Non-Threshold Equilibrium for Customers Joining an M/G/1 Queue

Eitan Altman
INRIA, BP93
2004 Route des Lucioles
06902 Sophia-Antipolis Cedex
France

Refael Hassin
Department of Statistics
and Operations Research
Tel-Aviv University
Tel-Aviv 69978, Israel

November 13, 2002

Abstract

An important research direction in the control of queueing systems has been to establish structural properties. In particular, there has been an intensive effort to obtain conditions in which threshold policies are optimal for socially and individually optimality. The goal of our paper is to provide a counter-example for the optimality (or even ϵ-optimality) of any threshold type policy in an admission control problem to a queue for an individually optimal criterion. We consider a situation where customers that arrive according to a Poisson process have to decide individually whether or not to join a queue with general service time. The available information is whether or not a customer is served, as well as the number of customers at the queue. The problem is modeled as a non-cooperative game. We argue that the non-optimality of a threshold policy is due to the fact that the queue length provides a signal to the arriving customers on the unknown residual service time. Instead of a threshold policy, we obtain an alternative ϵ-optimal policy with a different structure, using some fluid limit model that approximates our system. A correlated equilibrium for the limiting fluid model is identified and used to establish the ϵ (non-correlated) equilibrium Nash equilibrium for our original model.

1 Introduction

An important research direction in the control of queueing systems has been to establish structural properties [1, 2, 6, 9, 10, 12, 14, 16]. In particular, there has been an intensive effort to obtain conditions in which threshold policies are optimal for social and individual optimality. Little is known on cases where optimal policies do not have such a structure. In the case of social optimality some results on optimality of non-threshold policies are available when extra constraints are imposed, see [8]. The goal of our paper is to provide a counter-example for the optimality (or even ϵ-optimality) of any threshold type policy in an admission control problem to a queue for an individually optimal criterion. Some other interesting counter examples (related to structures other than threshold policies) can be found in [7] and [15].

We consider a service facility in which an arriving customer can observe the length of the queue upon arrival. Customers are identical. The value of service is r and the cost of spending time in queue (or in service) is c per time unit. An arriving customer can either join the queue or balk (leave without being served). The decision is made upon arrival. It is well known that in a $G/M/s$ system, the optimal customers' strategy is to join the queue if and only if its length is at most some 'threshold value' (see Naor [12] for $M/M/1$, Yechiali [16] for $G/M/s$ and [2, 6, 9] for surveys).

Moreover, this type of behavior corresponds to a Nash equilibrium in the sense that it is optimal given that others follow it (for example, by never reneging after joining the queue). Below, we show that this property doesn't hold in general when the service distribution is not exponential. In particular we discuss the $M/G/1$ case. We show by an example that customer's affect each other not only by directly increasing the length of the queue but also by transmitting relevant information on the residual service time. This type of 'positive externalities' may lead to more complicated equilibrium behavior.

We assume throughout that an arriving customer does not have the possibility of reneging. For other models in queues that include reneging, we refer the reader to [6].

Consider an $M/G/1$ queue and denote by $E(R|j)$ the expected residual service time of the customer in service given that there are j customers in the system. Mandelbaum and Yechiali [11] and Fakinos [4] have shown that

$$E(R|j) = \frac{1-\rho}{\lambda \pi_j}\left(1 - \sum_{k=0}^{j} \pi_k\right), \tag{1}$$

where π_k is the probability that there are k customers in the system.

Suppose (as in Whitt's model [15]) that the service time is 0 with an extremely high probability and 1 with the complement extremely small probability. Under these assumptions π_0 is approximately 1 and almost all customers are served without delay. If an arrival sees any queue it means that a customer with unit time service is being served. Consider an interval of unit length in which a 1-customer is served. The expected number of arrivals during the interval is λ. Suppose that customers enter the queue if the number of customers in the queue (not including in service) is strictly less than a threshold L. We assume that λ is large enough so that in a randomly chosen instant in this unit interval the probability that we see i ($i < L$) customers in the queue is approximately $1/\lambda$, for all $i \ll \lambda$.

Substituting in (1) we get

$$\begin{aligned} E(R|i) - E(R|i+1) = \\ \left(-\frac{\sum_{j=1}^{i} \pi_j}{\pi_i} + \frac{\sum_{j=1}^{i+1} \pi_j}{\pi_{i+1}}\right) \frac{1-\rho}{\lambda} \approx \\ \left(-\frac{i/\lambda}{1/\lambda} + \frac{(i+1)/\lambda}{1/\lambda}\right) \frac{1-\rho}{\lambda} = \frac{1-\rho}{\lambda} > 0. \end{aligned}$$

We conclude that the expected residual service decreases as the queue size increases in the above range. This is so since any queue indicates that the rare event of an arrival of a customer with a unit time service has occurred. A longer queue indicates that more time elapsed since this service started. Therefore a longer queue makes it more attractive for a customer to join, as long as the probability that the queue contains a 1-customer is small.

Consider a fixed value of $0.5 < r < 1$, say $r = 0.75$, and let $c = 1$.

- A threshold $L = 0$ (which means that no one ever joins) does not define an equilibrium since if everybody adopts this strategy then an arrival who sees one customer in service and none waiting faces an expected residual service of 0.5 and will certainly join.
- Consider a pure threshold $L > 0$, or more generally, any strategy that prescribes joining with probability 1 when there is a customer in service and nobody in queue. Then this strategy

does not define an equilibrium since if everybody adopts this strategy then an arrival who sees one customer in service and none waiting sees an expected residual service of approximately 1 and will certainly not join.

- We thus propose the following candidate for an equilibrium: In the case of a customer in service and none waiting, the arrival enters with probability $p(r)$. Otherwise it enters with certainty as long as the queue is not greater than a (large) threshold L (so that the probability of a 1-customer in the queue (in addition to the one in service) becomes significant). We call such policies "delayed threshold policies". The value $p(r)$ is such that the residual service when an arrival sees a customer in service and none waiting is exactly r so that the arriving customer is indifferent between joining or balking. When it sees also a waiting customer the residual service is smaller (as we have seen) and in particular it is less than r so that it is worth joining. When the queue is very large, the probability that it contains a 1-customer becomes significant and at a certain point the sum of the (small) residual service time and (larger) expected wait for service of the customers in queue exceeds r and balking is preferred.

The above explanation for our candidate equilibrium is quite heuristic. The goal of the following sections of the paper is to study in a rigorous way whether there are indeed Nash equilibria among delayed threshold policies, and to further study the performance obtained with such policies. It will turn out that the analysis of the above model is quite complex. Although we are able to give some characterization of the performance of delayed threshold policies, it is not possible to obtain closed form formulae for the performance, nor to conclude that an equilibrium exists among these policies.

We therefore turn to analyze the limiting case where the arrival rate is very large, using a fluid approach. After formulating the problem for this limiting regime, we are able to obtain a correlated equilibrium with a public signal for that regime with a similar delayed threshold structure. This then allows us to obtain delayed threshold policies for our original problem that are ϵ-equilibria for all positive ϵ and for all arrival rates sufficiently large.

The structure of the paper is as follows. In the next section we introduce the original discrete model. The fluid approximation is introduced in Section 3. The relation between the original model and the fluid model is described in Section 4. Conditions for the existence of equilibria among delayed threshold policies for the fluid model are then given in Section 5, and numerical examples are given in Section 6. Then we show in Section 7 that ϵ-equilibria are obtained among delayed threshold policies for our initial discrete problem for large input rates.

2 The discrete model

Customers arrive according to a Poisson process with rate λ. An arrival is a massive customer that requires one unit of service time with probability q and with probability $1-q$ it requires zero service time; we then call it a "tiny" customer.

Each customer can decide to enter the queue and wait for service or to renege. Getting served has a value of r units per customer. On the other hand, we assume that waiting is costly so that there is a cost proportional to the expected waiting time till service of the customer begins, $E[W]$. The decision of joining or not is made individually by each customer. A customer would join the queue if $cE[W]-r<0$, will renege if $cE[W]-r>0$, and will be indifferent if $cE[W]-r=0$. Without loss of generality we can take $c=1$ (by a proper change of units of the constant r). Obviously $E[W]$ is nonzero only if upon arrival there is a massive customer in service. The available information to the customer is whether or not there is a customer in service, as well as the queue length.

RAFI: I Added The following The time at which the first customer *joins* the queue will denote in the sequal the first time that a customer arriving at the queue actually decided to join it (as opposed to the time at which the first customer *arrives* at the queue; that one need not actually join the queue).

Note that the decision of a customer n depends on the strategies of previous arrivals through $E[W_n]$.

To deal with our transient setting we need the following assumption which we shall often make.

Assumption 2.1 *The system is initially empty and at time zero a massive customer starts service. We restrict our study of equilibrium to the policy of customers that arrive during time* $[0,1]$. *A customer that arrives at time* t *during that period does not know* t *(i.e. how much service has already been given to the customer in service). We assume, however, that he knows that* $t \in [0,1]$. *This assumption is common knowledge.*

This assumption implies that customers during the transient period $[0,1]$ have some notion that we are in a particular transient state.

Let $x(t)$ be the queue size at time t (not including the customer in service).

We consider the following candidate for a Nash equilibrium. A customer joins if there is no customer in service. Moreover,

- As long as there is no queue and there is a massive customer in service, an arriving customer joins the queue with probability p.
- There is an integer L, such that a customer arriving at time t joins the queue if $0 < x(t) < L$ and balks if $x(t) \geq L$.

We call this type of policy a *delayed threshold policy* with parameters (p, L) and denote it by $N(p, L)$.

We have not been able to prove that there exists a Nash equilibrium of this form, but we shall present conditions in Section 7 under which for any positive ϵ there exists some $\lambda(\epsilon)$ such that for any $\lambda > \lambda(\epsilon)$, there exists an ϵ-Nash equilibrium[1] with this structure .

Remark 2.1 Assume now that a delayed threshold policy with parameters (p, L) is used by all customers. Consider the time τ from the instant that there is a massive customer in service (with one unit service required) till either the first customer joins the queue, or the service of the customer ends (whichever occurs first) and assume that the queue is initially empty. Then τ is distributed according to the exponential distribution with rate $\theta = p\lambda$, truncated at 1, since it is the minimum between 1 and between a sum of a geometrically distributed number of elements (with parameter p) where the elements are exponential i.i.d. RVs with parameter λ. Since the effect of the parameter p is to delay the instant at which customers start joining the queue, we shall use $N(\theta, L)$, with some abuse of notation, to denote the policy $N(p, L)$ for which $p = \theta/\lambda$.

To proceed, we first introduce the random variable $\tau[t]$ which denotes the first time that an arrival joined the system given that it occurred during the interval $[0, t]$, $t \leq 1$. Then its probability density is given by

$$f_{\tau[t]}(s) = \frac{1}{1 - \exp(-\theta s)} \theta \exp(-\theta t).$$

[1] A multi-strategy is called ϵ-Nash equilibrium if no player can benefit by more than an amount of ϵ by deviating from his strategy.

In particular,

$$E(\tau[t]) = \frac{-\exp(-\theta t)\left[\frac{1}{\theta}+t\right]+\frac{1}{\theta}}{1-\exp(-\theta t)} = \frac{1}{\theta}+t-\frac{t}{1-\exp(-\theta t)}. \tag{2}$$

$$E(\tau[t]^2) = \frac{1}{1-\exp(1-\theta)}\left(-t^2\exp(-\theta t)+\frac{2}{\theta}\left(-\exp(-\theta t)\left[\frac{1}{\theta}+t\right]+\frac{1}{\theta}\right)\right). \tag{3}$$

We now focus on some arbitrary arrival during the interval $[0,1]$. We make the following observations:

- The probability that the arrival finds the system empty upon arrival (conditioned on $\tau[1]$ and on the fact an arrival during the interval [0,1] occurred) is equal to $\tau[1]$;
- Let ζ be the arrival time of a customer that finds the system empty. Then $E(\zeta|\tau[1]) = \tau[1]/2$.

Combining the two observations we get that

$$E[\zeta] = \frac{E(\tau[1]^2)}{2E(\tau[1])} = \frac{-\exp(-\theta)}{2[-\exp(-\theta)(1/\theta+1)+1/\theta]}+\frac{1}{\theta}. \tag{4}$$

Note that the left hand equality in (4) is the standard expected recurrence time of the variable $\tau[1]$.

Some extra notation:

- $\overline{R}_i$:=the residual time till the end of the service of the served customer under the following assumption: there are i customers in the queue, and before the first one arrived the buffer was empty. This means in particular that when the current customer in service has begun its service, then the system was empty. It is thus defined under the transient assumption 2.1.
- $\overline{W}_i$:= the time till all the customers present at the queue are served, given that their number now is i under the assumption that before the first one arrived the buffer was empty. It is thus defined under the transient assumption 2.1. $\overline{W}_i$ is the waiting time of a customer that arrives during the time interval $[0,1]$ and that finds i others at the queue.
- $Z_i = \tau+Y_1+Y_2+\cdots+Y_i$ where τ is an exponentially distributed random variable with parameter θ and where Y_j, $j=1,\dots,i$ are exponentially distributed random variables with parameter λ, and where all random variables are independent. Note that the probability distribution of Z_i is the convolution of the distribution of an exponential R.V. with parameter θ and an Erlang-i R.V. with parameter λ. Z_i is distributed like the time from arrival of a massive customer to an empty system till the i-th arrival, assuming an $N(\theta,L)$ strategy with $L \geq i$.

For the delayed threshold policy (θ,L) to be a Nash equilibrium, we first need that p (or equivalently θ) be chosen such that $E[\overline{W}_0] = r$. In view of (4) this condition becomes

$$E[\overline{R}_0] \quad = \quad E[\overline{W}_0] = 1-E[\zeta] = 1+\frac{\exp(-\theta)}{2[-\exp(-\theta)(1/\theta+1)+1/\theta]}-\frac{1}{\theta}. \tag{5}$$

Central equations for our analysis are

$$E(\overline{R}_i) = 1-E[Z_i|Z_i\leq 1], \quad E(\overline{W}_i) = E(\overline{R}_i)+iq, \quad t\in(0,1),\ i>0. \tag{6}$$

To understand the above equations, we first note that all customers found at the queue by the customer whose expected residual time we compute (and who arrived, say, at time t), have arrived during the service time of the currently served massive customer. This is due to the definition of $\overline{R}$. Thus t has to be smaller than one. On the other hand, it is easily seen that t is distributed like Z_i. This explains the first relation. To get the second relation, we note that in addition to the residual time, our customer has to wait also for the service time of all massive customers present at the queue. The expectation of this time is clearly iq.

A sufficient condition for (p, L) to be a Nash equilibrium is that for all i greater than L, $E[\overline{R}_i] > r$ and for all $i \leq L$, $E[\overline{R}_i] < r$. A sufficient condition for (p, L) to be an ϵ-Nash equilibrium is that for all i greater than L, $E[\overline{R}_i] > r - \epsilon$ and for all $i \leq L$, $E[\overline{R}_i] < r - \epsilon$. This could be obtained if $E[\overline{R}_i] < r$ for all i smaller than some i_0, and then it increases to a value greater than r for $i \geq i_0$.

These conditions could be checked numerically, since we know the distribution of Z_i in (6). Instead, we shall show that the condition holds in an approximating fluid model obtained as a weak limit when the arrival rate becomes large. We shall investigate another equilibrium concept that applies to the limit problem and that is much easier to compute. We will then be able to construct an ϵ-equilibrium policy for our original discrete model for any positive ϵ and for any sufficiently large arrival rates.

3 The fluid model

We consider again the transient setting. We thus assume that at time zero a massive customer starts to be served and that the queue size just prior to the beginning of the service is zero. We note that the larger the arrival rate is, the more precision we shall have on the time elapsed since the beginning of the service time. We therefore analyze a fluid model, which can be viewed (as will be explained below) as the weak limit of the system (scaled in a proper way) as the arrival rate goes to infinity. We shall begin by describing the fluid model and then show that it is the limit of the original one.

3.1 The arrival model and service process

Customers arrive at a queue according to the superposition of two processes.

- The first is a fluid with rate λ. The fluid can be viewed as water that arrives to a dam. Thus during a time interval $[s, t]$, the amount of customers that arrive is $\lambda(t - s)$ (for $t > s$). In this process, one unit of customer brings η unit of workloads. We shall take as in the initial discrete model $\eta = 0$.

- The second process is Poisson with rate λ'. At each arrival of that process a so called "massive" customer arrives with one unit of workload. Thus during a time interval $[s, t]$, the expected amount of workload brought by the second process is $\lambda'(t - s)$ (for $t > s$), which is also the expected number of customers that arrive during that interval.

For the service process, we assume that there is a server that serves workload at a rate of one workload unit per time unit according to the FIFO discipline. An arriving customer observes the amount of fluid in the system and then decides whether or not to join.

Equivalently this state may be understood a measure of the time since the first customer joined the system (after the massive customer started being served); having λt customers in queue is thus equivalent to saying that this time equals t. Then each arriving customer has the information on the value of that time. We shall later allow the joining decision to depend yet on another external signal.

3.2 A correlated equilibrium with public signals

We consider the game where the set of players is the (continuum) set of arriving customers. The set of players is a combination of atomless players (those corresponding to the fluid arrival, in which a single individual has a mass of zero in terms of its impact on the performance of other future individuals) together with other atomic players (as an example for such games, in which atomless and atomic players are combined, see [5]). The decision of each player is to either join the queue and wait for his turn to be served, or not to join it. The available information is the total *amount of customers in the queue*. The actual workload in the queue, and in particular, the number of massive customers in the queue, is not known to an arrival; it cannot distinguish between the two types of customers in the queue.

As in the original model, in this fluid model we do not have equilibrium threshold policies neither. The precise example with $r = 0.75$ (and $\lambda = 1$) can be used again to convince ourselves.

The delayed threshold Nash equilibrium proposed for the original model cannot be implemented directly for the fluid model. Indeed, a process in which a continuum of players randomize independently cannot be defined mathematically. Now even if we could: if arrivals randomize with probability $p > 0$ when finding the queue empty with a customer in service, then after a time whose expectation is zero, the first customer will eventually join the queue. So such a policy would lead to the same type of evolution of the queue as a threshold policy with threshold zero, which we know is not an equilibrium.

We thus propose an alternative definition for an equilibrium policy which would turn out to be very similar to the original proposed Nash equilibrium. We shall show how we can translate it back to the original discrete model.

- Whenever the service of a massive customer begins when the queue is empty, then a timer is set with an exponential duration with parameter θ. As long as this timer counts, a red signal is displayed to advise arrivals not to join the queue. Once the timer expires, a green light is displayed to advise arrivals to join the queue. If the service of the massive customer ends before the timer expired then the light is not needed anymore, since there are no waiting customers at the queue and no customer in service (light is only needed when the queue is empty but there is a customer in service). So we may assume that it is simply turned off. Thus, if a customer arrives and the light is off, he will join the system and receive service immediately with no waiting.

- Once customers start joining the queue, they will continue to join it until it reaches a threshold λL.

We call this policy a (correlated) delayed threshold policy and use the notation $T(\theta, L)$ to characterize it. (Note that the candidate Nash equilibrium in the discrete setting used also a threshold L, but the first argument was a probability.) The above policy is known to be a *correlated policy with public signal* see e.g. [13] and references therein. We shall show that it constitutes a correlated equilibrium, i.e. that no customer has an incentive to deviate unilaterally from it (nor to ignore the advice given by the signal). (A correlated equilibrium is a Nash equilibrium where the strategies of users are allowed to depend on a correlating external signal. Thus a multistrategy for all users is a correlated equilibrium if no user has an incentive to deviate from its strategy, where a strategy of each user may depend both on the state of the queue as well as on the signal).

Below we do not consider the steady state operation of the system, we restrict to t that is smaller than the first time a massive customer ends its service (and we assume that the queue is initially empty.)

4 Relation between the fluid and the original model

We show here how a sequence of discrete models M_n, $n = 1, 2, 3, \ldots$, can be constructed that converge to the fluid model. More precisely, we show how to scale the parameters in the discrete model so that it defines the fluid limiting model.

4.1 The arrivals

Customers arrive at the model M_n according to a Poisson process with rate $n\lambda$. An arrival is a massive customer that requires one unit of service time with probability

$$q_n = \frac{\lambda'}{n\lambda},$$

and with probability $1 - q_n$ it requires zero service time; and is a tiny customer. We conclude that the arrival of massive customers at the nth model is according to a Poisson process with rate λ'.

We have the following relation between the fluid limit and the original one. Let $A^M(s,t)$ and $A^T(s,t)$ be the number of massive arrivals and the amount of tiny arrivals in the fluid model during time interval (s,t). With an extra subscript n, these quantities will correspond to the model M_n.

Then the arrival process of massive customers $A^M(s,t)$ has the same distribution as the arrival process of massive customers in each one of the models M_n: $A^M(s,t) =_d A_n^M(s,t)$. The arrival of tiny customers is given as the limit

$$A^T(s,t) = \lim_{n\to\infty} \frac{A_n^T(s,t)}{n} = \lambda(t-s).$$

4.2 The timer

In view of Remark 2.1 we could use, with some abuse of notation, (θ_n, L_n) to denote a delayed threshold policy in model M_n, which then defines $p_n = \theta_n/\lambda_n$.

Note that if we take

$$p_n = \frac{\theta}{\lambda_n} = \frac{\theta}{n\lambda},$$

then $\theta_n = \theta$ for all n.

More generally, we may conclude that if θ_n converges to θ, then the delay period before customers start joining the queue in model M_n converges in distribution to the delay period in the fluid model.

Note that both in the M_n as well as in the fluid model, the delay period cannot exceed 1 (Assumption 2.1).

4.3 The queue length process

Assume now that the delayed policy $N(\theta_n, nL)$ is used in model M_n and the delayed policy $T(\theta, L)$ is used in the fluid model. Let $X_n(t)$ denote the number of queued customers in M_n at time t. Then one can show that if θ_n converges to θ, then the process $X_n(t)/n$ converges to the fluid queue length process $x(t)$ in some sense (we delay this to the section 7 and to the Appendix).

Recall that we restrict here to t that is smaller than the first time a massive customer ends its service and that we assume that in all models, the queue is initially empty.

5 Delayed threshold equilibrium for the fluid model

We show in this section how one can obtain θ and L such that the delayed threshold policy (θ, L) constitutes a correlated equilibrium in the fluid model. Similarly to the discrete model, we define here

- $\overline{R}_t$:=the residual time under the following assumption: t time units ago, the buffer was empty and since then arrivals joined it (at rate λ). This means in particular that when the current customer in service has begun its service, then the system was empty. It is thus defined under the transient assumption 2.1.
- $\overline{W}_t$:= the time till all the customers present at the queue are served, given that there are λt customers at the queue, under the assumption that before the first one arrived the buffer was empty. It is thus defined under the transient assumption 2.1. $\overline{W}_t$ equals the waiting time of the λtth customer that arrives during the interval $[0, 1]$.

As in the discrete model, for the delayed threshold policy $T(\theta, L)$ to be a correlated equilibrium in the fluid model we first need that θ be chosen such that $E[\overline{R}_0] = r$. In view of (4) this condition becomes

$$E[\overline{R}_0] = E[\overline{W}_0] = 1 + \frac{\exp(-\theta)}{2[-\exp(-\theta)(1/\theta + 1) + 1/\theta]} - \frac{1}{\theta} = r. \tag{7}$$

Secondly, we should have $E[\overline{W}_t] < r$ for all t sufficiently small. We shall show through numerical examples that $E[\overline{W}_t]$ often has the following desirable behavior:

Property 5.1 *$E[\overline{W}_t]$ is convex, it increases "sufficiently" so that for all t greater than some $L \in [0, 1]$, $E[\overline{W}_t] > r$ and for all $0 < t \leq L$, $E[\overline{W}_t] \leq r$.*

Examples that satisfy this property are given in Figs. 1 and 2.

The key equations for our analysis are:

$$E(\overline{R}_t) = 1 - E[\tau|\tau \leq 1 - t] - t = 1 - E(\tau[1 - t]) - t, \quad E(\overline{W}_t) = E(\overline{R}_t) + \lambda' t, \quad t \in (0, 1). \tag{8}$$

We shall now explain the equations:

When the current customer started service, the system was empty. Then τ (exponentially distributed truncated at one) time passed till the queue started to build up. Thus when the queue starts building up then the residual service time is $1 - \tau$. Now t time units later, the residual service time has decreased by t. During that time, the expected number of massive customers that have arrived is $\lambda' t$ (the massive customers have Poisson distribution with rate λ' both in the discrete as well as in the fluid model). Note however, that we gained new information about τ. Since we consider the case in which after time t the service time has not expired, (assumption 2.1 which holds by definition for the transient variables $\overline{R}$ and $\overline{W}$) this excludes the possibility that τ were larger than $1-t$. Therefore we should take the expectation of τ with respect to the exponential distribution truncated at $1-t$.

Remark 5.1 *Note that by comparing (7) to (8) (with $t \to 0$) we see that indeed, $E[\overline{W}_t] < r$ for t sufficiently small (which should hold for Property 5.1 to be valid).*

Substituting (2) into (8) we obtain:

$$E[\overline{W}_t] = -\frac{1}{\theta} + \frac{1-t}{1-\exp(-\theta(1-t))} + \lambda' t. \tag{9}$$

Note that as $t \to 1$ we have, as expected, $E[\overline{W}_t] = \lambda'$: indeed, if t is close to 1 then it means that τ was close to zero, and the service time of the current customer is about to end. Therefore a customer that arrives at this time will have to wait 1 unit for each massive customer who arrived during 1 unit, which is in expectation λ'.

As $t \to 0$ we have for θ very large $E[\overline{W}_t] = 1 - 1/\theta = 1 - E[\tau]$. Indeed this is what we expect since for large θ, $E[\tau]$ equals approximately $E[\tau|\tau \leq 1]$.

We conclude the following:

Theorem 5.1 *Consider the transient framework of Assumption 2.1. Assume that Property 5.1 holds. Then a delayed threshold policy $T(\theta, L)$ is a correlated equilibrium with θ given by the solution of (7) and L is given by the solution of $E(\overline{W}_L) = r$ where $E(\overline{W}_t)$ is given by eq. (9) and $L \in [0,1]$.*

6 Numerical examples

Having shown that a simple threshold policy is in general not a Nash equilibrium, the goal of this section is to illustrate that there are cases in which a delayed threshold policy is a correlated equilibrium for the fluid model. For such cases, we show in the next section how to construct an ϵ-Nash equilibrium for the original discrete model provided that the arrival rate is sufficiently large.

Consider $r = 0.9377$ (and $c = 1$). Then eq. (7) yields $\theta = 8$. Now let $\lambda' = 0.95$. Using Eq. (9), we obtain in Figure 1 the expression for $E[\overline{W}_t]$. (Clearly, the value of $E[\overline{W}_t]$ is independent of λ). We see that it indeed satisfies Property 5.1, and that it first decreases and then increases in t within the interval $[0,1)$. The threshold value is $L = 0.9713$, which yields indeed $E[\overline{W}_t] = r$.

We now make only one change, λ' is chosen to be 1 instead of 0.95. Now $E[\overline{W}_t]$ is increasing all along the interval $[0,1)$, as can see in Fig. 2. The equilibrium threshold is $L = 0.947$ yielding indeed $E[\overline{W}_t] = r$.

On the other hand, taking $\lambda' = 0.8$, we still have a curve of $E[\overline{W}_t]$ similar to that in Fig. 1, but it does not intersect r in the interval $[0,1)$. Thus for the transience analysis that we did, we can only conclude that the equilibrium is such that after the random delay of θ, all customers should enter. Further decreasing λ' to e.g. 0.5, $E[\overline{W}_t]$ becomes monotone decreasing all over the interval $[0,1)$.

7 Back to the discrete model

We show in this section how the Nash equilibrium in the fluid model can be used to construct an (almost) equilibrium in the original model for all large enough arrival rates, i.e. for model M_n

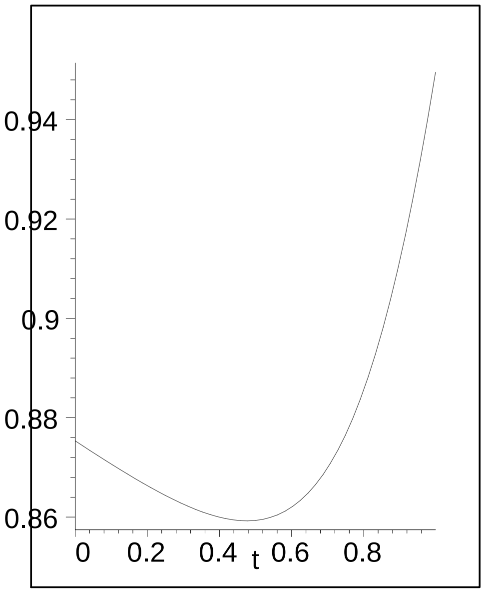

Figure 1: $E[\overline{W}_t]$ as a function of t, $\lambda' = 0.95$.

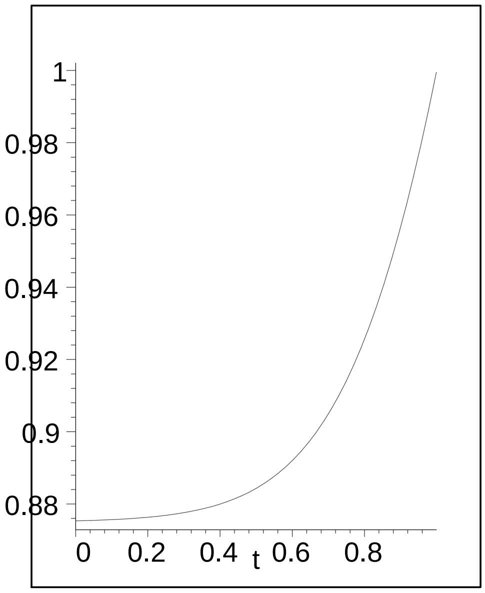

Figure 2: $E[\overline{W}_t]$ as a function of t, $\lambda' = 1$.

(defined in Section 4) for all n sufficiently large. We shall add a superscript n to the residual and waiting times corresponding to model M_n.

We need the following Key Lemma whose proof is given in the appendix.

Lemma 7.1

$$\lim_{n\to\infty} E[\overline{W}^n_{\lfloor \lambda n t\rfloor}] = E[\overline{W}_t]$$

uniformly over $t \in [0,1]$.

(In the above Lemma, $\lfloor x \rfloor$ means the largest integer smaller than x.)

We now present the following main result:

Proposition 7.1 *Assume that $\overline{R}$ satisfies property 5.1 for the limit fluid model, so that a policy of the type $T(\theta, L)$ is a correlated equilibrium in the fluid model for some θ and L. Then for any $\epsilon > 0$ there exists some $\mathcal{N}_\epsilon$ such that for all $n > \mathcal{N}_\epsilon$, the policy $N(\theta, n\lambda L)$ is an ϵ-equilibrium for the nth discrete model.*

Proof. Choose some $\epsilon > 0$. Due to Lemma 7.1, there exists some $\mathcal{N}_\epsilon$ such that

$$|E[\overline{W}^n_{\lfloor \lambda n t\rfloor}] - E[\overline{W}_t]| \leq \epsilon$$

for all $t \in [0,1]$ and $n > \mathcal{N}_\epsilon$.

Due to property 5.1, for any $s < L$, we have $E[\overline{W}_s] < r$. Combining this with the previous equation, we obtain

$$E[\overline{W}^n_{\lfloor \lambda n t\rfloor}] < nL\lambda + \epsilon, \text{ for all } n > \mathcal{N}_\epsilon.$$

Similarly, for all $s \geq L$ and $n > \mathcal{N}_\epsilon$, we have $E[\overline{W}^n_{\lfloor \lambda n t\rfloor}] \geq nL\lambda - \epsilon$. This establishes the proof. ∎

8 Appendix

Proof of Lemma 7.1. Note that $E(\overline{R}_t)$ in (8) can be rewritten as

$$E(\overline{R}_t) = 1 - E[\tau + t | \tau + t \leq 1].$$

We shall now construct a simple common probability space on which all models M_n are defined.

- First, let τ be an exponential random variable with parameter θ. This will be the random delay till the first customer enters the queue; it can be taken to be the same for all models M_n.
- Let Y_n be i.i.d. exponentially distributed random variables with parameter λ, and independent of τ. Denote
$$Y^i := Y_0 + Y_1 + ... + Y_{i-1},$$
where $Y_0 := 0$. The ith arrival $i \geq 1$ at model M_n is set to occur at time $Z_i^n := \theta + Y^i/n$.

It is easily checked that after the delay τ, the arrival process in model M_n is indeed Poisson with parameter $n\lambda$.

To obtain convergence of the discrete models to the fluid one, we extend the definition of the discrete models to functions of t as well; we define for $t \geq 0$:

$$y(t,n) := Y^{\lfloor \lambda n t \rfloor}, \qquad z(t,n) := \tau + y(t,n)/n.$$

Note that by the Strong Law of Large numbers, we have for each $t \geq 0$ P-a.s.:

$$\lim_{n\to\infty} \frac{y(t,n)}{n} = t.$$

Moreover, the convergence is *uniform over* $t \in [0,1]$. Indeed, for each n, $y(t,n)$ is increasing in t and the uniform convergence follows from e.g. [3, Lemma 4.1].

It then follows that P-a.s.,

$$\lim_{n\to\infty} z(t,n) = \tau + t \tag{10}$$

uniformly over $t \in [0,1]$.

This implies, in particular, that for any bounded function f on R,

$$\sup_{t\in[0,1]} \lim_{n\to\infty} |f(z(n,t)) - f(\tau + t)| = 0. \tag{11}$$

Hence, by the bounded convergence theorem, and by Jensen's inequality,

$$\lim_{n\to\infty} \sup_{t\in[0,1]} E[|f(z(n,t)) - f(\tau+t)|] \leq \lim_{n\to\infty} E[\sup_{t\in[0,1]} |f(z(n,t)) - f(\tau+t)|] = 0 \tag{12}$$

In particular, we get

$$\lim_{n\to\infty} E(Z^n_{\lfloor \lambda t n \rfloor} 1\{Z^n_{\lfloor \lambda t n \rfloor} \leq s\}) = E[(\tau+t)1\{\tau + t \leq s\}].$$

uniformly in $t \in [0,1]$.

We conclude from the above equations together with relations (6) and (8) that

$$\lim_{n\to\infty} E[\overline{R}^n_{\lfloor \lambda n t \rfloor}] = E[\overline{R}_t]$$

uniformly over $t \in [0,1]$. Again due to relations (6) and (8) used for the waiting times, one can then easily shows that this implies

$$\lim_{n\to\infty} E[\overline{W}^n_{\lfloor \lambda n t \rfloor}] = E[\overline{W}_t]$$

uniformly over $t \in [0,1]$ which establishes the Lemma. ∎

References

[1] E. Altman. Non zero-sum stochastic games in admission, service and routing control in queueing systems. *Queueing Systems*, 23:259–279, 1996.

[2] Altman, E. "Applications of dynamic games in queues", invited paper, to appear in *The Annals of Dynamic Games*, (2003).

[3] Dai, J.G. "On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models", *Annals of Applied Probability*, **5** (1995) 59-77.

[4] Fakinos, D. "The expected remaining service time in the single server queue," *Operations Research* **30** (1982) 1014-1018.

[5] Harker P. "Multiple equilibrium behaviors on networks", *Transportation Research B*, **22** (1988) 39-46.

[6] Hassin, R. and M. Haviv *To queue or not to queue: equilibrium behavior in queueing systems*, Kluwer (2002).

[7] A. Hordijk and G.M. Koole. The μc-rule is not optimal in the second node of the tandem queue: A counterexample. *Advances in Applied Probability*, 24:234–237, 1992.

[8] A. Hordijk and F. Spieksma. Constrained admission control to a queueing system. *Advances in Applied Probability*, 21:409–431, 1989.

[9] Johansen, S.G. and S. Stidham Jr. "Control of arrivals to a stochastic input-output system," Advances in Applied Probability **12** (1980) 972-999.

[10] G. M. Koole. Structural results for the control of queueing systems using event-based dynamic programming. *Queueing Systems*, 20:323–339, 1998.

[11] Mandelbaum, A. and U. Yechiali "The conditional residual service time in the $M/G/1$ queue", unpublished manuscript, (1979).

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

[13] Nowak A.S. and T.E.S. Raghavan "Existence of stationary correlated equilibria with symmetric information for discounted stochastic games", *Math. Oper. Res.*, **17** (1992) 519-526.

[14] S. Stidham. Socially and individually optimal control of arrivals to a $GI|M|1$ queue. *Management Science*, 24:1598–1610, 1970.

[15] Whitt, W. "Deciding which queue to join: some counterexamples," *Operations Research* **34** (1986) 55-62.

[16] Yechiali, U "Customers' optimal joining rules for the $GI/M/s$ queue," *Management Science* **18** (1972) 434-443.