The curse of the first-in-first-out queue discipline*

Trine Tornøe Platz, Lars Peter Østerdal†
Department of Business and Economics
University of Southern Denmark

September 13, 2012

Abstract

Every day people face waiting time in queues when going to the grocery store, checking in at the airport or leaving the stadium parking lot after a sports event. Often the facility where they are waiting for service will open at a specific time, and the agents will most likely be serviced in order of arrival using the first-in-first-out (FIFO) queue discipline. In this paper we consider a game in which agents choose their arrival time to a service facility where they can line up only after a given point in time. We show that in terms of equilibrium expected utility, the first-in-first-out queue discipline has inferior welfare properties in this model. In fact, in Nash equilibrium the first-in-first-out queue discipline is the *worst* in terms of utility and welfare among all work-conserving stochastic queue disciplines, while the last-in-first-out (LIFO) discipline comes off *best*.

JEL codes: C72, D62, R41
Keywords: queue discipline, Nash equilibrium, FIFO, LIFO, welfare, congestion.

1 Introduction

In many everyday situations, agents are serviced by a facility that opens at a specified time, for example when going to the bank or post-office, to a concert or sports event, or through check-in at the airport. In such situations, each agent must decide when to arrive at the server, taking into account his preferred service time and the waiting time he expects to face upon arrival. In some cases, the agent also faces the additional restriction that it is not possible to queue up at the server before it opens. This is for example true in the case of check-in at the airport when the counter number is not announced before the counter opens.

*We thank workshop participants at the Intelligent Road User Charging (IRUC) kick-off meeting at the Department of Economics, University of Copenhagen, November 21-22, 2011, and Refael Hassin for helpful comments. Jesper Breinbjerg has provided excellent research assistance. Financial support from the Danish Council for Strategic Research is gratefully acknowledged.

†Correspondence to: Lars Peter Østerdal, Department of Business and Economics, University of Southern Denmark, Campusvej 55, DK-5230 Odense M. E-mail: lpro@sam.sdu.dk

We consider the game of agents choosing arrival times in such situations, and we explore the welfare implications of employing specific queue disciplines.

The FIFO queue discipline (also known as first-come-first-served) is the most commonly analyzed queue discipline in the literature on queueing with endogenous arrival times. FIFO is generally considered as "fair" and is the focal discipline in many everyday situations, such as queueing at a grocery store or bank, as well as under more serious circumstances, such as in the allocation of donor organs to patients on the waiting list. However, while FIFO is intuitively fair and acceptable to most people, it may not be the best way of settling a queue. It is well known that settling a queue using FIFO will entail a cost compared to the socially optimal solution (with no queueing) that is generally not obtainable with endogenous arrivals. In this paper, however, we do not consider the loss incurred compared to the socially optimal solution but rather compare equilibrium welfare under different queue disciplines.

We model a facility that serves agents with a fixed capacity from a given point in time. Agents decide themselves when to line up for service, but they cannot line up before service begins. Besides the example already mentioned, situations that can be modeled in this way include passengers at a gate in the airport queueing up to board a flight, an audience exiting a venue after a concert or sporting event, and traffic where drivers are only allowed to line up for passing through a bottleneck from a given point in time (due to, say, environmental restrictions or bad weather).

We consider (pure strategy) Nash equilibria for a general family of stochastic queue disciplines with full capacity use, and we show that when agents are impatient, and the cost of queueing is linear in time, the FIFO discipline does the *worst* in terms of equilibrium utility and welfare while the Last-In-First-Out (LIFO) queue discipline does the *best*. Thus, these two queue disciplines provide an upper and lower bound for equilibrium utility and welfare under general stochastic queue disciplines.

Our setting is related to the classical bottleneck model of Vickrey (1969) that models congestion arising from the existence of a single bottleneck in the context of morning commute and trip timing. The original model was further analyzed and extended by Arnott et al. (1993), de Palma and Fosgerau (2009) and others, see de Palma and Fosgerau (2011) and references therein. Existing literature dealing with Vickrey's bottleneck model has largely assumed FIFO. An interesting exception is de Palma and Fosgerau (2009) who consider risk-averse agents and a family of stochastic queue disciplines that (to a vanishing degree) gives priority to early arrivals, ranging from FIFO to a completely random queue. They define a "no residual queue" property, which means that there is no queue at the time when the last user arrives at the queue, and they prove that this property holds in equilibrium under all queueing regimes that they consider. Remarkably, all queue disciplines within this family provide the same equilibrium utility and welfare. In this literature, it has commonly been assumed that the bottleneck facility is open at all times, while agents have a preferred time for passing the bottleneck and will incur a cost from being early or late. In contrast to these models, we consider, a setup in which the service facility opens at a specific point in time and where agent are not able to queue up before service opens.

Another strand of literature on queueing with strategic arrivals is based on the model from the seminal paper by Naor (1969). In this model, agents with exponentially distributed service times arrive at a service facility according to a Poisson distribution. At arrival they observe the length of the queue and decide whether to join the queue or balk. If they balk, they can not return. In this somewhat different setting, Hassin (1985) provides arguments to show that the LIFO queue discipline results in socially optimal behavior, see also Hassin and Haviv (2003, Ch. 2). Within this literature, Glazer and Hassin (1983) considered equilibrium arrival patterns to a server with opening (and closing) times. Subsequent extensions and variations include Jain et al. (2011) and Hassin and Kleiner (2011), and in line with Hassin and Kleiner, we consider a setup where early arrivals are not allowed.

The paper is organized as follows. In section 2, the model and key terms and assumptions are presented. Section 3 contains the results. Section 3.1 presents some preliminary results, in section 3.2, the FIFO discipline is considered, while 3.3 concerns the LIFO discipline. Section 4 contains concluding remarks.

2 Model

2.1 Basics

Suppose that at time 0 a facility opens that can service agents with a given fixed capacity. The capacity use of each agent is assumed to be negligible, and the set of agents is identified with $[0, 1]$, so that the total mass of agents is 1.

The bottleneck capacity at each unit of time is k. In what follows, we let T denote the earliest time before which it is possible to service all agents. We note that $T = 1/k$.

Let $R(t)$ be the *cumulative arrival distribution* (CAD) indicating for each t the share of agents that have arrived at the bottleneck up until time t. We assume that $R(t)$ is non-decreasing and right-differentiable everywhere. At any point in time where the CAD does not contain a jump we interpret the right-derivative $R'_+(t)$ as the rate ("speed") at which agents arrive. Figure 1 shows an example of a CAD reflecting the following arrival profile: One third of the agents arrive at time zero. From time zero agents arrive smoothly until time t_1 where another jump is seen in the figure. At t_1 a share of around one fifth of the agents arrive followed by another period of smooth arrivals. At t_2 the arrival rate increases, and by t_3 all agents have arrived.

2.2 Agent preferences

All agents have identical preferences. Each agent wants to be serviced as early as possible and spend a minimum of time in the queue. More precisely, we assume that there is a *waiting cost* c for each unit of time wasted in the queue. The *willingness-to-pay* for being served at time t is given by the bounded, continuous, and strictly decreasing function $w(t)$. The utility of an agent arriving at time s and being served at a time t after queuing

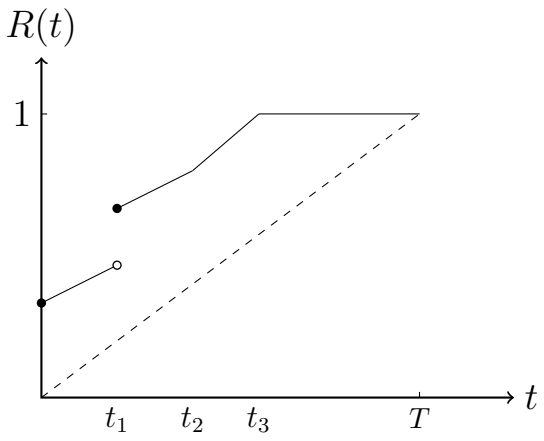

Figure 1: Example of a CAD

for $t - s$ units of time is then $w(t) - c(t - s)$. When time of service is stochastic, the agent maximizes expected utility.

2.3 Queue discipline

The cumulative serving time distribution for an agent joining the queue at time s is denoted $S^s(t)$. Thus, $S^s(t)$ is the cumulative probability that an agent arriving at time s has been serviced by time t. A profile of cumulative serving time distributions (one for each arrival time s) is *feasible* if the total "mass" of agents served in any interval of time does not exceed serving capacity.[1]

A *queue discipline* is a mapping that associates with a given cumulative arrival function $R(t)$ a feasible profile of cumulative serving time distributions. In other words, a queue discipline is a rule that describes when agents can expect to be served, given a specific arrival pattern.

A queue discipline is *work-conserving* if it does not waste capacity in the sense that if $R(t) \geq kt$ for all t, then all agents are served with certainty by time $T = 1/k$.

[1]Formally,

$$\sum_{\substack{t,\text{where } R(t) \text{ has jump } I_t, \\ \text{and } 0 \leq t \leq y}} I_t \cdot S^t(y) + \int_0^y R'_+(t) S^t(y) dt -$$

$$\sum_{\substack{t, \text{ where } R(t) \text{ has jump } I_t, \\ \text{and } 0 \leq t \leq x}} I_t \cdot S^t(x) - \int_0^x R'_+(t) S^t(x) dt \leq k(y - x),$$

for any $0 \leq x < y$. Note that the specification of the cumulative serving time distribution for agents who arrive at a point in time s where there is no jump and $R'_+(t) = 0$ is of no importance for feasibility.

2.4 Equilibrium and optimality

Given a work-conserving queue discipline, the expected utility of an agent arriving at time s is denoted by $EU[S^s(t)]$. Let t^* denote the point in time, where an agent is serviced. Then the expected utility of an agent arriving at time s will equal:

$$EU[S^s(t)] = E_s[w(t^*) - c(t^* - s)]. \tag{1}$$

Note that the expectation exists as long as the queue discipline is terminating.

Let $W^s = t^* - s$ denote the waiting time of an agent who joins the queue at time s and is serviced at time t^*, and let $E_s[W^s]$ denote the *expected waiting time* that can be induced from the cumulative serving time distribution $S^s(t)$. We may then write

$$EU[S^s(t)] = E_s[w(t^*)] + cE_s[W^s]. \tag{2}$$

Given a queue discipline, an arrival distribution, $R(t)$, is a *(Nash) equilibrium* if no agent can unilaterally improve his expected utility by choosing another arrival time. That is, for each s where $R(s)$ either jumps or $R'_+(s) > 0$ we have $EU[S^s(t)] \geq EU[S^{s'}(t)]$ for all s'. We then say that the arrival distribution is *supported* by the queue discipline.

A queue discipline is optimal (worst) if it supports an equilibrium arrival distribution that gives the highest (lowest) possible expected utility of among all equilibria supported by a work-conserving queue discipline.

3 Results

3.1 Preliminary results

First we state some general observations about queue disciplines and equilibrium cumulative arrival distributions. Note that throughout the paper, we limit ourselves to consider queue disciplines that are work-conserving.

We are interested in comparing the welfare that arises in equilibrium under different queue disciplines. Two observations regarding comparison of welfare between equilibria under different queue disciplines are stated in the lemma below. Figure 2 provides an illustration of the type of situation considered in the second part of the lemma.

Lemma 1. *Let $R(t) \geq kt$ be a Nash equilibrium under queue discipline 1, and let $Q(t) \geq kt$ be a Nash equilibrium under queue discipline 2. Then (a): If $R(t) = Q(t)$ for all t, equilibrium utility and hence welfare are the same under the two queue disciplines. (b) If $R(t) \geq Q(t)$ for all t, and the inequality is strict for some time interval, then equilibrium utility is higher under $Q(t)$.*

Proof. For the first part, let $R(t)$ be a given CAD where $R(t) \geq kt$ for all t.

Then the average expected utility (over all agents) is

$$\begin{aligned}
&\sum_{s, \text{ where } R(s) \text{ has jump } I_s} I_s EU[S^s(t)] + \int R'_+(s) EU[S^s(t)] ds \\
= &\sum_{s, \text{ where } R(s) \text{ has jump } I_s} I_s (E_s[w(t^*)] - cE_s[W^s]) + \int R'_+(s)(E_s[w(t^*)] - cE_s[W^s]) \\
= &\sum_{s, \text{ where } R(s) \text{ has jump } I_s} I_s E_s[w(t^*)] + \int R'_+(s) E_s[w(t^*)] ds \\
- &\sum_{s, \text{ where } R(s) \text{ has jump } I_s} I_s cE_s[W^s] - \int R'_+(s) cE_s[W^s] ds. \\
= &\int_0^T w(s) ds - c(\int_0^T R(s) ds - \frac{1}{2k}),
\end{aligned}$$

where the last line is due to the queue discipline being work-conserving whereby the cumulative amount of agents that have recieved service is the cumulative distribution function of the uniform distribution on $[0, T]$. Note that the content of the parethesis corresponds to the area between the $R(t)$ curve and the capacity line kt.

Since the average expected utility is thereby independent of the queue discipline, and the average expected utility in equilibrium is equal to the expected utility of any agent who arrives at time s where $R(s)$ either has a jump or $R'_+(s) > 0$, we conclude that if $Q(t) = R(t)$ for all t and $Q(t) \geq kt$ for all t, the equilibrium utility is the same for both CAD.

The second part follows immediately. □

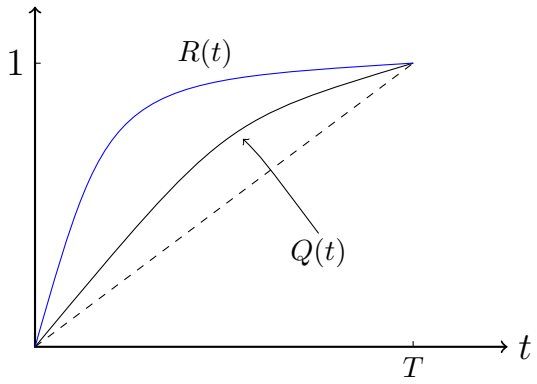

Figure 2: $R(t) \geq Q(t) \geq kt$

In the following subsections we consider more specifically the FIFO and LIFO disciplines.

3.2 First-in-first-out (FIFO) queue discipline

Consider the FIFO queue discipline under which agents are served in order of arrival and as soon as capacity becomes available.

For an agent arriving at a time t where $R(t)$ has no jump, the time of service is deterministic. He will be serviced as soon as every agent in the existing queue has been serviced. (Technically, the cumulative serving time distribution is a unit step function for each arrival time t). For agents arriving at a jump, time of service is uniformly distributed over an interval. Every agent already in the queue must be serviced before him, and some of the agents arriving at the jump will be serviced before him as well.

Next, we turn to the analysis of Nash equilibria in arrival patterns under FIFO.

Lemma 2. *Let $R(t)$ be a Nash equilibrium for the FIFO queue discipline. Then $R(t)$ is Lipschitz continuous.*

Proof. If $R(t)$ is supported by FIFO, the slope of $R(t)$ can not exceed k for any interval $[t_1, t_2]$, with $0 < t_1 < t_2 \leq T$. If this was the case, the waiting time would be increasing with t in this interval, implying that an agent with arrival time $t \in]t_1, t_2]$ could increase expected utility by arriving instead at t_1, since this would imply both a shorter waiting time and earlier service. Thus, the slope of $R(t)$ is less than k on the interval $]0, T]$, implying that $|R(t_2) - R(t_1)| \leq k|t_2 - t_1|$ for all t_1, t_2 with $0 < t_1 < t_2 \leq T$, and it follows that $R(t)$ is Lipschitz continuous. □

Considering the FIFO discipline, note that although Lemma 2 implies that $R(t)$ has no jumps for $t > 0$, we may have $R(0) > 0$, i.e. a non-zero fraction of the agents may arrive at time 0.

Let $R'_+(t)$ denote the right-derivative of R. Since $R(t)$ is Lipschitz continuous (and hence absolutely continuous) we have $R(t) = R(0) + \int_0^t R'_+(x)dx$.

In case not all agents arrive at time 0, the following result holds:

Lemma 3. *Suppose $R(t)$ is a Nash equilibrium for the FIFO queue discipline and $R(0) < 1$. Then the last agent who arrives is served immediately (i.e. the "no residual queue property" holds).*

Proof. Let $r = \min\{t | R(t) = 1\}$ and $R(0) < 1$, and assume that $r < T$. Then since $R(0) < 1$, and there are no jumps for $t > 0$, service time is deterministic and equal to T for a player arriving at r. This player could therefore increase expected utility by postponing arrival until $t = T$, in which case he would be serviced at time $t = T$ while avoiding waiting time entirely. □

Thus, due to $R(t)$ being work-conserving and Lemma 3, we have $R(T) = 1$, and if $R(0) < 1$ we have $R(t) < 1$ for all $t < T$.

Before we go any further, we would like to establish existence of an equilibrium arrival profile under FIFO.

Lemma 4. *There exists a cumulative arrival distribution supported by FIFO.*

Proof. Let $R(t)$ be the CAD for which $R(0) = 1$. Then, either the expected utility of arriving at time 0 is greater than the utility from arriving at $t = T$, in which case $R(t)$ is an equilibrium since no agent can profitably deviate, or the opposite is true, and $R(t)$ is not supported by FIFO.

Assume that all agents arriving at time 0 is not an equilibrium. In this case, we provide a constructive argument for the existence of an equilibrium.

Let I be the fraction of agents that would have to arrive at time 0 in order for the expected utility of these agents to equal $w(T)$. Note that I exists and is uniquely determined.

Note also that an agent arriving immediately after time 0 will at best be serviced at time $\frac{I}{k}$. For t sufficiently small, an agent arriving in the interval from 0 to t could therefore increase expected utility by arriving at 0 instead. This implies that the jump at 0 will be followed by a period of time where the density function vanishes, and no "mass" of agents arrives. Next, let t^* between 0 and I/k be the point where $w(I/k) - c(I/k - t^*) = w(T)$, i.e. we choose t^* such that an agent obtains the same expected utility from arriving at t^* and being served at time I/k as from arriving at time 0.

For any s with $t^* \leq t \leq T$, define $x(s)$ such that $s \leq s + x(s) \leq T$ and

$$w(s + x(s)) - cx(s) = w(T),$$

i.e. for an agent arriving at time s, $x(s)$ is the waiting time that gives the agent the same expected utility as the agents arriving at time 0. Note that $x(s)$ exists and is uniquely determined for each s. Moreover, $x(s)$ is strictly decreasing and continuous, $s + x(s)$ is strictly increasing, and $x(s) \to 0$ for $s \to T$.

Now, define $R(s)$ such that $R(s) = I$ for, $0 \leq s \leq t^*$, and $R(s) = k(s + x(s))$ for $t^* < s \leq T$. Then by construction $R(s)$ is an equilibrium arrival distribution function. □

In equilibrium under FIFO either every agent chooses to arrive at time 0, or some fraction of the agents arrive at time 0 followed by a period (from 0 to t^*) where the density function disappears (no arrivals) and finally a period where agents arrive smoothly until time T where $R(t) = 1$, see Figure 3.

Figure 3 shows an example of an equilibrium arrival distribution function under FIFO. Note that for an agent arriving at time $s > t^*$, the waiting time is given by the horizontal distance between the $R(t)$ curve and the kt line that shows the cumulative number of agents that has been serviced up until time t. The figure therefore also illustrates how the waiting time decreases with t (from t^*) until it reaches zero at T.

Next, we address the question of whether FIFO supports a unique CAD function.

Lemma 5. *Under the FIFO queue discipline there is at most one equilibrium.*

Proof. We prove this by way of contradiction. Let $R(t)$ and $Q(t)$ be two distinct cumulative arrival distributions supported by FIFO, and assume that $Q(0) < R(0) \leq 1$. From Lemma 3 and the queue disciplines being work-conserving, the equilibrium utility of an agent arriving at time 0 is at least $w(T)$ if $R(0) = 1$ and exactly $w(T)$ if $R(0) < 1$. However, since $Q(0) < R(0)$, the expected utility of an agent arriving at 0 is greater for

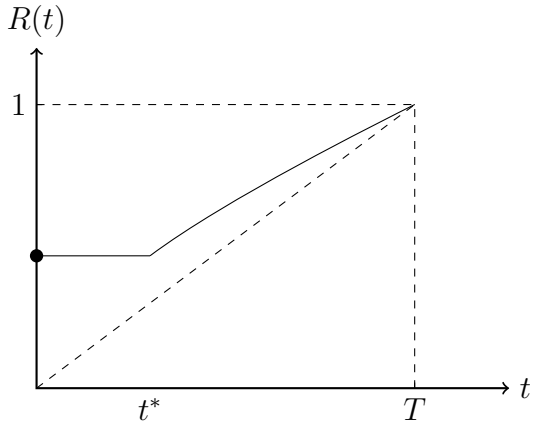

Figure 3: Equilibrium under FIFO

$Q(t)$ than for $R(t)$, i.e. a contradiction. Therefore, we must have $R(0) = Q(0)$. Then, given that $R(0) = Q(0)$, it readily follows that t^* is the same for the two arrival profiles. Further, we know that in both cases the expected utility of every player arriving from t^* and onwards equals $w(T)$, and therefore it must be that $R(t) = Q(t)$ for all t. Otherwise, two players arriving at time s between t^* and T would experience different waiting times, and hence, different expected utilities. □

Having established that FIFO always supports a unique equilibrium CAD, we move on to state the following negative result regarding the welfare properties of the FIFO queue discipline.

Proposition 1. *FIFO minimizes welfare.*

Proof. There are two cases: (a) all agents arrive at time 0, and (b) some fraction of the agents arrive at time 0 followed by a period where nobody arrives and then a period where agents arrive smoothly until time T when the last agent arrives.

In case (a) total waiting time is the highest possible among work-conserving queue disciplines. Thus, no other queue discipline can do worse, and FIFO therefore minimizes welfare among work-conserving queue disciplines.

In case (b), since $r = T$, equilibrium utility is equal to $w(T)$. Under every work-conserving queue discipline, all players are serviced by T. Thus, a player can always choose to arrive at $t = T$ and obtain utility $w(T)$. If, for some queue discipline $r < T$, then this implies that equilibrium utility is at least $w(T)$, since otherwise a player arriving at r could increase expected utility by postponing arrival to $t = T$. Thus, for no queue discipline can equilibrium utility be strictly lower than $w(T)$, and FIFO therefore minimizes equilibrium utility among work-conserving queue disciplines. □

3.3 Last-in-first-out (LIFO) queue discipline

Intuitively speaking, the problem with the FIFO discipline is that the strict queue discipline gives the agents an incentive to join the queue early, which in the end will hurt all agents in equilibrium. Below, we will show that the last-in-first-out (LIFO) queue discipline is not only better than the FIFO discipline, it is in fact welfare optimal among all queue disciplines that do not waste capacity. In our setting, the LIFO queue discipline works as follows. It always gives highest priority to those agents who have arrived the latest: When agents arrive continuously at a slower rate than capacity, they are all served immediately. If they arrive continuously at a higher rate than capacity, a fraction of agents are served immediately corresponding to capacity, while the rest must wait to be served until those arriving later have all been served.[2]

Before we turn to investigating equilibrium utility under LIFO, we provide some preliminary observations on equilibrium CADs and prove uniqueness and existence of an equilibrium under LIFO.

Lemma 6. *Let R be an equilibrium cumulative arrival distribution under LIFO, and let $r = \min\{t|R(t) = 1\}$. Then (a) $R(0) = 0$, (b) R is Lipschitz continuous on $[0, r[$, (c) $R'_+(t) > k$ for all $0 \leq t < r$, (d) $r < T$.*

Proof. (a) In equilibrium we cannot have $R(0) > 0$, since then there is a $\varepsilon > 0$ (sufficiently small) such that an agent arriving at time 0 would be better off by arriving ε later.

(b) First, observe that R is continuous on $[0, T]$, since if the CAD has a jump at time t, there is some $\varepsilon > 0$ (sufficiently small) such that an agent arriving at time t would be better off by arriving ε later. Let b be a fixed constant, where $0 < b < T$, and let $\theta > 0$ such that $0 < b - \theta < b$. We show that R is Lipschitz continuous on $[0, b - \theta]$. For this, note that since R is continuous, $R'_+(s)$ exists for each $s \in]0, T]$. It is therefore sufficient to show that $R'_+(s)$ is bounded. In (c) we show that $R'_+(t) > k$ for all $0 < t < r$, so here we focus on showing that $R'_+(s)$ is bounded from above, i.e. there is $K > 0$ such that $R'_+(s) < K$ for all $s \in [0, b - \theta]$. For this, suppose on the contrary that there is a sequence $s_1, s_2, ...$ in $[0, b - \theta]$ such that $R'_+(s_1) < R'_+(s_2) < ...$ and $R'_+(s_h) \to \infty$ for $h \to \infty$. Since $s_h \leq b - \theta$ an agent arriving at s_h who is not served immediately must wait for a period of time of at least θ to be served at some time after r. Since the probability of being served immediately goes to 0 as $h \to \infty$, the expected utility of an agent arriving at time s_h falls to a level below that of an agent arriving at time r (who is being served immediately with certainty). This contradicts that R is an equilibrium. Thus, $R'_+(s)$ is bounded from above on $[0, b - \theta]$ and the conclusion follows.

(c) If $R'_+(t) \leq k$ for some $0 \leq t < r$, it means that an agent arriving at time t is served immediately with certainty and thus obtains higher expected utility than an agent arriving at time r. This contradicts that R is an equilibrium CAD.

[2] Thus, in general an agent is facing a lottery over service times with two possible outcomes: either the agent is serviced immediately at arrival, or the agent is serviced when everyone who arrives later has been served.

(d) Since R is Lipschitz continuous (and hence absolutely continuous) we have $R(t) = \int_0^t R'_+(s)ds$. By (c), the desired conclusion follows. □

Uniqueness and existence of an equilibrium is established in the following lemmas.

Lemma 7. *Under the LIFO queue discipline, there is at most one equilibrium.*

Proof. Suppose, by contradiction, that R and Q are equilibrium cumulative arrival rates, $R \neq Q$.

Let $r = \min\{t|R(t) = 1\}$ and $q = \min\{t|Q(t) = 1\}$. We consider two cases:
(i) $r < q$
(ii) $r = q$.
(The case $r > q$ is symmetric to (i) and is thus omitted).

Ad. (i): Since the agents arriving at times r and q are served immediately in the two distributions, respectively, the expected utility for agents in R is greater than for the agents in Q.

Let $s = \max\{t|R(t) = Q(t), t < q\}$. Since $R(r) = 1 > Q(r)$, and Q and R are continuous functions, s is well defined. Moreover, we have $R'_+(q) > Q'_+(s)$, contradicting that expected utility is higher at R, since the agents not served at time s will be served at the same later time for both R and Q and the probability of being served at time s is lower in R than in Q.

Ad. (ii): Since the agents arriving at time $r(= q)$ are served immediately, the expected utility for agents arriving at time $r(= q)$ is the same for both arrival profiles, and thus expected utility in equilibrium is the same for both profiles. Since $R \neq Q$, $R(0) = Q(0)$, and $R(r) = Q(q)$, there is some t such that (a) $R(t) > Q(t)$ and $R'_+(t) < Q'_+(t)$ or (b) $Q(t) > R(t)$ and $Q'_+(t) < R'_+(t)$. If (a) then an agent arriving at t is served immideately with higher probability at R compared to Q, and if the agent is not served when arriving at t then he will be served earlier at R than at Q since less people will arrive afterwards. This contradicts that R and Q provide the same ex ante utility. A symmetric argument holds in case (b). □

Lemma 8. *There exists a cumulative arrival distribution that is a Nash equilibrium under LIFO.*

Proof. Let b be a fixed constant, where $0 < b < T$. Let v denote the straight line going through the points $(b, 0)$ and $(b, 1)$. Let l denote the straight line that goes through $(b, 1)$ with slope k.

Now, for each $s \in [0, T]$, we define a sequence of CAD functions $Q_b^1(s), Q_b^2(s)$,... as follows.

For each $s \in [0, b]$, let $p^1(s)$ be (uniquely) determined such that the expected utility of an agent arriving at time s and being served immediately with probability $p^1(s)$ and otherwise served at time T with probability $1 - p^1(s)$ is equal to the utility of an agent arriving at time b and being served immediately with certainty. Let $r^1(s) =$

$k/p^1(s)$. (Note that if agents arrive at the rate $\alpha(s) > k$, the probability of being served immediately is $k/\alpha(s)$).

Since $w(s)$ is continuous, $r^1(s)$ and $p^1(s)$ are continuous too. Note also that we have $r^1(s) \geq k$ for all s. Define $Q_b^1(s)$ on $[0,T]$ such that

$$Q_b^1(s) = \int_0^{\sigma_b(1)} r^1(t)dt$$

on $[0, \sigma_b(1)]$ where $\sigma_b(1)$ is defined as the first point s where the graph of $\int_0^s r^1(t)dt$ hits the line l, $Q_b^1(s)$ is identified as the line l from $\sigma_b(1)$ up to the point b, and $Q_b^1(s) = 1$ for $s \geq b$. By the Fundamental Theorem of Calculus, $Q_b^1(s)$ is differentiable (and hence Lipschitz continuous) on $]0, \sigma_b(1)[$.

Now, we define $Q_b^h(s)$ recursively as follows.

Suppose that a CAD $Q_b^{h-1}(s)$ and a point $\sigma_b(h-1)$ has been defined such that $0 < \sigma_b(h-1) \leq b$, $Q_b^{h-1}(s)$ is differentiable on $0 \leq s < \sigma_b(h-1)$ and the derivative on this domain is greater than or equal to k, $Q_b^{h-1}(s)$ is identified with the line l for $\sigma_b(h-1) \leq s \leq b$, and $Q_b^{h-1}(s) = 1$ for $s \geq b$. For each $s \in [0,b]$, let $\beta^{h-1}(s)$ denote the point in time where the straight line from $(s, Q_b^{h-1}(s))$ with slope k meets the horizontal line connecting $(0,1)$ and $(T,1)$. Now, let $p^h(s)$ be (uniquely) determined such that the expected utility of an agent arriving at time s and being served immediately with probability $p^h(s)$ and otherwise served at time $\beta^{h-1}(s)$ with probability $1-p^h(s)$ is equal to the utility of an agent arriving at time b and being served immediately with certainty. Let $r^h(s) = k/p^h(s)$. Define $Q_b^h(s)$ such that

$$Q_b^h(s) = \int_0^{\sigma_b(h)} r^h(t)dt,$$

where $\sigma_b(h)$ is the first point s where the graph of $\int_0^s r^h(t)dt$ hits the line l, $Q_b^h(s)$ is the line l from the point $\sigma_b(h)$ up to b, and $Q_b^h(s) = 1$ for $s \geq b$. Since $w(s)$ is continuous and $Q_b^{h-1}(s)$ is continuous on $[0,b[$, $r^h(s)$ and $p^h(s)$ are continuous on $[0,b]$. Note that $Q_b^h(s)$ is non-decreasing and by the Fundamental Theorem of Calculus it is differentiable (and hence Lipschitz continuous) on $]0, \sigma_b(h)[$.

Moreover, it follows from the recursive construction that $\sigma_b(1) \geq \sigma_b(2) \geq ...$, $Q_b^1(s) \leq Q_b^2(s) \leq ...$ for all $s \in [0,T]$, and $r^1(s) \leq r^2(s) \leq ...$ for all $s \in [0,b]$.

We have $\lim_{h\to\infty} \sigma_b(h) > 0$, since if $\lim_{h\to\infty} \sigma_b(h) = 0$, the highest derivative of $Q_b^h(s)$ on $[0, \sigma_b(h)]$ would go to infinity as $h \to \infty$ implying that the expected utility as calculated above of a h-sequence of agents $[0, \sigma_b(h)]$ would go to zero, contradicting the construction of the sequence $Q_b^1(s), Q_b^2(s), ...$

Also, for each $0 \leq s \leq \lim_{h\to\infty} \sigma_b(h)$, $\lim_{h\to\infty} r^h(s)$ is finite, since if $\lim_{h\to\infty} r^h(s) = \infty$ it implies that the expected utility for agents arriving at s as calculated above goes to zero, implying a contradiction.

In what follows, let $\overline{Q}_b(s) = \lim_{h\to\infty} Q_b^h(s)$ and $\overline{\sigma}_b = \lim_{h\to\infty} \sigma_b^h(s)$.

We now make the following observations:

(i) For b sufficiently close to T, $\overline{\sigma}_b < b$. This follows from observing that $Q_b^h(s)$ is increasing in b.

(ii) For b sufficiently close to 0, $\overline{\sigma}_b = b$. This follows (again) from observing that $Q_b^h(s)$ is increasing in b (since it means it decreases when b goes to zero).

(iii) $\overline{\sigma}_b$ is continuous in b. This follows since $w(s)$ is continuous and by the construction of $\overline{\sigma}_b$ and $\overline{Q}_b$.

Combining (i),(ii) and (iii), we get that there exists b such that $\overline{\sigma}_b = b$ and $\overline{Q}_b(s)$ is continuous on $[0, T]$. By construction, with the CAD $\overline{Q}_b(s)$ the expected utility for an agent arriving at any time s with $\leq s < b$ is equal to the utility of an agent arriving at b (who is served immediately with certainty) and thus $\overline{Q}_b(s)$ is an equilibrium CAD under LIFO. □

Having established existence and uniqueness of LIFO, we now move on to establish the welfare properties.

Lemma 9. *Let $R(t)$ be an equilibrium arrival distribution under LIFO, and let $Q(t)$ be an equilibrium under some other queue discipline that gives higher welfare. Then $q < r$, where $q = \min\{t|Q(t) = 1\}$, and $r = \min\{t|R(t) = 1\}$, i.e., the latest arriving agent according to $Q(t)$ arrives earlier than the latest arriving agent according to $R(t)$.*

Proof. An agent arriving at r under LIFO is served immediately. Thus, since equilibrium utility is lower under the LIFO discipline, the agent arriving at s under the alternative discipline must be served (and hence must have arrived) earlier than r, as illustrated in Figure 4. □

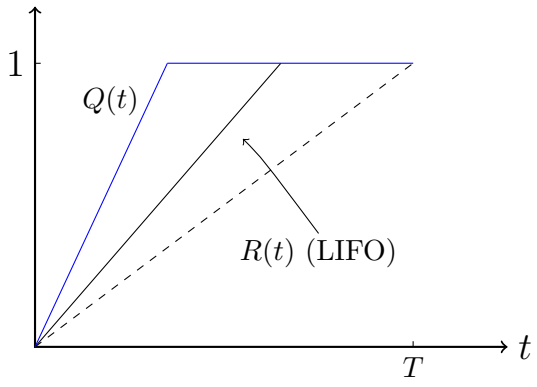

Figure 4: LIFO and possible eq. under other discipline giving higher welfare

Lemma 10. *Let $R(t)$ be an equilibrium CAD under LIFO, and let $Q(t)$ be an equilibrium under some other queue discipline, where $Q(\bar{t}) = R(\bar{t})$ for some $\bar{t} < T$ and $Q(t) \geq R(t)$ for all $\bar{t} \leq t \leq T$ with strict inequality for some interval of time. Then the equilibrium utility (and welfare) is higher under LIFO .*

Proof. Under LIFO, the agents arriving from $\bar{t}$ or later get priority over those who arrived earlier but have not yet been served. Thus, the equilibrium utility for this group of agents must be at least as high as the equilibrium utility for the group of agents arriving from $\bar{t}$ or later under the queue priority supporting distribution $Q(t)$. □

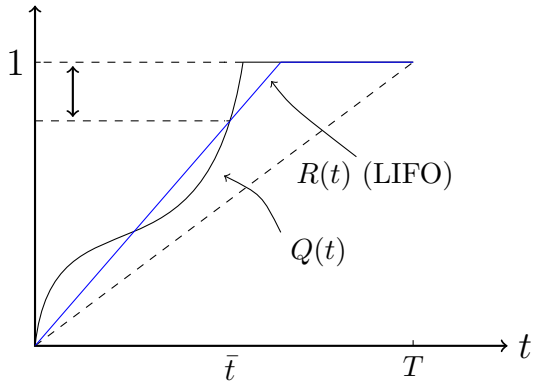

Figure 5: Higher welfare under LIFO

It follows readily from lemmas 9 and 10 that no queue discipline supports an equilibrium giving higher welfare, and we can therefore state the following proposition.

Proposition 2. *LIFO maximizes welfare.*

3.4 Welfare properties

Theorem 1. *The equilibrium utility (and welfare) of any work-conserving stochastic queue discipline is bounded from above by the LIFO and from below by the FIFO queue discipline. In other words, within the family of work-conserving queue disciplines, FIFO gives the lowest equilibrium utility while LIFO gives the highest equilibrium utility.*

Proof. Follows immediately from propositions 1 and 2. □

4 Concluding remarks

In this paper, we modeled a service facility with a given opening time that serves impatient agents with a fixed capacity. With linear waiting costs our conclusion is as follows: FIFO minimizes the equilibrium welfare while LIFO maximizes it.

Within the scope of the model, our results imply that the traditional use of FIFO as queuing discipline for common service facilities, such as airport passengers waiting to board a flight and morning traffic where drivers arrive at the bottleneck after a given time, provides the worst possible level of welfare. In principle, LIFO should instead be used as it guarantees an optimal queue discipline. In a sense, this shows why the traditional queue discipline might, counterintuitively perhaps, be a curse rather than a blessing.

However, in real-world applications, issues like fairness and practical service setup for facilities cannot be ignored, and in this respect implementation of LIFO features some obvious challenges.

In terms of perceived fairness, LIFO might be considered inferior to FIFO due to the cultural and traditional habit of using FIFO as the common queuing discipline. However, as no agents are discriminated on individual characteristics, LIFO is arguably not unfair by construction. From an ex-post perspective, agents may have different perceptions of fairness, depending on whether they are serviced immediately at arrival (thereby reaping the benefits of the priority to late arrivals), or whether they are forced to wait for a longer period while later arrivals are serviced before them. In the former case, the agent will most likely appreciate the fast service and still perceive LIFO as fair, while he might feel unfairly treated in the latter case when later arriving agents are being served before him. In any case, the possible difference in the perception of fairness strictly depends on the cultural consensus of fair queuing.

In terms of implementing the LIFO discipline, some types of facilities would obviously face difficulties in connection with the practical execution of service. These would include some queuing facilities where agents have to psychically line up as in the case of agents exiting a parking lot. In that situation, LIFO may not be a relevant alternative due to physical restrictions when cars line up at the bottleneck. In some of these situations, a compromise between FIFO and LIFO, like service in random order, might be a preferred alternative. However, LIFO is more easily implemented into other types of service facilities. Take for example telecommunication queues, e.g., customers calling a support center, or patients calling their doctor to make an appointment in the period after telephones open. In such cases, the phone support software can more easily be set up to service the latest arriving customer in the phone queue rather than the first arriving customer.

In our framework, we considered agent preferences to be identical and associated with a linear cost of waiting. A possible generalization of the model could be to allow for different types of agents. One approach is to characterize an agents utility function with type specific parameters for cost and willingness-to-pay. All agents are assigned their type randomly from a known probability function, and the type is private information only to the agent. A similar approach was taken by Jain et al. (2011), although they only assigned the cost functions with type specific parameters. Furthermore, one could consider a general functional form of the cost function. As an example, it seems plausible that an agents marginal cost of waiting will increase with time spent in a queue. Thus, the construction of a cost function that increases over time might add to the relevancy

of the model.

As part of future research, further extensions could be of interest, e.g. extension to multi-service facilities, allowing for early arrivals, and considering service facilities with both an opening and closing time. Moreover, one could study the equilibrium welfare for other queuing disciplines. This paper only studied LIFO and FIFO, while disciplines like Priority Service and Service in Random Order (SIRO) could be of interest. Finally, empirical research based on behavioral experiments or numerical simulations could be of interest. We hope this paper will motivate further research in the welfare analysis of stochastic queuing systems.

References

[1] Arnott R.A., de Palma A., and R. Lindsey, 1993. A structural model of peak-period congestion: A traffic bottleneck with elastic demand. *The American Economic Review*, 83, 161-179.

[2] de Palma, A. and M. Fosgerau, 2009. Random queues and risk averse users. Working paper.

[3] de Palma, A, and Fosgerau, M, 2011. Dynamic Traffic Modeling, in A de Palma, R Lindsey, Quinet, E. and R. Vickerman (eds), A Handbook of Transport Economics. Edward Elgar Publishing, Incorporated.

[4] Glazer, A. and R. Hassin (1983) ?/M/1: On the equilibrium distribution of customer arrivals. *European Journal of Operations Research*, 13, 2, 146-150.

[5] Hassin, R., 1985. On the optimality of first come last served queues. *Econometrica*, 53, 1, 201-202.

[6] Hassin, R. and M. Haviv, 2003. To Queue or not to Queue: Equilibrium Behavior in Queueing Systems.

[7] Hassin, R. and Y. Kleiner, 2011. Equilibrium and optimal arrival patterns to a server with opening and closing times. IEE transactions, 43, 3, 164-175.

[8] Jain, R., Juneja, S., and N. Shimkin, 2011. The concert queueing problem: to wait or to be late. *Discrete Event Dynamic Systems*, 21, 103-138.

[9] Naor, P. 1969. The regulation of queue size by levying tolls. *Econometrica*, 37, 1, 15-24.

[10] Vickrey, W.S., 1969. Congestion theory and transport investment. *The American Economic Review*, 59, 2, 251-260.