Queueing Systems 47, 117–146, 2004

Rational Abandonment from Tele-Queues: Nonlinear Waiting Costs with Heterogeneous Preferences

NAHUM SHIMKIN shimkin@ee.technion.ac.il
Department of Electrical Engineering, Technion, Haifa 32000, Israel

AVISHAI MANDELBAUM avim@tx.technion.ac.il
Department of Industrial Engineering, Technion, Haifa 32000, Israel

Received 4 June 2002; Revised 18 November 2003

Abstract. We consider the modelling of abandonment from a queueing system by impatient customers. Within the proposed model, customers act rationally to maximise a utility function that weights service utility against expected waiting cost. Customers are heterogeneous, in the sense that their utility function parameters may vary across the customer population. The queue is assumed invisible to waiting customers, who do not obtain any information regarding their standing in the queue during their waiting period. Such circumstances apply, for example, in telephone centers or other remote service facilities, to which we refer as *tele-queues*. We analyse this decision model within a multi-server queue with impatient customers, and seek to characterise the Nash equilibria of this system. These equilibria may be viewed as stable operating points of the system, and determine the customer abandonment profile along with other system-wide performance measures. We provide conditions for the existence and uniqueness of the equilibrium, and suggest procedures for its computation. We also suggest a notion of an equilibrium based on sub-optimal decisions, the *myopic* equilibrium, which enjoys favourable analytical properties. Some concrete examples are provided to illustrate the modelling approach and analysis. The present paper supplements previous ones which were restricted to linear waiting costs or homogeneous customer population.

Keywords: tele-queues or invisible queues, abandonment, impatient customers, Nash equilibrium, telephone call centers, contact centers, multi-server queues

AMS subject classification: 90B22, 91A80.

1. Introduction

A rich interplay exists between the performance of a service system and its customer characteristics. Performance is obviously affected by customer characteristics such as arrival rate and service requirements; but this dependence goes both ways, as customer characteristics may be affected by the perceived performance measures, such as the anticipated delay. One obvious relation is the effect that "quality of service" may have on the arrival rate, through the fraction of returning customers and reputation effects. Indeed, a number of studies on queueing systems (e.g., [4,11,14,15]) have incorporated a "demand curve" approach, whereby arrival rate depends on congestion (and possibly also on external pricing). The system operating point must then be determined through an equilibrium analysis, which takes into account the variability of customer characteristics.

Similar observations hold with respect to the abandonment characteristics of impatient customers. We shall focus on the modelling of the customer abandonment profile, or patience, and its dependence on system performance. The relevant aspects of the system performance are captured here by the queueing delay, namely the distribution of the waiting time before admitted to service. The queue is assumed invisible, in the sense that waiting customers do not have any information regarding the queue condition, so that their estimates of the remaining waiting time rely solely on prior beliefs and the elapsed waiting time. To model the presumed dependence of patience on delay, we consider a rational decision model in which each customer seeks to maximise an appropriate utility function. Specifically, the abandonment time is chosen to maximise an individual utility function which weights the utility of the required service with the expected waiting cost. The system equilibrium now determines the abandonment profile. Our main interest here here is in this equilibrium point and its properties – existence, uniqueness, and computation.

The rational viewpoint for abandonment modelling has been considered in several previous studies. The papers [5,7] consider the model with homogeneous preferences, so that the utility functions are identical for all customers. In [5], the authors consider an $\mathrm{M}/\mathrm{M}/1$ queue with linear waiting costs and strict due times for service commencement, and show that the induced equilibrium is a probabilistic split between an immediate abandonment and none at all. In [7], the authors consider the multi-server $(\mathrm{M}/\mathrm{M}/m)$ queue with nonlinear waiting costs, and show that the equilibrium is given by a randomised abandonment time, with identical distribution for all customers. The generalisation to heterogeneous preferences is taken up in [10], also in the context of the $\mathrm{M}/\mathrm{M}/m$ queue but with linear waiting costs.

For the basic queue model and utility function, it is first shown in [10] that an optimal decision for each customer, who encounters all servers busy upon arrival, will always be to either abandon immediately or else wait until being served. In other words, the option of abandonment during wait is never optimal. This follows after observing that the hazard-rate function for the waiting time in such a queue is non-decreasing; hence, with linear waiting cost, as time progresses it only becomes less worthwhile to abandon. As this theoretical result does not conform with reality, the model was modified by adding a fault state, real or subjective, so that with some probability an arriving customer might never get served. This gives rise to a hazard rate function which is eventually decreasing and thus facilitates a non-trivial abandonment profile, which turns out unique under equilibrium conditions.

We revisit here the heterogeneous preferences model for the $\mathrm{M}/\mathrm{M}/m$ queue [10], this time allowing nonlinear waiting costs. Non-trivial abandonment times now arise when the waiting costs are super-linear, without resort to the fault state. Obviously, nonlinear costs allow greater flexibility in modelling different components of the waiting cost function, both from the economic and psychological viewpoints. On the down side, nonlinear costs require more challenging analysis, and might possibly result in non-uniqueness of the equilibrium point.

We mention that a simplified model for adaptive customer patience was considered in [16]. This is a descriptive model which directly describes the aggregate population behavior as a function of a single parameter, the expected waiting cost. That paper also contains a brief discussion of waiting costs and their characteristics in the context of the abandonment problem.

The analysis of the proposed model can be divided into two parts. In the first, we consider the nonlinear-cost model with minor restrictions on the cost structure. In this generality, little can be said about the equilibrium point; in particular, it need not be unique, and a computation procedure is not available. These difficulties arise as the individual utility functions (as induced by the waiting time distribution, which itself needs to be determined) need not be unimodal, and the possibility of local maxima renders the equilibrium analysis intractable. In order to arrive at a tractable equilibrium concept, we allow for suboptimal decisions of the customers, in the form of the *myopic decision rule*: abandon at the first local maximum of the utility function, namely as soon as the utility starts decreasing. The precise definition and a discussion of this sub-optimal decision rule are presented in section 2.3. The equilibrium that is induced by the myopic decision rule is accordingly termed *myopic equilibrium*. We shall establish the uniqueness of the myopic equilibrium under very weak assumptions, and provide computational procedures for its calculation.

The second part of our analysis concerns conditions which guarantee uniqueness for the *global* (as opposed to myopic) equilibrium. Essentially, the required conditions are a complete ordering of the waiting cost functions of the different customer types, and a concavity-like requirement (assumption B2 in section 6) on the marginal waiting costs. Under these conditions it is shown that the global equilibrium exists, is unique, and in fact coincides with the myopic equilibrium.

The paper is organised as follows. In section 2 we present our basic model, including the queueing system, customer abandonment model, and system equilibria. Section 3 presents some preliminary analysis, which includes the characterisation of extreme points of individual utility functions in terms of the hazard rate function associated with the waiting time distribution, and certain properties of this hazard rate which are central in our analysis. Some ideas and difficulties related to equilibrium computation are outlined on this basis. The myopic equilibrium and its properties are explored in section 4. The complementary notion of the *farsighted equilibrium*, which is mainly of interest for computational purposes, is briefly considered in the subsequent section. Section 6 presents the results concerning the global equilibrium, focusing on sufficient conditions for existence and uniqueness. In section 7 we illustrate, through some examples, the computational and modelling scope of our framework. Section 8 discusses the application of our results (which were obtained for a continuum of customer types) to models with discrete types. We conclude, in section 9, with some suggestions for future research directions.

2. The model

We proceed to describe the queueing system, and characterise the abandonment time of a waiting customer in terms of an appropriate utility function. The notions of global equilibrium and myopic equilibrium are then introduced.

2.1. The queueing system

Consider an $M/M/m$ queue, with m servers and Poisson arrivals at rate λ. Service times are i.i.d. and exponentially distributed with mean $1/\mu$. The service discipline is first come first served (FCFS), and the queue size is unlimited.

While waiting in queue, customers may decide to abandon the queue and give up the demanded service. The decision whether to abandon the queue or not and the precise instant of abandonment are determined individually by each customer, based on a decision model which is described next.

2.2. Individual utility and rational decisions

After joining the queue, a customer may abandon at any time $T \geqslant 0$ before being admitted to service ($T = 0$ is the arrival instant). It is assumed that no information is conveyed to the customer during the waiting period regarding the status of the queue and his or her position in it. Thus, an abandonment policy for each customer is simply the time T he or she is willing to wait for service before abandoning the queue.

Observe that a decision to abandon at $T = 0$ differs from not approaching the system at all, as in the former case the customer will not abandon if admitted to a free server upon arrival.

We now define an individual utility function for each customer. Customers will be categorised into different types according to their utility function parameters. Let $z \in Z$ denote the type, with Z the set of possible types. Further, a probability distribution P_Z is prescribed over the set of customer types, so that the type z of an arriving customer is randomly and independently determined according to P_Z.

A customer of type z is characterised by the pair (R_z, C_z), where

(i) $R_z(t)$, the *service utility* function: $R_z(t)$ is the utility (or reward) which the customer expects to obtain by entering service, having waited t time units beyond arrival to the queue. We assume that $R_z(t)$ is strictly positive and continuous in t. One naturally expects R_z to be non-increasing, but we do not need to impose that assumption.

(ii) $C_z(t)$, the *waiting cost* function: $C_z(t)$ is the disutility of a customer who waits in queue for t time units. Let $C'_z(t) := \mathrm{d}C_z(t)/\mathrm{d}t$ denote the marginal waiting cost function. We assume that $C_z(t)$ is positive and increasing in t, and that $C'_z(t)$ is continuous in t.

Besides these cost parameters, a customer's abandonment time will also depend on a third quantity:

(iii) $F_z(\cdot)$, a probability distribution on $[0, \infty)$, which reflects the customer's belief about the *offered waiting time* V, namely, the time he or she would have to wait in queue (without abandoning) before being admitted to service.

Observe that F_z is in general a subjective quantity, conceived by each customer based on prior experience, beliefs, and relevant information. In this paper we shall impose the following assumption, that underlies the definition of system equilibrium.

Consistency assumption. For each customer type $z \in Z$, the subjective distribution F_z coincides with the *actual* distribution of the offered waiting time, which we denote by F.

Since $F_z \equiv F$, and we omit the subscript z from F_z. Implicit in the above assumption is the requirement that the virtual waiting time distribution for an arriving customer will be well defined and stationary. Indeed, this will be a consequence of assumption A1 below, which implies that all customers have finite patience, hence that the system is stable.

Define the cost-to-reward ratio, or simply the *cost-ratio*, as

$$\gamma_z(t) \triangleq \frac{C_z'(t)}{R_z(t)}, \quad t \geqslant 0.$$

The function γ_z will play a key role in our analysis. It is reasonable to expect that $\gamma_z(t)$ is non-decreasing in t, but this will not be imposed.

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

$$\begin{aligned} U_z(T) &= E\big(R_z(T)\mathbf{1}\{T \geqslant V\} - C_z\big(\min\{V, T\}\big)\big) \\ &= \int_{0-}^{T} R_z(t)\,\mathrm{d}F(t) - \int_{0-}^{\infty} C_z\big(\min\{t, T\}\big)\,\mathrm{d}F(t), \end{aligned} \tag{2.1}$$

where E stands for the expectation with respect to the distribution F of the offered waiting time V. Note that F may include a point mass at the origin, thus representing the probability of finding a free server immediately upon arrival. Therefore, $U_z(0) = R_z(0)F(0)$.

Denote by T_z the abandonment time of a type-z customer (we assume that all customers of the same type chose the same T). The *rational* choice of T_z is the value T which maximises the utility function $U_z(T)$, over $T \geqslant 0$. If the maximiser is not unique, a specific one may be assigned arbitrarily.

A note about randomised choices is in order here. In case the maximiser T_z in not unique, then any probability distribution over the maximising set of $U_z(T)$ may in fact be chosen; this corresponds to the game-theoretic concept of randomised (or mixed) strategies, which are often required to ensure existence of the Nash equilibrium. In this paper we shall not require randomised choices, as we assume a continuum of user types

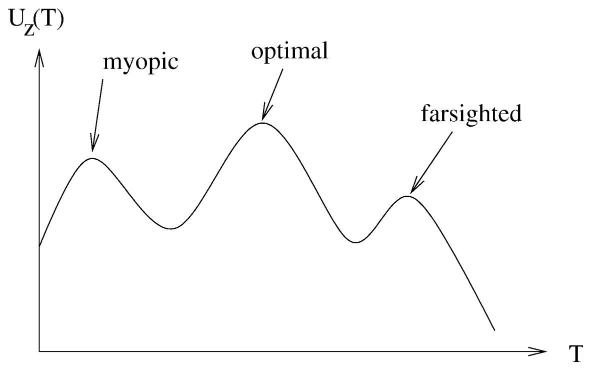

Figure 1. Global and local optima of the utility function.

(assumption A2 below). Existence of equilibrium in pure (non-randomised) choices is thereby facilitated. This may be contrasted with the model of [7], where equilibrium with a single customer type is inherently randomised.

2.3. *Myopic decisions*

So far, we have defined the optimal (or rational) choice for each customer as the abandonment time which globally maximises this customer's utility function. In the following we shall consider also *locally* optimal choices, namely the possibility of abandonment at local maxima of the utility function. Of particular interest will be the notion of myopic choices.

The *myopic decision rule* chooses the abandonment time as the *first* local maximum of the utility function. By convention, we refer to a *weak* local maximum in this definition. Thus, the myopic decision is the smallest time T at which the utility function $U_z(T)$ is *not* strictly increasing (see figure 1). The concept of myopic decisions, and the induced myopic equilibrium, will prove most useful in the analysis to follow.

As a solution concept of independent interest, the myopic decision rule may be advocated on the following grounds:

1. It is plausible that abandonment decisions are taken online (cf. [10]), based on the customer's assessment of the current situation and the utility of further wait. The online choice at each point of time is then whether to wait "a little longer" or abandon immediately. Such considerations would indeed lead to abandonment as soon as the utility starts declining.

2. Customers may lack precise information regarding the waiting time distribution (or its hazard rate) for long waits, especially if they are inclined to abandon earlier times. They may therefore base their assessment of the utility of longer waits on their short wait experience, and will tend to extrapolate a local utility decrease.

2.4. System equilibria

Suppose we are given an *abandonment profile* $\mathcal{T} = \{T_z\colon z \in Z\}$, which assigns an abandonment time T_z to each customer type z. Together with the type distribution P_Z, this determines the patience function, namely a probability distribution G on the abandonment time T of an arriving customer. Specifically,

$$G(t) \triangleq \mathit{Prob}\{T \leqslant t\} = P_Z\{z\colon T_z \leqslant t\}, \quad t \geqslant 0. \tag{2.2}$$

Equivalently, the survival function associated with G is $\overline{G}(t) \triangleq \mathit{Prob}\{T > t\} = P_Z\{z\colon T_z > t\}$.

Assuming rational choices, we can now define a mapping from the set of abandonment profiles into itself. We have just seen how $\mathcal{T}$ determines the patience function G. Given G, the system is an $\mathrm{M}/\mathrm{M}/m + G$ queue, in which one can calculate the distribution function F of its offered waiting time in steady state [2,6]. Invoking the consistency assumption described above now yields the utility function (2.1) of each customer type. The mapped-into profile is finally given as the collection of optimal points of the respective utilities.

A *system equilibrium* is a fixed point of this map. Under rational decisions this coincides with the Nash equilibrium as each customer is maximising his or her utility given the choices of all others. For concreteness we refer to this equilibrium as a *global* equilibrium point. Note that one can use a similar procedure to that of the previous paragraph to define a mapping from the set of offered waiting time distributions onto itself, and the equilibrium point is then equivalently defined as a fixed point of this map. This latter definition will be useful in the ensuing analysis.

A *myopic* equilibrium is defined similarly, except that the abandonment time of each customer is determined according to the myopic decision rule.

2.5. Additional assumptions

In order to ensure that each customer eventually abandons, we shall make use (in section 3.2) of the following assumption.

Assumption A1 (All-leave). For each $z \in Z$, the cost ratio γ_z satisfies $\liminf_{t\to\infty} \gamma_z(t) > m\mu$.

Assumption A1 will be imposed throughout this paper, without further mention. We note that much of the analysis below can be carried out under the weaker condition of queue stability, namely $\lambda\overline{G}(\infty) < m\mu$ [2]. From assumption A1 we will deduce that $\overline{G}(\infty) = 0$, so that stability is trivially implied. However, the implication that all customers eventually abandon is quite natural for reasonable customers, and, furthermore, it simplifies some of our arguments and computational procedures.

To establish existence of a (myopic) equilibrium with a continuum of types, we shall require certain continuity properties of optimal decisions. In particular, we need to prevent a small change in the waiting time distribution from resulting in a sharp change in the customer abandonment profile (or patience distribution). For that purpose, we shall require that the cost-ratio curves $\gamma_z(\cdot)$ will not be too concentrated around one point or curve. This is made precise as follows.

Assumption A2 (Continuity). There exists a constant $K > 0$ such that the following holds. For any continuous function $h(t)$, $t \geqslant 0$, and any $\epsilon \geqslant 0$,

$$P_Z\Big\{z\colon \sup_{t\geqslant 0}\big[\gamma_z(t) - h(t)\big] \in [-\epsilon, \epsilon]\Big\} \leqslant K\epsilon. \tag{2.3}$$

The functions h can be further restricted to be non-decreasing, with range in $(0, m\mu]$, and with derivative bounded by $\dot{h} \leqslant (m\mu)^2$. Indeed, assumption A2 is used in lemma 4.3, where h stand for the hazard rate function $H(t)$ and may inherit its properties.

The following observations apply to assumption A2.

1. The probability in (2.3) is taken over those types z for which γ_z is upper-bounded by $h + \epsilon$, while $\gamma_z(t)$ is ϵ-close to $h(t)$ at some point t. That is, γ_z enters a sleeve of size ϵ around h, but does not exceed it.
2. Assumption A2 implies, in particular, that the probability of any single type z is null. This follows by taking $h(t) = \gamma_z(t)$. Obviously, then, the set of types Z cannot be discrete.
3. Moreover, with $\epsilon = 0$, we obtain that

$$P_Z\Big\{z\colon \sup_{t\geqslant 0}\big[\gamma_z(t) - h(t)\big] = 0\Big\} = 0.$$

The last set consists of those functions that touch h from below. Refer to a set of functions that satisfies this condition (with some h) as an *exposed* set. Any exposed set must have zero probability: that is the essence of assumption A2.

An intrinsic characterisation of an exposed set can be simply given by taking h as the upper envelope (the supremum at each t) of the functions in that set. Put another way, any function in an exposed set must be larger than all others at some point t.

To illustrate, the set $\{\gamma_z(t) = 1 - (t - z)^2\colon z \in X\}$ is exposed (for any $X \subset \mathbb{R}$), since each γ_z is undominated from above at $t = z$. An appropriate "test function" here is $h(t) = 1$. On the other hand, a set which consists of mutually dominated functions cannot be exposed, unless it is a singleton. Thus, if $\{\gamma_z\}$ consists of mutually dominated functions, as assumed in section 6, then the requirement is simply that any single function (or type) will have zero probability. This also holds when $\{\gamma_z\}$ consists of a finite union of sets of mutually dominant function; see example 3 in section 7.

In the special case where the functions $\gamma_z(t)$ are all linear in t, a subset $\{\gamma_z\}$ is exposed if and only if all lines $\gamma_z(t)$ in it are tangent to a single convex function $h(t)$. Indeed, the upper envelope (or supremum) of a set of linear functions is convex, and any line that touches it from below is tangential to it.

4. As noted, the set of types Z cannot be discrete (hence finite) under assumption A2. However, any finite set of types can be slightly perturbed so as to satisfy this continuity assumption. This will be further discussed in section 8.

3. Preliminary analysis

We collect in this section a few properties that will be used in the subsequent analysis. First we show (following [10]) that the extremal points of the utility function occur at those points in time when the hazard-rate of the offered waiting-time equals the cost ratio. We follow with some basic queueing relations for the M/M/$m + G$ queue that lead to a key differential relation for the hazard rate function. We then provide a brief preview of equilibrium computation.

3.1. Extremal points of the utility function

Recall that F is the distribution of the offered waiting-time, as perceived by all customers. Suppose that $F(t)$ is continuously differentiable for $t > 0$ (that is, it has a continuous density F', except possibly for a point mass at $t = 0$), that F' has a right-limit at 0, and that $F(t) < 1$ for all $t < \infty$. (These properties indeed follow from the expression (3.3) for F' that holds in our queue.) Differentiating the utility function (2.1) with respect to $T > 0$ gives

$$U_z'(T) = R_z(T)F'(T) - C_z'(T)\overline{F}(T) = R_z(T)\overline{F}(T)\big[H(T) - \gamma_z(T)\big], \tag{3.1}$$

where $\gamma_z = C_z'/R_z$ is the cost-ratio defined previously, $\overline{F} = 1 - F$ denotes the survival function of F, and H is the hazard rate function associated with the offered waiting time distribution F, namely

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

We define $H(0) = H(0+)$. The first order condition for a local extremum of $U_z(T)$ at $T > 0$ is $U_z'(T) = 0$, which is is equivalent to

$$H(T) = \gamma_z(T). \tag{3.2}$$

For a (strict) local maximum, $H - \gamma_z$ should change sign from positive to negative at T. In general there may be several local extrema of the utility function. Hence, a local characterisation is not sufficient to establish the global maximum. This accounts for much of the difficulty in the analysis of the (Nash, or global) equilibrium and its properties in this model. As we shall see, the notion of the myopic equilibrium works around these difficulties.

3.2. Basic queueing relations

Consider the M/M/m queue with a given patience distribution $\overline{G}(t)$. Under the stability condition $\lambda\overline{G}(\infty) < \mu$, the probability density function of the offered waiting time at $t > 0$ is given by [2,6]:

$$F'(t) = \lambda\pi_{m-1}\exp\left(-\int_0^t I(s)\mathrm{d}s\right), \quad t > 0 \tag{3.3}$$

where π_{m-1} is a normalisation constant, and

$$I(t) = m\mu - \lambda\overline{G}(t). \tag{3.4}$$

By differentiating $H = F'/\overline{F}$, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}H(t) = H(t)\big[H(t) - I(t)\big], \quad t > 0. \tag{3.5}$$

The following properties of the hazard rate function $H(t)$ will prove to be useful.

Lemma 3.1.

(i) $H(t)$ is non-decreasing in t. Furthermore, it is strictly increasing up to the point $t_0 \in [0, \infty]$, where $G(t_0) = G(\infty)$, and constant thereafter: $H(t_0) = H(\infty)$.

(ii) $H(\infty) = I(\infty) \leqslant m\mu$. If all customers have finite patience, namely $\overline{G}(\infty) = 0$, then

$$H(\infty) = I(\infty) = m\mu.$$

Proof. Part (i) is from [10], proposition 3.2. As for (ii), note that by its definition in (3.4), $I(t)$ is non-decreasing and upper bounded by $m\mu$, hence converges to some limit $I(\infty)$. It is easily verified (either analytically from (3.3), or simply by noting that $H(t)$ cannot be larger than the service completion rate mμ) that $H(t)$ is upper bounded by $m\mu$, hence converges to a finite limit $H(\infty)$. Invoking (3.5) again it follows that $H(\infty) = I(\infty)$. It is further seen from (3.4) that $I(\infty) = m\mu$ when $\overline{G}(\infty) = 0$. □

Note that $H(\infty) = m\mu$ provides a terminal condition for the differential relation (3.5). The finite patience condition which is required for this equality is enforced in our decision model through assumption A1 above. Indeed, since $H(t) \leqslant m\mu$, it follows from (3.1) and A1 that the utility function $U_z(t)$ is strictly decreasing for t large enough, hence the optimal decision (and, similarly, the myopic and far-sighted ones) is always to abandon at some finite time: $T_z < \infty$, for every type z.

3.3. Preview of equilibrium computation

Our plan is to use the differential relation in (3.5) as a starting point to establish key properties of the equilibrium such as existence and uniqueness, and as a means to compute the equilibrium point. Referring to the equilibrium problem introduced in the previous section, the patience function G is not a-priori given but rather determined by the customer decision profile. Hence $I(\cdot)$ is unknown, and must be determined together with $H(\cdot)$. In fact, at each point t, $I(t)$ may in general depend on the entire function $H(\cdot)$, so that (3.5) is a functional differential equation which may be quite intractable.

In order to be able to directly integrate (3.5) to compute both $H(\cdot)$ and $I(\cdot)$, one of the following properties would be required:

(F) $I(t)$ is a function of $\{H(s),\ s \leqslant t\}$. In that case we could integrate (3.5) forward in t.

(B) $I(t)$ is a function of $\{H(s),\ s \geqslant t\}$. In that case we could integrate (3.5) backward in t.

Unfortunately, for the optimal decision rule neither one of these needs to hold. We will, however, show that under the myopic decision rule, property (F) does hold. This will provide the key to the analysis of the myopic equilibrium in section 4. Later we shall also consider the complementary concept of *farsighted* equilibrium, for which property (B) is applicable. A somewhat more intricate argument will be required for the analysis of the global equilibrium.

4. Myopic equilibrium analysis

We consider here the system equilibrium under the assumption that all customers follow the myopic decision rule. Our main result concerning the related equilibrium point is the following. Recall that assumption A1 is imposed throughout.

Theorem 4.1.

(i) A myopic equilibrium is unique.

(ii) Assume A2. Then the myopic equilibrium exists.

To formulate the proof, let us first collect the basic relations that apply to our model under myopic decisions. Assume that the system is in myopic equilibrium. Recall that a myopic customer abandons at the first local maximum of the utility function. Equivalently, a customer abandons as soon as the marginal utility function $U_z'(t)$ becomes nonpositive. Recall further that the utility function of a type-z customer is given by (2.1). Since the sign of $U_z'(t)$ is the same as that of $H(t) - \gamma_z(t)$, as seen in (3.1), then the myopic decision rule can be expressed as follows: abandon as soon as $H(t) \leqslant \gamma_z(t)$. Equivalently, T_z is the smallest t for which $H(t) \leqslant \gamma_z(t)$. Since H is continuous, this means:

- If $H(0) < \gamma_z(0)$, abandon immediately at $t = 0$.
- Otherwise, abandon as soon as $H(t) = \gamma_z(t)$.

A critical observation is that the question of whether or not a customer abandons by time t depends only on $\{H(s),\ s \leqslant t\}$. Specifically,

$$G(t) = P_Z\{z\colon\ T_z \leqslant t\} = P_Z\bigl\{z\colon\ H(s) \leqslant \gamma_z(s)\text{for some } s \leqslant t\bigr\}. \tag{4.1}$$

This expression may be substituted in (3.4) to give

$$I(t) = (m\mu - \lambda) + \lambda P_Z\bigl\{z\colon\ H(s) \leqslant \gamma_z(s) \text{ for some } s \leqslant t\bigr\}. \tag{4.2}$$

Observe further that $H(\infty) = I(\infty) = m\mu$ (see lemma 3.1 and the subsequent discussion). Combined with (3.5), we obtain the following characterisation for $H(t)$:

$$\frac{\mathrm{d}}{\mathrm{d}t}H(t) = H(t)\bigl[H(t) - I(t)\bigr], \quad t > 0, \tag{4.3}$$

$$H(\infty) = m\mu, \tag{4.4}$$

with $I(t)$ given by (4.2). For future reference we note that $I(t)$ is non-decreasing (by its definition). Recall also that $H(t)$ is monotonically non-decreasing and bounded in $0 \leqslant H \leqslant m\mu$ (lemma 3.1).

As $I(t)$ depends only on $H(s)$ for $s \leqslant t$, we can in principle forward integrate the differential relation (4.3) to obtain H. This proceeds as follows:

1. Choose a candidate $H(0)$.
2. Compute $I(0) = (m\mu - \lambda) + \lambda P_Z\{z\colon\ H(0) \leqslant \gamma_z(0)\}$.
3. Use the differential equation (4.3) and the expression (4.2) for $I(t)$ to compute $H(t)$ from $t = 0$ to ∞.
4. Now check $H(\infty)$. If it equals $m\mu$ then this is an equilibrium solution, otherwise "try" another $H(0)$.

We can obviously restrict attention to those solutions $H(t)$ that satisfy the above-noted properties of the hazard rate function, namely: non-decreasing in t, and bounded in $0 \leqslant H \leqslant m\mu$. The questions of existence and uniqueness of the equilibrium now reduce to the following:

- Existence: does there exist $H(0)$ so that the procedure above yields $H(\infty) = m\mu$?
- Uniqueness: can there be more than a single such $H(0)$?

The key to uniqueness is in the following monotonicity property.

Lemma 4.2. Let $H(t, H_0)$ be a solution of (4.2)–(4.3) with initial conditions $H(0) = H_0$. Restrict attention to those values of $H_0 > 0$ for which $H(\cdot)$ is non-decreasing and bounded. Then, for each $t > 0$, the difference $H(t, H_0) - H_0$ is monotone increasing in H_0.

Proof. Denote $\dot{H} = \mathrm{d}H/\mathrm{d}t$. As we restrict attention to non-decreasing H, the right-hand side of (4.3) is non-negative; thus $\dot{H}(t)$ is increasing in $H(t)$ and decreasing in $I(t)$. Starting at $t = 0$, observe that $I(0)$ depends on $H(0)$ alone and is non-decreasing in it.

Thus, (4.3) implies that $\dot{H}(0)$ increases in $H(0)$. It can further be seen that, for any t, if H increases over an entire interval $[0,t]$ then $I(t)$ decreases, hence $\dot{H}(t)$ increases. From that we can conclude that if $H(0)$ increases then so does $\dot{H}$ on any interval $[0,t]$, hence $H(t)-H(0)$ increases. □

Taking the limit in t, it follows that $H(\infty)-H(0)$ is increasing in $H(0)$, hence $H(\infty)$ is *strictly* increasing in $H(0)$. This immediately implies that there exists at most one function H of the required form that satisfies $H(\infty)=m\mu$. This established the *uniqueness* of the myopic equilibrium.

We turn next to the proof of existence of a myopic equilibrium, under the continuity assumption A2.

Lemma 4.3. Assume A2. Then

(i) For any $H(0)\in(0,m\mu)$, the differential equation (4.2)–(4.3) has a unique solution $H(t)$, which extends at least up to the point where $H(t)=0$ or $H(t)=m\mu$.

(ii) For each $t>0$, $H(t)$ is continuous in $H(0)$.

Proof. Both properties follow by standard results on differential equation, by showing that $I(t)$ satisfies an appropriate Lipschitz condition. Observe, from (4.2), that $I(t)=I(H(s),\ s\leqslant t)$. We start by showing that $I(t)$ is Lipschitz continuous in its argument $(H(s), s\leqslant t)$, with respect to the sup-norm. Fix t and H, and let H_ϵ be an ϵ-perturbation of H on $[0,t]$, namely,

$$\left|H_\epsilon(s)-H(s)\right|\leqslant\epsilon,\quad s\leqslant t.$$

Define $I_\epsilon(t)$ similarly to $I(t)$, but with respect to H_ϵ. Then

$$\begin{aligned}&\left|I_\epsilon(t)-I(t)\right|\\&\quad=\lambda\left|P_Z\left\{z\colon\ H_\epsilon(s)\leqslant\gamma_z(s),\ \text{some } s\leqslant t\right\}-P_Z\left\{z\colon\ H(s)\leqslant\gamma_z(s),\ \text{some } s\leqslant t\right\}\right|\\&\quad\leqslant\lambda P_Z\left\{z\colon\ \gamma_z(t)\leqslant H(s)+\epsilon\text{ for all } s\leqslant t,\gamma_z(t)>H(s)-\epsilon\text{ for some } s\leqslant t\right\}\\&\quad\leqslant\lambda K\epsilon,\end{aligned}$$

where the last inequality follows directly by assumption A2, with $h(t)=H(t)$.

Using this functional bound in the differential equation (4.3), the stated properties follow as in standard results for ordinary differential equations under a uniform Lipschitz condition. See, e.g., [8], theorem 2.2.1 (Lipschitz uniqueness theorem) and theorem 3.1.1 (continuity in initial conditions). □

While $H(t)$ is continuous in $H(0)$ for any finite t, this property fails to hold for the limiting value $H(\infty)$. Still, using the limiting properties of $I(t)$ we show next that the required final value of $m\mu$ is obtained for some initial conditions.

Lemma 4.4. Assume A2. Then there exists a solution $H_0(t)$ to (4.2)–(4.3) which is positive non-decreasing and with $H_0(\infty)=m\mu$.

Proof. Observe first that the solutions $H(t)$ of (4.2)–(4.3) (with any initial condition $H(0) > 0$) have the following properties:

(a) Once $\dot{H}$ becomes strictly negative, equivalently $H - I < 0$, it remains that way (since then H is decreasing while I non-decreasing, hence $H - I$ remains negative), and eventually decreases to $\dot{H}(\infty)$=0.

(b) Once $H(t)$ becomes larger than $I(\infty) = m\mu$, it increases to ∞. This actually happens in finite time (since $\dot{x} = x^2$ blows up in finite time).

(c) The only remaining option is $H(\infty) = I(\infty)$.

Hence, we have that:

- For $H(0)$ small, $H(\infty) = 0$.
- For $H(0)$ large, $H(t) \to \infty$ (in finite time).
- In between, we may have a *single* point $H(0)$ so that $H(\infty) = \infty$. Uniqueness of such $H(0)$ follows from the monotonicity property in lemma 4.2.

We can now establish the existence of the required solution. Let $H(t; h)$ denote the solution that corresponds to initial conditions $H(0) = h$. Let K be the set of initial conditions for which the solution blows up: $K = \{h\colon \lim_{t\to\infty} H(t; h) = \infty\}$. Let k be then infimum of this set; note that $k > 0$. We first claim that $k \notin K$. Indeed, by monotonicity (cf. lemma 4.2) we know that the solution $H(t; k)$ is a lower bound on any other solution that blows up. If $H(t; k)$ blows up (namely $k \in K$), then at some time s we have $H(s; k) = m\mu + 2$. We also know that for $H(0) = h$ small enough we have $H(s; h) < m\mu$. Hence, by the continuity property of the last lemma there exists some $h_0 < k$ for which $H(s; h_0) = m\mu + 1$. But this solution must also blow up (as it is above the threshold $m\mu$ as discussed above), so that k is not the infimum of K.

It follows then that $H(t; k)$ does not blow up. On the other hand, we know that any solution $H(t; h)$ with $h \in K$ is monotone increasing, hence $H(t; h) \geqslant k$ for all t and $h \in K$. Invoking again the continuity property of lemma 4.3, if follows that $H(t; k) \geqslant k > 0$ for all t. As noted before this implies that $H(\infty; k) = I(\infty) = m\mu$, and the claim is satisfied with $H_0(\cdot) := H(\cdot; k)$. □

We can now conclude the existence proof. Given the function $H_0(t)$ of the last lemma, define the abandonment time for type z customers according to the myopic rule, $T_z = \min\{t\colon H_0(t) \leqslant \gamma_z(t)\}$. To show that this is an equilibrium profile, we need to show that this leads to a hazard rate function $H(t)$ which coincides with $H_0(t)$ from which we started. Indeed, the above decision profile induces the patience function $G(t)$ as per (4.1) (with $H(t) := H_0(t)$). The waiting time distribution in now given by (3.4) and (3.3). Using the fact that $H_0(t)$ by its definition satisfies (3.5) and $H_0(\infty) > 0$, it may be verified by substitution and integration that

$$\overline{F}(t) := \int_t^\infty F'(s)\,\mathrm{d}s = H_0(t)^{-1}F'(t)$$

hence $H(t) := F'(t)/\overline{F}(t) = H_0(t)$, as required. □

We finally summarise the computational procedure for the myopic equilibrium that falls off the previous analysis. This computation involves the forward integration of the differential equation (4.2)–(4.3), and a search procedure on $H(0)$ so that $H(\infty) = m\mu$. Specifically:

- Starting with some arbitrary $H(0)$, calculate $H(t)$.
- If $H(t)$ becomes larger than $m\mu$, then stop; $H(0)$ should be increased.
- If $\dot{H}(t)$ becomes non-positive, then stop; $H(0)$ should be decreased.
- Otherwise, $H(t) \to m\mu$ and this is the correct $H(0)$.

5. The farsighted equilibrium

We shall consider briefly an additional notion of a local equilibrium, the farsighted equilibrium, which is complementary to the myopic one. While this equilibrium can hardly be justified as a reasonable solution concept, it will turn out to be useful computationally.

The *farsighted decision rule* selects the abandonment time as the *last* (largest) local maximum of the utility function. More precisely, if $U_z(t)$ is the utility function, then the farsighted decision is the largest time t at which $U_z'(t) \geqslant 0$ (and $T_z = 0$ if such t does not exist). Since assumption A1 is in effect, so that $U_z'(t) < 0$ for t large enough (as indicated at the end of section 3.2), it follows that T_z is the largest time at which $U_z'(t) = 0$. The farsighted equilibrium is the system equilibrium induced by the farsighted decision rule.

As in the case of the myopic equilibrium, we can formulate this decision rule in terms of the hazard rate function. Indeed, T_z is the largest time t for which

$$H(t) = \gamma_z(t)$$

(and $T_z = 0$ if this inequality is nowhere satisfied).

It follows that the question of whether a customer abandons at time t according to the farsighted decision rule depends only on the *future* values of the hazard rate function, namely on $\{H(s), s \geqslant t\}$. That will allow us to integrate the differential equation for $H(t)$ backwards in time. To that end, note that,

$$\overline{G}(t) = P_Z\{z\colon\ T_z > t\} = P_Z\big\{z\colon\ H(s) = \gamma_z(s) \text{ for some } s > t\big\} \tag{5.1}$$

and

$$I(t) = m\mu - P_Z\big\{z\colon\ H(s) = \gamma_z(s) \text{ for some } s > t\big\}. \tag{5.2}$$

This expression for $I(t)$ may be combined with the differential equation (4.3) and terminal condition (4.4) to compute the farsighted equilibrium, using backward integration from $t = \infty$. In contrast to the myopic equilibrium, no search procedure is required here, as the terminal conditions actually serve as explicit initial conditions for backward integration.

Note that if T_0 is an upper-bound on the abandonment times of all customer types, then backward integration may start from T_0 with terminal condition $H(T_0) = I(T_0) =$

$m\mu$ (see lemma 3.1). Such an upper bound is given for instance by the minimal time T for which $\gamma_z(T) \geqslant m\mu$ for all z.

Existence and uniqueness of the farsighted equilibrium may be established similarly to the myopic equilibrium. In fact, uniqueness is easier to establish here since the search procedure for $H(0)$ is not required. We shall therefore not dwell on the details of these proofs.

We close this section by pointing to some useful relations between the two types of "local" equilibria – the myopic and the farsighted – and the global equilibrium. First note that either one of the local equilibria may turn out to be a global one. In the myopic equilibrium, for example, this would be the case if for each customer type the myopic decision rule is in fact the optimal one, namely if the first local maximum of the equilibrium utility function turns out to be the global maximum. As this may not be simple to verify directly, the following observation may be useful.

Recall that the myopic decision rule selects the first local maximum of the utility function, and the farsighted rule selects the last local maximum. Obviously, if the two coincide for a given utility function, then this function is unimodal and both are, in fact, a global maximum. This leads to the following straightforward but potentially useful result.

Proposition 5.1. Suppose the myopic and farsighted equilibria coincide. Then they constitute a global equilibrium as well.

Indeed, recall that the myopic decision rule selects the first local maximum of the utility function, and the farsighted rule selects the last local maximum. Obviously, if the two coincide for a given utility function, then this function is unimodal (it has exactly one local maximum) and this local maximum is in fact a global maximum. The converse is also trivially true: if in a global equilibrium the utility function of each customer is unimodal, then this equilibrium coincides with both the myopic and the farsighted equilibria. Since each of the latter is unique, there may exist at most one global equilibrium with the property that the utility function for each customer type is unimodal.

An interesting conjecture is that the condition of the last proposition also leads to the *uniqueness* of the global equilibrium. This is yet to be verified or disproved. In the forthcoming section we shall formulate explicit conditions on the problem data that guarantee the uniqueness of the global equilibrium.

6. Global equilibrium

For the general model discussed so far, uniqueness of the global equilibrium is not guaranteed. In the present section, we shall impose additional conditions that guarantee the uniqueness of the global equilibrium. Further, as this unique global equilibrium coincides with the myopic and farsighted ones, it can be computed using the procedures outlined above for these equilibria. The imposed conditions are essentially that customer types will be strictly ordered in terms of their cost functions, and the marginal waiting cost will satisfy certain concavity properties.

Our first assumption concerns the complete ordering of customer types according to their cost functions.

Assumption B1 (Ordering). There exists a (complete) order on the set Z of customer types, so that for any $y, z \in Z$ with $y < z$, at least one of the following holds:

(i) $R_y(t) \geqslant R_z(t)$ and $C'_y(t) < C'_z(t)$, for all $t \geqslant 0$.

(ii) $R_y(t) = R_z(t)$ and $\gamma_y(t) < \gamma_z(t)$, for all $t \geqslant 0$.

The following three remarks concern this last assumption.

1. Assumption B1, together with (3.1), implies that $U'_y(t) > U'_z(t)$ (for any F).
2. Since $\gamma_z = C'_z/R_z$, either condition in B1 implies that $\gamma_y(t) < \gamma_z(t)$.
3. The equality $R_y(t) = R_z(t)$ in the second condition is actually required to hold up to a constant scaling factor. (Indeed, the utility function U_z may be rescaled without affecting the optimal decision, and such rescaling is obtained by scaling both the service reward R_z and the waiting cost function C_z by the same factor. Note that this does not affect the cost-ratio γ_z.) For example, it holds if each R_z is constant in time.

In addition, we shall impose the following requirement on the cost-ratio functions.

Assumption B2 (Concavity). $\ddot{\gamma}_z(t) \leqslant \dot{\gamma}_z(t)\gamma_z(t)$ for any z and t such that $\dot{\gamma}_z(t) \geqslant 0$.

An obvious sufficient condition for B2 is that $\gamma_z(t)$ is (weakly) concave in t, namely $\ddot{\gamma}_z(t) \leqslant 0$. We note that this condition is slightly stronger than the one that appears in [7] (in relation to the homogeneous customer problem). Indeed, the latter requires the function $\gamma(t) - \dot{\gamma}(t)/\gamma(t)$ to be monotone increasing, which can be differentiated to give $\ddot{\gamma} \leqslant \dot{\gamma}\gamma + (\dot{\gamma})^2/\gamma$. We conjecture that the same condition would suffice here, however, the details of the proof would be more involved.

The continuity condition A2 will be also used here. Under the order condition B1, the functions h in A2 can be restricted to the set $\{\gamma_z\}$. The requirement is, essentially, that the set of functions $\{\gamma_z\}$ will be sufficiently "spaced apart" under P_z.

The following theorem summarises the main results of the present section.

Theorem 6.1. Assume A2 (continuity), B1 (ordering) and B2 (concavity). Then the global equilibrium is unique, and coincides with both the myopic and farsighted equilibrium points.

The proof is presented in the remainder of this section. Let us first sketch the main ideas in the proof of uniqueness. Our goal is to show that, at any global equilibrium, each utility function $U_z(t)$ is unimodal; hence a global equilibrium coincides with the myopic equilibrium (and also with the farsighted one), whose uniqueness has already been established. To show unimodality, we note that at any (local) maximum t of U_z we have $U'_z(t) = 0$, with U'_z increasing; since U'_z is sign-equivalent to $H - \gamma_z$, these properties are shared by the latter. To show that there exists at most one point with these properties, it would suffice that γ_z is concave in t (by assumption), while H is convex, so that the difference $H - \gamma_z$ is convex. Unfortunately, H is *not* a convex function; indeed, it is increasing and upper bounded, hence must be concave on some part of its domain. More refined analysis is therefore required. A basic observation will be that H does in fact exhibit convex-like properties on those parts of the time axis at which no abandonment occurs, namely that contain no points from the decision profile $\{T_z\}$ (see lemmas 6.3 and 6.4 below). Together with some monotonicity properties, this will lead to uniqueness of the maximum point for each U_z, and hence to the uniqueness of the global equilibrium.

For the rest of this section we impose, without further note, the conditions of the last theorem.

Lemma 6.2. Let $\{T_z\}$ be an equilibrium profile corresponding to a global equilibrium. Then for any $y, z \in Z$, $y < z$ implies that $T_y \leqslant T_z$. Furthermore, if $T_z > 0$ then $T_y < T_z$.

Proof. As noted, assumption B1 together with (3.1) imply that $U_y'(t) > U_z'(t)$. Recalling that T_z is a global maximum of U_z over $t \geqslant 0$, this implies the assertion of the lemma. Indeed, for any $t < T_y$ we have $U_z(t) - U_z(T_y) < U_y(t) - U_y(T_y) \leqslant 0$, implying that such t cannot be a maximum of U_z, hence $T_z \geqslant T_y$. Finally, if $T_z > 0$ then $0 = U_z'(T_z) < U_y'(T_z)$, hence T_z is not a maximum of U_z, namely $T_y \neq T_z$. □

Lemma 6.3. Let $\gamma(t) > 0$ satisfy the requirement of assumption B2, namely $\ddot{\gamma}(t) \leqslant \dot{\gamma}(t)\gamma(t)$ whenever $\dot{\gamma}(t) \geqslant 0$. Let $H(t)$ be a positive and strictly increasing function that satisfies $\dot{H} = H(H - I_0)$ on some interval (t_0, t_1) and for some constant I_0. Then the difference $f := H - \gamma$ satisfies the following property: if $f(t_0) = f(t_1) = 0$ then $f(t) < 0$ for $t \in (t_0, t_1)$.

Proof. Note first that

$$\ddot{H} = \dot{H}(H - I_0) + H\dot{H} > H\dot{H}. \tag{6.1}$$

Strict inequality follows since for H to be strictly increasing on (t_0, t_1) we must have $\dot{H} = H(H - I_0) > 0$ there, while $H > 0$ by assumption.

Suppose $f(t) > 0$ for some $t \in (t_0, t_1)$. Then $f(t)$ has a maximum point t^* in (t_0, t_1), and at that point we have:

$$f(t^*) \geqslant 0, \qquad f'(t^*) = 0, \qquad f''(t^*) \leqslant 0.$$

Since $f = H - \gamma$, this gives

$$H(t^*) \geqslant \gamma(t^*), \qquad \dot{H}(t^*) = \dot{\gamma}(t^*), \qquad \ddot{H}(t^*) \leqslant \ddot{\gamma}(t^*).$$

Using these relations together with (6.1), we obtain at $t = t^*$ that

$$\ddot{\gamma} \geqslant \ddot{H} > H\dot{H} \geqslant \gamma\dot{\gamma}$$

while $\dot{\gamma}(t^*) = \dot{H}(t^*) > 0$, which contradicts the assumed property of γ. □

Lemma 6.4. Let $\{T_z\}$ be an equilibrium profile corresponding to a global equilibrium. Let $I(t)$ be given by (3.4) and (2.2), and let $\dot{I}$ denote its time derivative (whenever it exists). Then, for any $t > 0$, one of the following holds:

(a) t is interior to an interval over which $\dot{I} = 0$.

(b) There exists a sequence (T_{z_i}) which converges to t. Consequently, $H(t) - \gamma_y(t) < 0$ for any $y \in Z$ such that $t > T_y$, and $H(t) - \gamma_y(t) > 0$ for any y such that $t < T_y$.

Proof. It is evident from (2.2) that I is constant over each open interval which does not contain points from $\{T_z\}$. Thus, if t is interior to such an interval then (a) follows. Otherwise, there must exist a sequence (T_{z_i}) which converges to t. Consider first the special case where $t = T_z$, for some z. Take $y \in Z$ for which $T_y < t = T_z$. From lemma 6.2 it follows (using the contrapositive implication) that $y < z$, and assumption B1 now yields $\gamma_z < \gamma_y$. Therefore,

$$H(t) - \gamma_y(t) = H(T_z) - \gamma_y(T_z) < H(T_z) - \gamma_z(T_z) = 0,$$

where the equality follows since $T_z > 0$ maximises U_z, cf. (3.1). The proof for $T_y > t = T_z$ is identical. Finally, if $T_z = t$ does not exist but a sequence $T_{z_i} \to t$ is available, a simple limit argument based on the continuity of H and γ_z establishes the same relations. □

Lemma 6.5. Assume that the system is in global equilibrium. For any $z \in Z$, suppose T_z is a maximum point of the utility function U_z. Then

(i) $U_z' > 0$ for $t < T_z$.

(ii) $U_z' < 0$ for $t > T_z$.

Consequently, U_z is strictly unimodal, in the sense that it has a single local maximum.

Proof. The outline of the proof is as follows. From (3.1) we know that U_z' is sign equivalent to $H - \gamma_z$, so that the two assertions of the lemma can be equivalently stated in terms of the latter. Next, we consider separately the two cases in lemma 6.4, which essentially correspond to time instances where $\dot{I} = 0$ and $\dot{I} \neq 0$, respectively. In case (b), the required claim regarding the sign of $H - \gamma_z$ follows directly from that lemma. In case (a), where t is internal to an interval where $\dot{I}(t) = 0$, we will show that the convex-like properties of $H - \gamma_z$ on an interval where $\dot{I} = 0$ (lemma 6.3) prevent $H - \gamma_z$ from changing sign, hence it retains the same sign as in case (b). We now proceed with the detailed argument.

Let $\{T_z, z \in Z\}$ be an equilibrium profile corresponding to a global equilibrium. Recall from (3.1) that $H(T_z) = \gamma_z(T_z)$ if $T_z > 0$, while $H(T_z) \leqslant \gamma_z(T_z)$ if $T_z = 0$. Assumption A1 implies that $T_z < \infty$ for every z. We will often make use of the fact that $U_z'(t)$ is sign equivalent to $H(t) - \gamma_z(t)$ for any t and z.

Fix $z \in Z$. We first establish part (ii) of the lemma, namely that $U_z' < 0$ (equivalently, $H - \gamma_z < 0$) for $t > T_z$. Fixing t and referring to lemma 6.4, either case (a) or case (b) of that lemma must hold. If (b) holds, then indeed $H - \gamma_z < 0$ is implied by that lemma. Consider next the case that (a) holds, namely t is interior to an interval over which $\dot{I} = 0$. Extend that interval (t_0, t_1) as much as possible, while keeping $t_0 \geqslant T_z$. Let $f(t) := H(t) - \gamma_z(t)$, as in lemma 6.3. Note that $f(t)$ is continuously differentiable on that interval, as $H(t)$ is given by (3.5) with I constant. In order to apply lemma 6.3 we again consider two possibilities:

(i) $t_0 = T_z$. First we have $f(t_0) \equiv H(T_z) - \gamma_z(T_z) = 0$. Turning to the end point t_1, if $t_1 < \infty$ then case (b) of lemma 6.4 holds (by definition of t_1), and as $t_1 > T_z$ we have $f(t_1) < 0$. If $t_1 = \infty$, then since $H(t) \to m\mu$ while assumption A1 holds, then $f(t_1') < 0$ for t_1' large enough. In either case it follows from lemma 6.3 that $f \equiv H - \gamma_z < 0$ on (t_0, t_1), and in particular at t as required.

(ii) $t_0 > T_z$. Here again case (b) of lemma 6.4 holds (by definition of t_0), so that $f(t_0) < 0$. As before we have that $f(t_1) < 0$. Then again from lemma 6.3 it follows that $f = H - \gamma_z < 0$ at $t \in (t_0, t_1)$.

We have thus shown in either case that $U_z' < 0$ for $t > T_z$.

We next establish part (i), namely that $U_z' > 0$ (equivalently, $H - \gamma_z > 0$) for $t < T_z$. Some extra care is required here to prepare the conditions for application of lemma 6.3. As T_z is a maximum of U_z, then $U_z'(t)$ to the left of T_z should be initially negative. More precisely, there exists an interval (t_2, T_z) such that either (i) $U_z'(t) = 0$ on that interval, or (ii) $U_z'(t) < 0$ on that interval. We first show that (i) is impossible. We claim in that case that there are no points from $\{T_y\}_{y \in Z}$ in (t_2, T_z). Indeed, if $y > z$ then $T_y > T_z$ (by lemma 6.2); if $y < z$ then $\gamma_y < \gamma_z$ implies that $H(t) - \gamma_y(t) < H(t) - \gamma_z(t) = 0$, hence $U_y'(t) < 0$ on (t_2, T_z). It follows that U_y cannot have a maximum point T_y on that interval. As there are no points from $\{T_y\}_{y \in Z}$ in (t_2, T_z), it follows by lemma 6.4 that $\dot{I}(t) = 0$ there. Recalling that $U_z' = 0$ implies that $H_z - \gamma_z = 0$ on that interval, this contradicts lemma 6.3.

We are thus left with option (ii): $U_z'(t) < 0$ on the non-empty interval (t_2, T_z). Extend t_2 as much as possible to the left. If $t_2 = 0$ and $U_z(0) < 0$ then $U_z(t) < 0$ on $[0, T_z)$ and we are done; otherwise, $U_z(t_2) = 0$. Assuming that the latter holds, we will show that it leads to a contradiction. (Unfortunately, a direct application of lemma 6.3 to obtain a contradiction is impossible since $\dot{I} = 0$ need not hold on that interval. However, we will be able to apply the lemma to another type $x \leqslant z$ on a sub-interval where $\dot{I} = 0$ does hold.)

Let $t_3 = \inf\{t \in (t_2, T_z]\colon t \in \{T_y\}_{y \in Z}\}$. Note that $\dot{I} = 0$ on (t_2, t_3). Assume first that $t_3 \in \{T_y\}$, namely $t_3 = T_x$ for some $x \in Z$. We will next show the existence of a sub-interval $(t_0, t_1) \subset (t_2, t_3)$ on which $H - \gamma_x > 0$, while $H - \gamma_x = 0$ at t_0 and t_1. But this contradicts lemma 6.3.

We thus proceed to establish the existence of such (t_0, t_1). Note that $T_x \leqslant T_z$, so that $x \leqslant z$. We first claim that $T_x \neq t_2$, so that the interval (t_2, t_3) is nonempty. Indeed, if $x = z$ then $T_x = T_z > t_2$, while if $x < z$ then $H(t_2) - \gamma_x(t_2) < H(t_2) - \gamma_z(t_2) = 0$, while $H - \gamma_x = 0$ must hold at T_x which maximises U_x. Next, as shown above with respect to T_z, there exists an interval (t_0, T_x) to the left of T_x on which $H - \gamma_x > 0$, while $H - \gamma_x = 0$ at $t = T_x$ and at $t = t_0$ (note that $t_0 \geqslant t_2$: if $x = z$ then $t_0 = t_2$, while if $x < z$ then $t_0 > t_2$ since $H(t_2) - \gamma_x(t_2) < 0$ as just shown). Recalling that $\dot{I} = 0$ on (t_2, t_3), hence on its sub-interval (t_0, T_x), we can apply lemma 6.3 with $t_1 = T_x$ to obtain that $H - \gamma_x < 0$ on (t_0, T_x), which yields the required contradiction. We have thus completed the argument under the assumption that $t_3 \in \{T_y\}_{y \in Z}$.

Assume next that $t_3 \notin \{T_y\}_{y \in Z}$. By definition of t_3, there exists a decreasing sequence (T_{x_i}) that converges to t_3. We may now apply a continuity argument to obtain the same result as before. Define a new customer type x with cost-ratio function $\gamma_x(t) := \sup_{x_i} \gamma_{x_i}(t)$. Note that the functions γ_{x_i} are increasing in i, so that the last supremum can be replaced by a limit. If we assign a P_Z-measure zero to the new type then the system equilibrium is not modified. It may be verified by continuity that $T_x := t_3$ is a global maximum of the corresponding utility function $U_x(t)$. It may be further shown that the conclusion of lemma 6.3 continues to hold for the new type (even though the derivatives of γ_x may not be everywhere well defined). The argument above can therefore be repeated to obtain a contradiction with the assumption that $U_z(t_2) = 0$. This establishes part (i), and completed the proof of lemma 6.5. □

Proof of theorem 6.1. Lemma 6.5 immediately implies that any global equilibrium is also a myopic (and farsighted) one. But uniqueness of the latter was established is theorem 4.1, so that the global equilibrium is unique. Existence of the global equilibrium will also be inferred from that of the myopic one. First, repeating the argument of lemma 6.5 for the *myopic* equilibrium, it follows similarly that the utility functions of all customers are unimodal at the myopic equilibrium as well. Hence the myopic equilibrium is, in fact, a global one, and existence of the former has been established is theorem 4.1 under assumption A2. □

Theorem 6.1 thus established the uniqueness of the global equilibrium under the stated assumptions, and enables its computation as either the myopic or the farsighted equilibrium. We mention again that even if the conditions of this theorem are not satisfied then either one of these "local" equilibria may still present a a global equilibrium, as may be verified after the explicit computation of the local equilibria; see also the discussion at the end of the previous section and proposition 5.1 there. However, no guarantee of uniqueness is available in that case.

7. Illustrative examples

In this section we present several examples that illustrate the computational process and the modelling flexibility offered by our proposed model. We should point out, however, that these examples are strictly meant to illustrate the mathematical framework, and do not attempt to capture parameters of a realistic scenario.

In all the examples below, the cost-ratio γ_z is taken for convenience to be linear. We shall further assume that the service utilities $R_z(t)$ are all constant, so that $\gamma_z \equiv C_z'/R_z$ completely describes the cost structure. The queue parameters are fixed at $\lambda = m\mu = 2$.

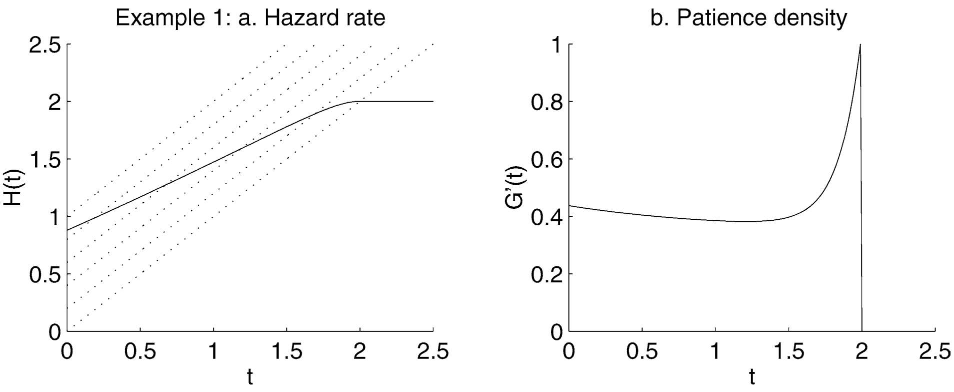

Figure 2. The hazard rate and the patience probability density function for example 1, computed with the farsighted procedure. The cost-ratio functions $\gamma_z(t) = z + t$ are also illustrated for $z \in [0, 1]$, with their density corresponding to that of the type z.

The first two examples illustrate the computational procedures and results for two different types of linear costs. The third example considers a mixture of two customer classes that violates the assumptions of theorem 6.1. The last example addresses the issue of inverse modelling, namely of determining a cost or population structure that gives rise to an a-priori given patience distribution.

7.1. Example 1: uniformly spaced costs with equal slopes

Let the cost-ratio function for type z customers be

$$\gamma_z(t) = z + t, \quad t \geqslant 0,$$

with z uniformly distributed in [0, 1]. This z represents the initial cost-ratio for a type z customer, and these costs increase at a uniform rate of 1 for all customers. The assumptions of theorem 6.1 are satisfied, hence the equilibrium point is unique and can be computed as either the farsighted or myopic equilibrium. We start by computing the hazard rate function $H(t)$ at equilibrium using the farsighted procedure. Here we backward integrate the differential equation (4.3) with the expression (5.2) for $I(t)$, starting with the boundary condition $H(\infty) = m\mu = 2$. In fact, it is easily verified that all abandonment times are bounded by $T_0 = 2$, since $\gamma_z(2) \equiv 2 + z \geqslant m\mu$ for all $z \in [0, 1]$. We can therefore take the terminal condition as $H(2) = 2$ (as per the remark in section 5). In the numeric computation, $H(2.5) = 2$ was used.

Numeric integration was carried out using simple Euler approximations, with a resolution of $N = 100$ points per unit time. The computed function $H(t)$ is shown in figure 2, alongside with an illustration of the functions $\gamma_z(t)$ that were used in this example. The patience distribution function $G(t)$, and the corresponding density $G'(t)$,

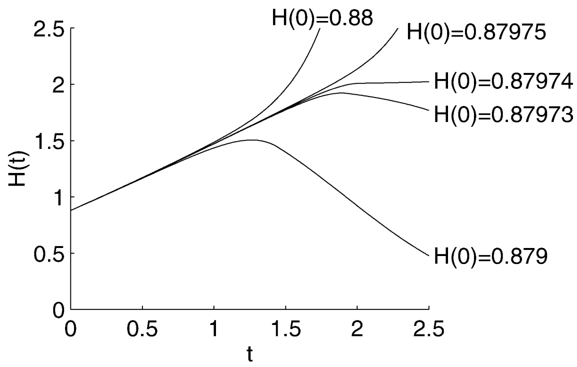

Figure 3. The hazard rate function $H(t)$ for example 1, computed with the myopic equilibrium procedure, for different initial conditions $H(0)$.

may be directly obtained using (5.1). More simply, it may be calculated from (3.4), namely,

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

where $I(t)$ is obtained during the computation of $H(t)$. $G'(t)$ is depicted in figure 2(b). As expected, T is bounded in $[0, 2]$. Its distribution is close to uniform, with some increase near the higher end.

For comparison, we compute $H(t)$ for this example also using the myopic equilibrium procedure. Recall that this is based on forward-integrating (4.3) with $I(t)$ from (4.2), and employing a scalar search procedure on $H(0)$ so that $H(\infty) = m\mu = 2$. The results for different initial conditions are shown in figure 3, where the integration was carried out up to $T = 2.5$ with the same resolution $N = 100$ as before. Note the high sensitivity of the final value to $H(0)$. The required terminal condition was obtained for $H(0) = 0.87974$, while the farsighted computation above yielded $H(0) = 0.87958$. Evidently, the results of these two computations coincide.

7.2. Example 2: uniformly distributed slopes

For the second example we take a family of cost-ratio functions with variable slope:

$$\gamma_z(t) = zt.$$

Obviously the type parameter z coincides with the slope. We choose its distribution so that the angular density is uniform in the first quadrant: $z = \tan(\theta)$, with θ uniformly distributed in $[\epsilon, \pi/2]$. The offset ϵ is introduced for numerical convenience, and taken here as $\epsilon = 0.01$. Again, the assumptions of theorem 6.1 are satisfied. The hazard rate function at equilibrium (computed with the farsighted procedure) is shown in figure 4, along with the patience density.

The abandonment profile in this example is of course quite different than in the previous one, and can be related to the density of the cost functions that intersect the

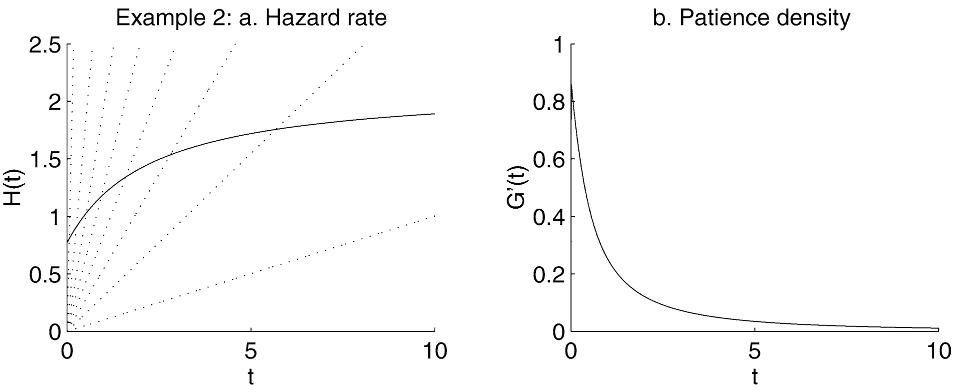

Figure 4. $H(t)$ and $G'(t)$ for example 2. The cost-ratio functions are $\gamma_z(t) = zt$, with uniform angular density (note the different axis scales).

hazard rate function at a given time. Similar changes may be induced by modifying the type density.

7.3. *Example 3: unordered cost ratios*

Assume next that the customer type is obtained as a mixture of two customer classes. The first class is the same as in example 1:

$$\gamma_{z_1}(t) = z_1 + t, \quad z_1 \sim U[0, 1].$$

The second class is given by

$$\gamma_{z_2}(t) = -2z_2 + 2t, \quad z_2 \sim U[0, 1].$$

Customers in this class are seen to have a lower initial cost, but it increases more rapidly in time. The customer type z is then the class designator (1 or 2, with equal probabilities), and the corresponding parameter z_1 or z_2. The situation is illustrated by the dotted lines in figure 5. Obviously, the cost-ratio functions are not ordered, as required in theorem 6.1 so that existence of a unique global equilibrium is not assured. Still, a unique *myopic* equilibrium does exist by theorem 4.1.

The myopic equilibrium was calculated and is depicted in figure 5. The jump in the patience density corresponds to the point where the second class start abandoning. It is easily seen that each cost-ratio function intersects the hazard rate at most once, so that this equilibrium is also a global one; this was indeed verified by computing the farsighted equilibrium that coincided with the above (proposition 5.1).

7.4. *Example 4: inverse modelling*

In our modelling framework, the problem of inverse modelling concerns the the construction of an appropriate customer cost structure to fit a given abandonment profile $G(t)$. Obviously, this is a basic step in fitting our model to empirical data. We shall

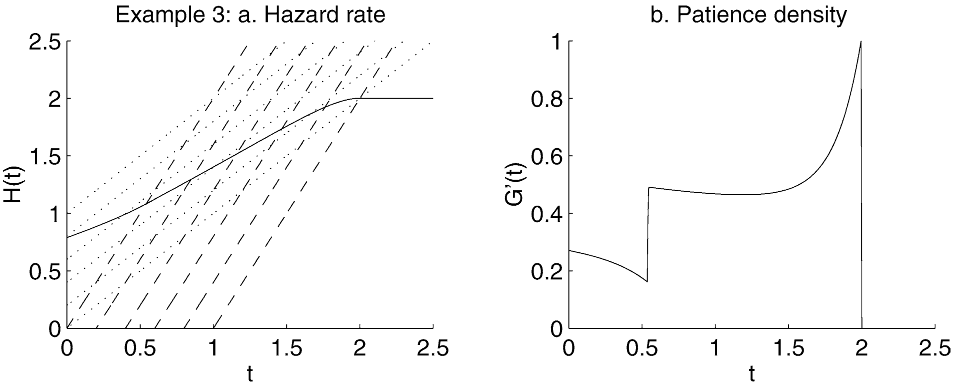

Figure 5. $H(t)$ and $G'(t)$ for example 3. The cost-ratio functions are a mixture of two classes with different slopes.

illustrate here the following variant of this problem. Suppose that the cost-ratio functions $\gamma_z(t)$ are given. Find a probability distribution P_Z on the type parameter z so that the induced equilibrium point gives rise to a-priori given abandonment profile (patience distribution).

The queueing system is the same as above, and the cost functions are as in example 1:

$$\gamma_z(t) = z + t.$$

Here the distribution of z (on the real line) is not given, and needs to be determined. Let the required patience distribution be exponential with unit parameter:

$$\overline{G}(t) = \mathrm{e}^{-t}, \quad t \geqslant 0.$$

To outline the computation procedure, we first note the $\overline{G}$ uniquely determines the (required) hazard rate function $H(t)$, through (3.3), (3.4) and $H = F'/\overline{F}$. Next, we compute the distribution P_Z on z that gives equality in (4.1). This density need not exist in general (see the comments below). However, if it does, it leads to a *myopic* equilibrium with the required patience profile. In general, it needs to be verified whether this equilibrium is a global one. This is assured, however, if the cost structure satisfies the ordering assumption of theorem 6.1.

Furthermore, if the cost-ratio functions are ordered, say increasing in z, then equation (4.1) reduces to

$$G(t) = P_Z\{z\colon z \geqslant z_0(t)\} \equiv 1 - F_Z\bigl(z_0(t)\bigr), \tag{7.1}$$

where $z_0(t)$ is the minimal z which satisfies the inequality in (4.1). The function $z_0(t)$ is increasing in t by its definition, and is easy to calculate once $H(t)$ is given. Thus, the distributions $G(t)$ and $F_Z(z)$ are related through the "scale change" $z_0(t)$.

The resulting hazard rate $H(t)$ and type probability density $f_Z(z)$ are shown in figure 6. The support of f_Z corresponds to the first γ_z that does not intersect H. At the

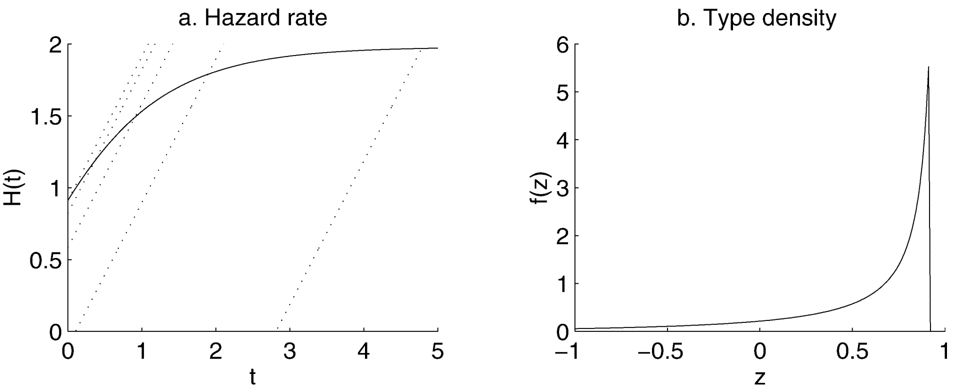

Figure 6. The probability density function of z for example 4 (“inverse modelling”).

other end, when H becomes constant the above-mentioned scale change $z_0(t)$ becomes linear, so that the lower end of f_z becomes exponential (as G').

A few additional comments are in order, in light of the last example. Recall that an abandonment (with $T > 0$) occurs at an intersection of the cost and the hazard rate functions. Thus, the required hazard rate function determines only the values of the cost function at these intersection points. The form of the cost functions away from these intersection points is therefore arbitrary. At the intersection point, however, the slope of the cost must be larger than that of $H(t)$ (for otherwise we would have a maximum rather than a minimum of the utility function). This sets a limitation on the hazard rate functions that are feasible with a given family of costs: for example, the above choice of unity-slope γ_z cannot give rise to $H(t)$ with slope larger than 1. In that case, cost functions with larger slopes must be introduced.

A more general modelling problem is that of inferring the *shape* of the cost functions in additions to their distribution. A detailed discussion of this problem is outside the scope of the present paper.

8. A finite number of customer types

Our analysis so far was mainly carried out under assumption A2, which requires the set of types to be non-discrete. From a descriptive viewpoint this seems reasonable, since it is hardly likely that different customers, or even the same customer on subsequent visits to a queue, will have exactly the same cost parameters. However, in modelling practice it may be convenient to divide the customer population into a *finite* number of types (in terms of their cost functions). We briefly address this model here.

An important characteristic of the discrete case is the essential role of randomized decisions. In equilibrium, customers of the same type will typically be required to choose different abandonment times according to some probability distribution (which is computed in [7] for the single-type model). Consequently, a direct analysis of the discrete case would require a somewhat different framework than the one used above. In place of repeating the lengthy analysis, we indicate below how the discrete-type model can be embedded within the continuous model, and demonstrate how this embedding allows computation of the equilibrium profile to required precision.

Consider a finite set $I = \{1, 2, \ldots, n\}$ of types, specified by the cost parameters $(\widetilde{C}_i(t), \widetilde{R}_i(t))$, with corresponding probabilities p_i, $i \in I$. Thus, an arriving customer will be of type i with probability p_i. As a first step, we embed this model within a continuous-parameter one. Let z (the type of an arriving customer) be a uniform random variable on $Z = [0, 1]$. Partition the set $[0, 1]$ into consecutive intervals of length p_i, and identify z with i if it falls in the corresponding interval; that is, define

$$(C_z, R_z) = \big(\widetilde{C}_i, \widetilde{R}_i\big) \quad \text{if } \rho_{i-1} \leqslant z < \rho,$$

where $\rho_i = \sum_{j=1}^{i} p_i$, and $\rho_0 = 0$. It is obvious that the two models are equivalent.

Still, the new model does not obey the continuity requirement in assumption A2. We therefore introduce a small perturbation in the cost parameters. For concreteness, define the following *positively perturbed model*:

$$C_z^{\epsilon} = \widetilde{C}_i + \epsilon(z - \rho_i) \quad \text{if } \rho_{i-1} \leqslant z < \rho,$$

where ϵ is a small positive number (and R_z as before). The corresponding cost-ratio is then $\gamma_z(t) = (C_z^{\epsilon})'(t)/R_z(t)$. We similarly define the *negatively* perturbed model, with ϵ replaced by $-\epsilon$.

The perturbed model "smears" the waiting cost function $\widetilde{C}_i$ over a small vertical band (of size ϵp_i). It may be easily verified that now this model does satisfy assumption A2. Moreover, if the discrete model satisfies the ordering and concavity properties from section 6, then so does the perturbed model (for ϵ small enough). We can apply our previous results to the perturbed model in order to compute the equilibrium point, which gives an approximation to the equilibrium of the discrete model. Moreover, by using both positive and negative perturbations we can contain the perturbation error.

A formal line of analysis would proceed by establishing the following properties:

(i) Monotonicity: The abandonment profile $\{T_z\}$ is monotone decreasing in the perturbation parameter ϵ. (Note that increasing ϵ increases the waiting cost of all customers.)

(ii) Continuity of the abandonment profile in ϵ.

Together these properties allow computation of the equilibrium to required precision, and also a proof of uniqueness under the assumptions of section 6. We do not pursue the analysis here, but merely provide an example to demonstrate the application of this framework.

8.1. Example 5: a two-type model

We consider the same queueing model as in the previous section, with cost-ratio functions

$$\gamma_1(t) = t, \qquad \gamma_2(t) = t + 0.5$$

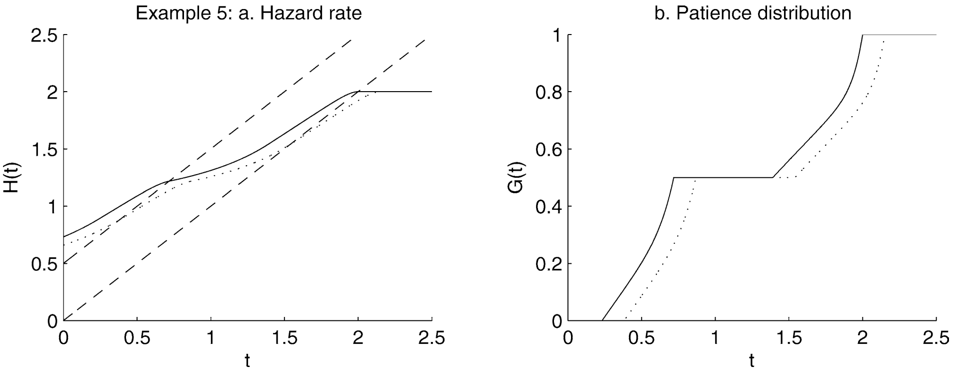

Figure 7. Equilibrium profiles for the perturbed models in example 5, with $\epsilon = 0.3$. The positively perturbed model is in solid line, the negatively perturbed one is dotted. The two straight lines are the cost-ratio functions of the two customer types.

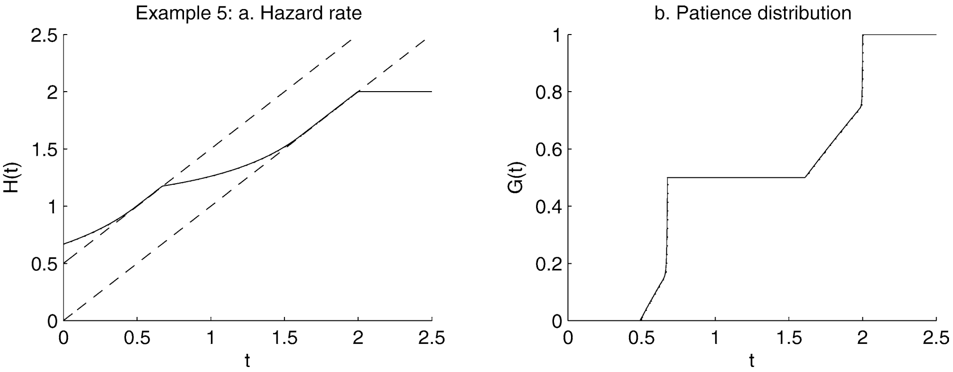

Figure 8. The same perturbed models with $\epsilon = 0.01$.

corresponding to customers of types 1 and 2. The two types are equally probable: $p_1 = p_2 = 0.5$. Figure 7 depicts the hazard rate and equilibrium profiles corresponding to both the positively and negatively perturbed models, with $\epsilon = 0.3$. A relatively large value of ϵ was chosen in order to demonstrate the monotonic relation between the two profiles (the cumulative distribution of G is depicted here rather than its density, to make this relation more apparent). Figure 8 shows the same quantities for $\epsilon = 0.01$. Here the results for the two perturbed models practically coincide, and give the equilibrium of the discrete model.

It is interesting to note in the last figure that the equilibrium hazard-rate function $H(t)$ follows along one of the cost-ratio functions in the two time intervals where abandonments occur (namely, when $G(t)$ is strictly increasing). This follows from the necessary condition (3.2), which implies that $H(t) = \gamma_i(t)$ whenever customers of type i abandon. In-between those intervals, the hazard rate function evolves according to equation (3.5), with $I(t)$ a constant.

9. Concluding remarks

We have considered in this paper a decision-theoretic model for customer abandonments in invisible queues. The model is inherently *adaptive*, in the sense that the optimal decisions of the customers depend on the waiting time distribution offered by the system. This decision model was incorporated and analysed within a basic $(\mathrm{M}/\mathrm{M}/m)$ queueing system. We have demonstrated the existence and uniqueness of the system equilibrium point under certain conditions on the waiting cost functions. Similar properties were established, under much broader conditions, for the modified concept of the myopic equilibrium. It was further demonstrated how these equilibria may be calculated and related to observed system characteristics. These results extend previous ones by incorporating nonlinear waiting costs and heterogeneous customer population, which may be essential for realistic modelling.

It should be emphasised that the present paper merely provides a mathematical framework for the proposed approach, while leaving a great deal of modelling flexibility in specifying the shape of the waiting cost functions and their distribution. The usefulness of such a model for queueing practice should be measured in its ability to provide reliable predictions for system characteristics under varying system conditions. To obtain that goal, the present work should be complemented with a methodology for determining the model parameters. This entails both empirical methods for estimating these parameters from measurements of a specific system, and the development of general guidelines regarding the shape of the waiting cost functions in typical waiting scenarios. To the best of our knowledge, these issues have not yet been addressed systematically. Some related quantitative work on the affective response to waiting may be found in [3,12,13]; see also [16] for further references and discussion.

Several key issues remain for further study. In modern call centers, selective information regarding the queue status is supplied to waiting customers. It should be of major interest to incorporate such information into our model, and investigate its consequences. As mentioned, the calibration and validation of the proposed model is essential for its application. More broadly, further study of the abandonment phenomena and its interaction with queueing performance is required, based on empirical data and human decision modelling.

Acknowledgements

The authors would like to thank an anonymous referee for constructive and useful comments on this paper. The reported research was partially supported by the Israeli Science Foundation, Grants No. 388/99-3 and 126/02-1, by the Niderzaksen Fund, and by the Technion V.P.R. fund for the promotion of sponsored research, and by the Fund for Promotion of Research at the Technion.

References

[1] M. Armony and C. Maglaras, On customer contact centers with a call-back option: Customers decisions, routing rules, and system design, Oper. Res. 52(2) (2004) 271–292.

[2] F. Baccelli and G. Hebuterne, On queues with impatient customers, in: *Proceedings of Performance'81*, ed. F.J. Kylstra (North-Holland, Amsterdam, 1981) pp. 159–179.

[3] Z. Carmon and D. Kahneman, The experienced utility of queuing: experience profiles and retrospective evaluations of simulated queues, Working Paper, Fuqua School of Business, Duke University (1998).

[4] S. Dewan and H. Mendelson, User delay costs and internal pricing for a service facility, Management Sci. 36(12) (1990) 1502–1517.

[5] R. Hassin and M. Haviv, Equilibrium strategies for queues with impatient customers, Oper. Res. Lett. 17 (1995) 41–45.

[6] R.B. Haugen and E. Skogan, Queueing systems with stochastic time out, IEEE Trans. Commun. 28 (1980) 1984–1989.

[7] M. Haviv and Y. Ritov, Homogeneous customers renege from invisible queues at random times under deteriorating waiting conditions, Queueing Systems 38 (2001) 495–508.

[8] E. Hille, *Lectures on Ordinary Differential Equations* (Addison-Wesley, Reading, MA, 1969).

[9] A. Mandelbaum, A. Sakov and S. Zeltyn, Empirical analysis of a call center, Technical Report, Technion, Faculty of Industrial Engineering (August 2000).

[10] A. Mandelbaum and N. Shimkin, A model for rational abandonment from invisible queues, Queueing Systems 36 (2000) 141–173.

[11] H. Mendelson and S. Whang, Optimal incentive-compatible priority pricing for the M/M/1 queue, Oper. Res. 38(5) (1990) 870–883.

[12] E.E. Osuna, The psychological cost of waiting, J. Math. Psychology 29 (1985) 82–105.

[13] C. Palm, Methods of judging the annoyance caused by congestion, Tele 2 (1953) 1–20.

[14] C.M. Rump and S. Stidham, Stability and chaos in input pricing for a service facility with adaptive customer response to congestion, Management Sci. 44(2) (1988) 246–261.

[15] W. Whitt, How multiserver queues scale with growing congestion-dependent demand, Oper. Res. 51(4) (2003) 531–542.

[16] E. Zohar, A. Mandelbaum and N. Shimkin, Adaptive behavior of impatient customers in tele-queues: theory and empirical support, Management Sci. 48(4) (2002) 566–583.