int^z \frac{-y^{-1}}{I_y - y}\,\mathrm{d}y. \tag{10}$$

We note that $I_z - z > 0$ must hold for $z < \theta$, since $(\mathrm{d}T_z/\mathrm{d}z) < 0$ there by lemma 9. See also a comment below lemma 12 concerning the positivity of $I_z - z$.

(ii) Starting again with the optimality condition

$$\frac{F'(T_z)}{\overline{F}(T_z) + g} = z,$$

multiplying both sides by $-\mathrm{d}T_z/\mathrm{d}z$ we obtain

$$\frac{\mathrm{d}}{\mathrm{d}z}\log\big(\overline{F}(T_z) + g\big) = -z\,\frac{\mathrm{d}T_z}{\mathrm{d}z} = (I_z - z)^{-1}.$$

Together with the initial conditions $\lim_{z\to 0}\overline{F}(T_z) = \overline{F}(\infty) = 0$, this equation uniquely defines $\overline{F}(T_z)$ as a function of z, namely,

$$\overline{F}(T_z) = -g + g\exp\left(\int_0^z (I_y - y)^{-1}\,\mathrm{d}y\right) := \overline{F}_z. \tag{11}$$

$F'(T_z)$ can now be determined by differentiation, or more simply via the optimality condition:

$$F'(T_z) = z\big(\overline{F}(T_z) + g\big) := F'_z. \tag{12}$$

□

Remark. An alternative proof to lemma 11 could start with the basic differential relation (7) for the hazard rate $H_q(t)$. Together with the equalities $I(T_z) = I_z$ and $z = H_q(T_z)$ it implies that H_q is a solution of the following autonomous first-order differential equation: $H'_q = H_q(H_q - I_{H_q})$, where $I_{H_q(t)}$ is simply I_z evaluated at $z = H_q(t)$. Then (9) can be deduced from $H_q(T_z) = z$, namely, that T_z is the inverse function of $H_q(t)$.

Let us briefly consider the options for the structure of the equilibrium profile, in view of our results so far. Referring to lemma 9, we can distinguish two cases which give rise to different equilibrium structure: either $T_\theta = 0$, or $T_\theta > 0$. In the former case the equilibrium is completely determined by the single parameter θ, since the equilibrium profile for positive abandonment times ($z < \theta,\ T_z > 0$) is determined by lemma 11. In the latter case, however, there seem to be two independent parameters θ and T_θ, where the latter represents a jump in the equilibrium profile from $T_{\theta+} = 0$ to

a positive value $T_\theta > 0$. We now examine the second case more closely, and show that such a jump can occur only at a specific value of θ. Furthermore, an interesting property of the hazard rate function is established for this case.

Lemma 12. Refer to the equilibrium structure as established in lemma 9, and the cutoff value z^o defined in lemma 10. Suppose $T_\theta > 0$. Then $\theta = z^o$, and $H_q(t) = z^o$ for $0 \leqslant t \leqslant T_\theta$.

Proof. From lemma 9(c), $T_\theta > 0$ implies that

$$U_\theta(T_\theta) - U_\theta(0) = 0. \tag{13}$$

Recall that by (3)

$$U'_\theta(t) \text{ is sign-equivalent to } \big[H_q(t) - \theta\big]. \tag{14}$$

Also, recall from the proof of lemma 10 that $H'_q(t)$ is sign-equivalent to $[H_q(t) - I(t)]$. Now, on $0 \leqslant t \leqslant T_\theta$, since there are no abandonments between 0 and T_θ we have $\overline{G}(t) := P_Z\{z\colon\ T_z > t\} = \overline{G}(T_\theta)$, hence, $I(t) := m\mu - \lambda_q \overline{G}(t) = I_\theta$, so that

$$H'_q(t) \text{ is sign-equivalent to } \big[H_q(t) - I_\theta\big] \quad \text{on } 0 \leqslant t \leqslant T_\theta. \tag{15}$$

It follows from this sign equivalence that $[H_q - I_\theta]$ (and H'_q) must keep the same sign on $[0, T_\theta]$ – in fact, it must be either strictly positive, or strictly negative, or zero on that entire interval.

We are now ready to show that $\theta = z^o$. From lemma 10 and the definition of θ it is obvious that $\theta \leqslant z^o$, so that it is enough to show that $\theta < z^o$ is not possible. But if $\theta < z^o$, then $\theta < I_\theta$ follows by the definition of z^o and the strict monotonicity of $(z - I_z)$. Invoking the optimality condition at T_θ gives

$$H_q(T_\theta) = \theta < I_\theta.$$

But then by (15), $H'_q(T_\theta) < 0$, and the sign preservation property established above implies that $H'_q(T_\theta) < 0$ on $[0, T_\theta]$. Together with $H_q(T_\theta) < \theta$ this means that $H_q(t) - \theta > 0$ on $[0, T_\theta]$, and by (14) this implies that $U'_\theta(t) > 0$ on that interval. But this contradicts (13). It follows that $\theta < z^o$ cannot hold, hence, $\theta = z^o$ is established.

Consider next the hazard rate function given that $\theta = z^o$. By definition of z^o we then have $I_\theta = \theta$. Now, if $H_q(t) - \theta \neq 0$ at $t = 0$, it follows by the above sign preservation property that is must keep the same sign on $[0, T_\theta]$, hence, so does U'_θ. But this again contradicts (13), which establishes that $H_q(0) - \theta = 0$, and by the sign preservation property this must hold on the entire interval $[0, T_\theta]$, as asserted. □

A few comments are due regarding the last result. In the case of $T_\theta > 0$ (hence, $\theta = z^o$), the utility function $U_{z^o}(T)$ is constant (at its maximal value) for $0 \leqslant t \leqslant T_\theta$; see (14). It follows that any choice of T in the interval $[0, T_\theta]$ is optimal for type z^o customers in this case.

The fact that H_q is constant on $0 \leqslant t \leqslant T_\theta$ is of particular interest, since it will allow to conclude that the hazard rate function in equilibrium is always non-increasing.

The last proof also shows that the threshold θ satisfies $I_\theta - \theta \geqslant 0$ (with equality if $\theta = z^{\mathrm{o}}$, and strict inequality if $\theta < z^{\mathrm{o}}$). If follows by monotonicity that $I_z - z > 0$ for $z < \theta$, which is consistent with the observation that the function $\tau(z)$ in (10) is strictly decreasing for $z < \theta$.

We summarize our findings regarding the structure of an equilibrium point in the following proposition. The two possible forms of equilibrium are illustrated in figure 2.

Proposition 13. Consider the M/M/m(q) model with rational abandonments. In any equilibrium point:

(i) The equilibrium profile has one of the following two alternative forms, with z^{o} as defined in lemma 10:

(a) For some $\theta \leqslant z^{\mathrm{o}}$, we have $T_z = 0$ on $z > \theta$, and

$$T_z = \tau(z) - \tau(\theta) := \int_z^\theta \frac{y^{-1}}{I_y - y}\,\mathrm{d}y \quad \text{for } z \leqslant \theta. \tag{16}$$

(b) For some constant $T_{z^{\mathrm{o}}} > 0$, $T_z = 0$ for $z > z^{\mathrm{o}}$ (hence, $\theta = z^{\mathrm{o}}$), and

$$T_z = T_{z^{\mathrm{o}}} + \int_z^{z^{\mathrm{o}}} \frac{y^{-1}}{I_y - y}\,\mathrm{d}y \quad \text{for } z \leqslant z^{\mathrm{o}}. \tag{17}$$

(ii) The associated hazard rate function H_q is non-increasing. In fact, in case (b), $H_q(t) = z^{\mathrm{o}}$ for $0 \leqslant t \leqslant T_{z^{\mathrm{o}}}$.

Proof. The stated form of the equilibrium follows from lemma 9, combined with lemmas 10 and 11, where the the function τ is specified in (10). The fact that the hazard rate function is non-increasing follows from lemma 9(ii) (for $t > T_\theta$) and lemma 12 (for $0 \leqslant t \leqslant T_\theta$), where the latter also established that $H_q = z^{\mathrm{o}}$ on the indicated interval. □

Given these structural characteristics of the equilibrium, we have essentially obtained a one-dimensional parameterization of all possible equilibrium points. It should be noted that these candidate equilibrium profiles are completely dominated by each other; that is, the profile $\{T_z\}$ is (weakly) increasing as θ increases from 0 to z^{o}, and then as $T_{z^{\mathrm{o}}}$ increases from 0 to infinity.

Uniqueness of the equilibrium may now be established by applying an appropriate normalization condition.

Theorem 14. For the M/M/m(q) model with rational abandonments, a consistent equilibrium exists and is unique.

Proof. Recall that F' is the virtual waiting time distribution in the active (M/M/m) part of the M/M/m(q) system, and must satisfy the normalization condition (5). Observe that $\sum_{j=0}^{m-1} \pi_j = B_m^{-1}\pi_{m-1}$, where the constant B_m is given by

$$B_m = \frac{(1/(m-1)!)(\lambda_q/\mu)^{m-1}}{\sum_{j=0}^{m-1}(1/j!)(\lambda_q/\mu)^j}. \tag{18}$$

(Note that this coincides with the Erlang-B formula.) Also, (4) implies that $F'(0) = \lambda_q \pi_{m-1}$, so that (5) may be written as

$$\frac{1}{\lambda_q B_m} F'(0) + \int_0^\infty F'(t)\,\mathrm{d}t = 1. \tag{19}$$

We will show that only one of the candidate equilibrium points suggested by the previous theorem satisfies this condition.

As already noted, the set of candidate equilibrium profiles may be considered a function of a single parameter, which first increases (as θ) from 0 to z^{o}, and then increases (as $T_{z^{\mathrm{o}}}$) from 0 to infinity. Refer to this parameter as the equilibrium parameter. Using relations implied by the optimality conditions, we shall associate with each candidate equilibrium profile a virtual waiting time density $F'(t)$, and show that the latter is an increasing function of the equilibrium parameter (at every t), so that only one candidate F' can satisfy the normalization condition above. Existence can be established by noting that F' is actually continuously increasing in the equilibrium parameter, so that the normalization condition is satisfied by one of the candidate equilibria, which is therefore an equilibrium point. Here we shall take a more direct approach, and derive explicit expressions for the normalization condition on F' which will turn out monotonic and continuous, and which will also be useful for computational purposes.

To start, observe that any $F'(t)$ associated with an equilibrium profile must be strictly decreasing in t. Indeed, $F''(t) = -I(t)F'(t)$ and

$$I(t) \geqslant I(0) = I_\theta \geqslant I_{z^{\mathrm{o}}} = z^{\mathrm{o}} > 0;$$

here the first relation is by definition of I, the second by definition of θ as the cutoff value, the third since $\theta \leqslant z^{\mathrm{o}}$ by lemma 10, and the last two by definition of z^{o}.

Consider first a candidate equilibrium in form (a), parameterized by $0 < \theta \leqslant z^{\mathrm{o}}$. Observe, from (16), that for a given θ, T_z decreases continuously from ∞ to 0 as z increases from 0 to θ. Furthermore, T_z is strictly increasing in θ at any z for which $T_z > 0$. Also recall, from lemma 11(ii), that $F'(T_z) = F'_z$, independent of the specific equilibrium, whenever $T_z > 0$. But since $F'(t)$ is strictly decreasing in t, as observed above, it is now easily shown that $F'(t)$ is strictly increasing in θ at every t. Indeed, refer to two candidate equilibria with corresponding parameters $\theta < \hat{\theta}$. Denote by $\widehat{T}_z$ and $\widehat{F}$ the quantities related to $\hat{\theta}$. Then for any $t > 0$ there exists z so that $T_z = t$, and consequently

$$F'(t) = F'(T_z) = F'_z = \widehat{F}'\big(\widehat{T}_z\big) < \widehat{F}'(T_z) = \widehat{F}'(t),$$

where the inequality follows from $T_z < \widehat{T}_z$.

Let us write explicitly the normalization condition for a candidate equilibrium in form (a). Note that (19) may be written as $1/(\lambda_q B_m)F'(0) + \overline{F}(0) = 1$. Using expressions (11) and (12) for $\overline{F}$ and F' at time $T_\theta = 0$, we obtain

$$\frac{g}{\lambda_q B_m}\,\theta\, J(\theta) + \big(-g + g\, J(\theta)\big) = 1,$$

where

$$J(\theta) := \exp\left(\int_0^\theta (I_z - z)^{-1}\,\mathrm{d}z\right). \tag{20}$$

Collecting terms and noting that $g = (1-q)/q$ gives

$$\left(\frac{1}{\lambda_q B_m}\theta + 1\right) J(\theta) = (1-q)^{-1}. \tag{21}$$

Observe that the left-hand side of this equality condition is continuously and strictly increasing in $\theta \in [0, z^{\mathrm{o}}]$, from 1 to a positive value.

Consider next a candidate equilibrium in form (b), parameterized by $T_{z^{\mathrm{o}}} \geqslant 0$. Treat $F'(t)$ separately on $0 < t \leqslant T_{z^{\mathrm{o}}}$ and $t > T_{z^{\mathrm{o}}}$. On the latter interval it may be shown that $F'(t)$ is increasing in $T_{z^{\mathrm{o}}}$, using the same argument as in form (a). On the former interval, since we have there $I(t) = I_{z^{\mathrm{o}}} = z^{\mathrm{o}}$ (see above (15), and the definition of z^{o}), it follows by (4) that $F'(t) = F'(0)\exp(-z^{\mathrm{o}}t)$ there, so that

$$F'(t) = F'(T_{z^{\mathrm{o}}})\exp\big(z^{\mathrm{o}}(T_{z^{\mathrm{o}}} - t)\big) \quad \text{for } 0 \leqslant t \leqslant T_{z^{\mathrm{o}}}. \tag{22}$$

But since z^{o} and $F'(T_{z^{\mathrm{o}}}) = F'_{z^{\mathrm{o}}}$ are (positive) constants, it obviously follows that $F'(t)$ is strictly increasing in $T_{z^{\mathrm{o}}}$ on this interval as well.

We proceed to express explicitly the normalization condition for a candidate equilibrium in form (b). Here we start with (19) written as

$$\frac{1}{\lambda_q B_m}\,F'(0) + \int_0^{T_{z^{\mathrm{o}}}} F'(t)\,\mathrm{d}t + \overline{F}(T_{z^{\mathrm{o}}}) = 1.$$

Using expressions (11) and (12) for $\overline{F}(T_{z^{\mathrm{o}}})$ and $F'(T_{z^{\mathrm{o}}})$, together with (22), we obtain after integration and rearranging terms,

$$\left(\frac{z^{\mathrm{o}}}{\lambda_q B_m} + \mathrm{e}^{z^{\mathrm{o}}T_{z^{\mathrm{o}}}}\right) J\big(z^{\mathrm{o}}\big) = (1-q)^{-1}, \tag{23}$$

where $J(z^{\mathrm{o}})$ is defined in (20). Again, the left-hand side of this condition is a continuously increasing function of $T_{z^{\mathrm{o}}}$, from a positive value (which coincides with the left hand side of (21) for $\theta = z^{\mathrm{o}}$) up to infinity as $T_{z^{\mathrm{o}}}$ increases from 0 to infinity.

It follows that the normalization condition in (21) and (23) will be satisfied for a unique equilibrium parameter. Thus, one and only one candidate equilibrium is consistent with the normalization condition (19), and is, therefore, the unique equilibrium point of the system considered. □

We shall now use the expressions obtained in the last proof in order to compute the equilibrium parameter, which specifies the actual equilibrium point in the set of candidate equilibria.

Proposition 15. Let z^{o} be defined as in lemma 10, $J(z)$ as in (20), and the candidate equilibrium profiles defined in proposition 13. If $(1/(\lambda_q B_m)z^{\mathrm{o}} + 1)J(z^{\mathrm{o}}) \geqslant (1-q)^{-1}$, then the equilibrium point is of the form (a), with θ given by the solution to (21). Otherwise, the equilibrium point is in form (b), with $T_{z^{\mathrm{o}}}$ obtained explicitly from (23).

Proof. The specified condition for selecting between the equilibrium forms is just the normalization condition (21) for $\theta = z^{\mathrm{o}}$, which coincides with (23) for $T_{z^{\mathrm{o}}} = 0$. The rest is a consequence of the previous proof. □

The next result shows that the equilibrium will be in form (a) in most cases of practical interest. Recall, however, that in example 1 above it was shown that form (b) may arise under certain conditions.

Proposition 16. Assume that the density $f_Z = \mathrm{d}P_Z/\mathrm{d}z$ of z is bounded in magnitude. Then the equilibrium profile is in form (a), as defined in proposition 13.

Proof. We show that $J(z^{\mathrm{o}}) = \infty$, which implies that the equilibrium is in form (a) by proposition 15. From (20), $J(z^{\mathrm{o}}) = \exp(\int_0^{z^{\mathrm{o}}} (I_z - z)^{-1}\,\mathrm{d}z)$. Recall that $I_{z^{\mathrm{o}}} - z^{\mathrm{o}} = 0$, and by strict monotonicity, $I_z - z > 0$ for $z < z^{\mathrm{o}}$. Furthermore, for $z < z^{\mathrm{o}}$,

$$I_z - z = (I_z - z) - \big(I_{z^{\mathrm{o}}} - z^{\mathrm{o}}\big) = \lambda_q P_Z\big\{z' \colon\ z \leqslant z' < z^{\mathrm{o}}\big\} + \big(z^{\mathrm{o}} - z\big).$$

Let $B < \infty$ be an upper bound on $f_Z(z)$; then

$$I_z - z \leqslant \lambda_q B\big(z^{\mathrm{o}} - z\big) + \big(z^{\mathrm{o}} - z\big) = (\lambda_q B + 1)\big(z^{\mathrm{o}} - z\big),$$

so that

$$J\big(z^{\mathrm{o}}\big) \geqslant \exp\left((\lambda_q B + 1)^{-1} \int_0^{z^{\mathrm{o}}} \big(z^{\mathrm{o}} - z\big)^{-1}\,\mathrm{d}z\right) = \infty. \qquad \square$$

Theorem 8 is now a compendium of theorem 14 and propositions 13, 15 and 16. Note that this theorem is stated in terms of the general parameterization of the type variable, so that the canonical parameterization $\gamma_z = z$ which was assumed for convenience at the beginning of this section is not imposed; the formulas for the general case are obtained simply by substituting γ in place of z at the appropriate places.

6. Modeling choices and options

Let us now briefly discuss some of the features of the models that have been considered in this paper, and point out some alternative and additional elements which may be of interest, and should be considered as part of future work.

The starting point for our study was the M/M/m queue with rational abandonments, a utility function based on a linear waiting cost, and a consistent equilibrium solution. As we have seen, for this model abandonments occur either upon arrival or none at all, which is obviously contradictory to our common experience and, perhaps, common wisdom. Within the rational abandonment model, several elements may cause this mismatch:

- Linearity of the waiting cost.
- The queueing model.
- The consistency assumption.

Costs. The assumption of a linear waiting cost is amenable to analysis, but may be lacking an important component. The waiting cost may be reasonably divided into two components: an *alternative* waiting cost and a *psychological* cost. The first reflects the actual value of time, and may be viewed as the amount a customer is willing to pay beforehand for someone else to wait in her place. This component may be argued to be approximately linear. The additional psychological component refers to the subjective feeling of impatience that develops while waiting, and can be argued to be strictly convex. One can check that strictly convex costs will induce abandonments in finite time. The equilibrium analysis, however, may be considerably more difficult and less explicit than in the linear case, and is not available at present.

The second and third points are centered around the shape of the hazard rate function associated with the virtual waiting time. Even for nonlinear waiting costs, and in fact under any abandonment profile, the hazard rate in any M/M/m queue is increasing. As already pointed out, this seems to be at odds with the subjective interpretation of the waiting time distribution. Indeed, excessive waits will often be interpreted by waiting customers as an indication that the system performs below its standard performance, thus leading to a decrease in the subjective hazard rate as perceived by the customer.

The queueing model. In this paper we have approached this discrepancy by assuming that the system actually deviates from the basic M/M/m model. This has been done in the simplest possible way that captures the desired effect of a decreasing hazard rate – namely, the inclusion of a fault state which is hit by arriving customers with certain probability. More involved models of resource deficiency and congestion may be of interest here, such as variable number of servers, varying arrival rates, priorities, and variable number of servers. The latter is the closest one to the model of this paper, and can perhaps be analyzed using similar methods. But either one of these factors tends to decrease the hazard rate in time, as the relative (posterior) weight of possible unfavorable circumstances increases while waiting. We finally note that heavy-tailed service distributions (in an M/G/m queue model) could also lead to decreasing hazard rate functions.

Consistency. An alternative approach for inducing a decreasing hazard rate tail, is to attribute it to the subjective beliefs of customers, which need not coincide with actual system performance. It may be argued that the virtual waiting time distribution in a given system is never learned perfectly by the customers, due to, say, limited experience, variation in time, prior belief, experience with other systems, etc. This is especially relevant for the tail of the distribution, since exceptionally long waiting times are rarely reached. We are thus lead to the concept of a *partially consistent equilibrium*, which may be of independent interest – where the subjective waiting time distribution is influenced by the actual one in some specified manner, but does not necessarily coincide with it. One option may be to specify some parametric form for the subjective distributions, and assume that this parameter is determined by some characteristics (e.g., the mean) of the actual system performance.

We next point out some additional issues that have not been dealt with in the present paper.

Retrials. These are obviously an important issue when abandonments are concerned. Besides their effect on the arrival process, the option of retrial may play a significant role in the abandonment decision. The incorporation of retrials within the rational model is an important subject for future work.

Demand elasticity. An additional concern is the arrival rate, which was assumed constant. In fact, we may expect the system performance (viz. the virtual waiting time) to affect not only the abandonment decisions, but also the decisions of some customers regarding whether to try to approach the system at all. This may be accommodated within the current rational framework, simply by appending some arrival cost to each customer type, and assuming that each customer joins the system only if his utility for approaching the system (and abandoning optimally) surpasses the arrival cost. This would lead the system to stabilize on a new effective arrival rate, but should not affect the uniqueness and structure of the equilibrium.

Real-time decisions. In our model formulation, abandonment times were considered as decision policies which are determined by customers upon arrival. These policies may be easily reinterpreted as real time decisions, which may seem more natural for the problem at hand. Specifically, while waiting a customer continouosly considers whether to abandom immediately, or wait further and possibly abandon at some later time. Once the former becomes preferable, in terms of residual utility, the customer leaves the queue. More formally, consider a z-type customer who has been waiting for t time units in the queue. Let $F_z(\cdot|t)$ denote this customer's subjective distribution on his *remaining* virtual waiting time $V - t$. Possible decisions for this customer are to leave immediately ($T = 0$) or stay, in which case he can leave at any time $T > 0$ in the future. The (residual) utility associated with a T-abandonment would be

$$U_z(T|t) = = E_{z|t}\big(r_z \mathbb{1}\{T \geqslant (V-t)\} - c_z \min\{(V-t), T\}\big).$$

An optimal decision at time t would then be to abandon immediately if $T = 0$ maximizes $U_z(T|t)$, and stay otherwise.

As may be expected, this real-time decision pattern coincides with the initial policy formulation, provided that customers are temporally consistent (cost parameters are not modified, and $F_z(\cdot|t)$ is obtained from $F_z(\cdot)$ via Bayes' rule). The real-time formulation may become useful in more complicated situations, where partial on-line information is supplied to customers concerning their remaining waiting time.

Asymptotic analysis. Queueing theory enjoys some universal laws which are valid under very broad assumptions. An outstanding example is Kingman's discovery [14] that waiting-times in heavily-congested $\mathrm{G}/\mathrm{G}/1$ queues tend to an exponential distribution. This fundamental law has been extended to cover the $\mathrm{G}/\mathrm{G}/\mathrm{m}$ queue [13], and much more. It is of interest to identify analogous universal laws that pertain to customers' patience. (Asymptotic analysis of queues with abandonments has been carried out only under the very restrictive assumptions of the $\mathrm{M}/\mathrm{M}/\mathrm{m}+\mathrm{M}$ queue, namely, exponentially distributed patience; see [7].)

Queueing science. Our paper could be viewed as an initial theoretical step, in an attempt to understand and model the patience (or impatience) of delayed individuals, as reflected in common queueing situations. A natural next step is a validation of the theory, either via laboratory experiments (as in [4]), or real-world measurements (in the spirit of [3,17]). This validation is likely to be followed by refinements or modifications of our theory, until a satisfactory understanding of the phenomenon of abandonment is achieved.

7. Conclusion

This paper suggests a rational decision framework for determining the abandonment times of waiting customers, assuming that these customers have no information regarding their standing in the queue. We focused here on the consistent equilibrium solution, which supposes that customers' expectations regarding their waiting time in the queue coincide with actual system performance. The utility function assumes a marginal waiting cost and service utility which are constant in time, but may vary among customers.

Our main results concern the existence, uniqueness, structure and computation of the equilibrium in the $\mathrm{M}/\mathrm{M}/\mathrm{m}$ queue, and in the extended $\mathrm{M}/\mathrm{M}/\mathrm{m}(q)$ system. In the former case it was shown that, due to an intrinsic increasing hazard rate property, rational decisions are either to leave immediately if not admitted to service upon arrival, or not to abandon at all. By introducing a possible fault state into this basic system, a nontrivial abandonment profile has been obtained in equilibrium.

In both cases, it turns out that the hazard rate function related to virtual waiting time tends to become non-increasing in equilibrium: in the $\mathrm{M}/\mathrm{M}/\mathrm{m}$ case it is (weakly) increasing in general but becomes flat in equilibrium, while in the $\mathrm{M}/\mathrm{M}/\mathrm{m}(q)$ case

it is increasing–decreasing in general but becomes decreasing under the equilibrium abandonment profile. This points to a general tendency which deserves further study.

We have pointed out several directions in which our basic models can and should be generalized. Of immediate interest to us are the incorporation of convex waiting costs, and the generalization of the fault state formulation to queues with more general failure (or congestion) modes. At present it is not clear whether a unique equilibrium exists in these models. The effect of intentionally supplied status information to customers is of great importance in practice, and appropriate methods for its incorporation and investigation within the rational model are yet to be explored.

Naturally, the practical utility and further evolution of the models suggested in this paper need to be evaluated in light of actual applications. A methodology is required to estimate the basic model parameters (and especially the customer parameters) from attainable measurements, and test the predictive capability of this model under varying conditions. All in all, it is apparent that much remains to be done in this area.

Acknowledgements

This research was supported by the Israel Science Foundation founded by the Israel Academy of Sciences and Humanities. The first author's research was supported by the Fund for Promotion of Research at the Technion, and by the Technion V.P.R. Funds – Smoler Research Fund and B.G. Greenberg Research Fund (Ottawa). We wish to thank an anonymous referee for constructive remarks concerning the presentation of the paper.

References

[1] E. Altman and N. Shimkin, Individual equilibrium and learning in a processor sharing system, Oper. Res. 46 (1998) 776–784.
[2] F. Baccelli and G. Hebuterne, On queues with impatient customers, in: *Performance '81*, ed. F.J. Kylstra (North-Holland, Amsterdam, 1981) pp. 159–179.
[3] A.J. Brigandi, D.R. Dargon, M.J. Sheehan and T. Spencer, AT&T's call processing simulator (CPPS) operational design for in-bound call centers, Interfaces 24(1) (1994) 6–28.
[4] Z. Carmon and D. Kahneman, The experienced utility of queuing: Experience profiles and retrospective evaluations of simulated queues, Working paper, Fuqua School Duke University, Durham, NC (1999).
[5] J.E. Cohen and F.P. Kelly, A paradox of congestion in a queuing network, J. Appl. Probab. 27 (1990) 730–734.
[6] C.F. Daganzo, The uniqueness of a time-dependent equilibrium distribution of arrivals at a single bottleneck, Transport. Sci. 19 (1985) 29–37.
[7] O. Garnet, A. Mandelbaum and M. Reiman, Design of large call centers with impatient customers, submitted (1999).
[8] R. Hassin and M. Haviv, Equilibrium strategies and the value of information in a two line queueing system with threshold jockeying, Comm. Statist. Stochastic Models 10 (1994) 415–435.
[9] R. Hassin and M. Haviv, Equilibrium strategies for queues with impatient customers, Oper. Res. Lett. 17 (1995) 41–45.

[10] M. Haviv and Y. Ritov, Homogeneous customers reneg at random times when waiting conditions deteriorate, preprint (August 1999).
[11] Help Desk and Customer Support Practices Report, May 1997 surgery results, Help Desk Institute (www.HelpDeskInst.com).
[12] M.K. Hui and D.K. Tse, What to tell customers in waits of different lengths: an iterative model of service evaluation, J. Marketing 60 (1996) 81–90.
[13] D.L. Iglehart and W. Whitt, Multiple channel queues in heavy traffic, I, Adv. in Appl. Probab. 2 (1979) 150–177.
[14] J.F.C. Kingman, The single server queue in heavy traffic, Proc. Cambridge Philos. Soc. 57 (1961) 902–904.
[15] S. Li and T. Basar, Distributed algorithms for the computation of noncooperative equilibria, Automatica 23 (1987) 523–533.
[16] H. Mendelson and S. Whang, Optimal incentive-compatible priority pricing for the $M/M/1$ queue, Oper. Res. 38 (1990) 870–883.
[17] C. Palm, Methods of judging the annoyance caused by congestion, Tele 2 (1953) 1–20.
[18] J.B. Rosen, Existence and uniqueness of equilibrium points for concave N-person games, Econometrica 33 (1965) 520–534.
[19] C.M. Rump and S. Stidham, Stability and chaos in input pricing for a service facility with adaptive customer response to congestion, Managm. Sci. 44 (1998) 246–261.
[20] M.J. Smith, The existence of a time-dependent equilibrium distribution of arrivals at a single bottleneck, Transport. Sci. 18 (1984) 385–394.
[21] W. Whitt, Improving service by informing customers about anticipated delays, Managm. Sci. 45(2) (1999) 192–207.