

# The Concert Queueing Game: To Wait or To be Late\*

Rahul Jain<sup>†</sup>, Sandeep Juneja<sup>‡</sup> & Nahum Shimkin<sup>§</sup>

July 2, 2010

## Abstract

We introduce the concert (or cafeteria) queueing problem: A finite but large number of customers arrive into a queueing system that starts service at a specified opening time. Each customer is free to choose her arrival time (before or after opening time), and is interested in early service completion with minimal wait. These goals are captured by a cost function which is additive and linear in the waiting time and service completion time, with coefficients that may be class dependent. We analyze the system in the many-customer asymptotic regime and develop a fluid limit for the resulting queueing system. We consider a fluid model of this system, which is motivated as the fluid-scale limit of the stochastic system. In the fluid setting, we explicitly identify the unique Nash-equilibrium arrival profile for each class of customers. Our structural results imply that, in equilibrium, the arrival rate is increasing up until the closing time where all customers are served. Furthermore, the waiting queue is maximal at the opening time, and monotonically decreases thereafter. In the simple single class setting, we show that the price of anarchy (PoA, the efficiency loss relative to the socially optimal solution) is exactly two, while in the multi-class setting we develop tight upper and lower bounds on the PoA. In addition, we consider

---

\*A preliminary version of this paper appeared as [10].

<sup>†</sup>EE & ISE departments, Viterbi School of Engineering, University of Southern California, Los Angeles, CA. email: rahul.jain@usc.edu

<sup>‡</sup>School of Technology and Computer Science, Tata Institute of Fundamental Research, Mumbai, India. email: juneja@tifr.res.in

<sup>§</sup>Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel. e-mail: shimkin@ee.technion.ac.il

several mechanisms that may be used to reduce the PoA. The proposed model may explain queueing phenomena in diverse settings that involve a pre-assigned opening time.

# 1 Introduction

In this paper, we introduce *the concert queueing game*. This model is motivated by the following scenario. Before going to, say a rock concert in a theater with unassigned seats where a large queue may be anticipated, one faces the following dilemma: Should one go late when queues are smaller and hence waiting is less but the best seats are already taken, or go early to secure the best seats but possibly wait in a long queue? Similarly, in cafeterias that open at a specific time for lunch, should one go early when one is really hungry but the queues may be longer, or go hungry longer to avoid the long queues? Similar trade-offs govern customer decisions in many queueing situations such as visiting a retail store on the day of a huge sale, visit to the DMV office, to a movie theater, etc. In communication networks, for example, congestion is typically not uniform throughout the day. There is greater traffic during working hours than at night. Similar, although not identical, tradeoffs may be found in diverse areas from transportation networks, regarding choice of travel timing, to Internet file downloads (which take longer during peak congestion hours).

The proposed modeling approach is meant to study the emerging system behavior when strategic players are faced by such utility vs. queueing delay trade-offs. We assume that there is a large but finite number of customers that need to be served by a server in a first-come first-served manner. The server at the queue becomes active at a particular time. Customers can choose to arrive and queue up both before and after that time. The cost structure of each customer is additive and linear in the waiting time and in the service completing time. Alternatively, a customer may be interested in the number of users served before her rather than the service completion time (see Remark 5 below). Multiple classes of customers are allowed that differ in their cost coefficients. We primarily focus on a finite number of classes, but also address briefly the same model with a continuum of classes.

The analysis in this paper is carried out within a fluid model, which is motivated as the fluid-scale limit of the stochastic queueing system with *prescribed* arrival timing. This fluid model offers a great deal of analytical simplification. The game of arrivals defined over this model belongs in the class of non-atomic games [16], where each customer is infinitesimal

and therefore his effect on the others is negligible. We show that this game has a *unique* Nash equilibrium point (in terms of the aggregate arrival profile), and explicitly identify this point.

An important property of any equilibrium solution is the social efficiency loss it entails, when compared to the social optimum. A popular measure of this loss is the *price of anarchy* (PoA), which equals the maximum of the ratio of the social cost of the equilibrium solution to that of the socially optimal one among all equilibria. In our model, we show that the PoA in the single class setting is exactly 2 for all parameter values. In the multi-class setting, we develop tight upper bounds and corresponding lower bounds on the PoA that depend only on the range of the cost parameters across customer classes. Furthermore, we consider several mechanisms that can be used to reduce the PoA; the analysis is carried out within the single-class setting for simplicity. These mechanisms include service time restrictions, assigning priorities to certain segments of the populations, or charging tariffs that depend on the time of service. We observe that by sufficiently segmenting the population optimally along any of these three ways, the PoA can be brought arbitrarily close to one and we identify the associated rate of decline.

Strategic queueing problems that involve self-optimizing customers have been extensively studied for over four decades, spanning problems of admission control, routing, renegeing, choice of priorities, pricing, and related issues. A sizable part of this literature is summarized in the monograph [4]. A central issue in this context is the comparison of the individual equilibrium and the socially optimal solution. This may be traced back to Naor’s seminal paper [12], which considers these two solutions in the context of admission control to a single-server queue, and suggests pricing as a means to induce the social optimum. Recently, [6] provided bounds on the PoA for the problem of routing into  $n$  parallel servers, and [2] studied the PoA in Naor’s model. In [7], an interesting perspective is taken wherein time of arrival of customers is determined through a first-price auction.

Equilibrium arrival patterns to queues with a finite service period were apparently first considered in [3], where a Poisson-distributed number of homogeneous customers may choose their arrival times with the goal of minimizing their waiting time. In this model, customers are indifferent to the time-of-day when their service is completed. Several extensions and variations of this model have been considered, e.g., in [14, 8, 5], and are further described in [4] (Chapter 6) and [5]. The model of [19] incorporates preferences for early service within a multiple shift scheme, where the service period is divided into evenly-spaced shifts, and the waiting time in each shift is determined by the number of customers who choose this

shift.

A related body of research exists in the transportation literature, where equilibrium trip-timing patterns were extensively studied in the context of the so-called bottleneck or morning commute problem. [18] introduced a fluid flow model, where homogeneous commuters choose their departure time for travel through a single bottleneck of fixed flow capacity. The cost function for each commuter includes a penalty for arriving early or late to the destination (relative to the desired arrival time), in addition to the cost of delay in the bottleneck. Pointers to the extensive ensuing literature regarding this model and its generalizations may be found in [11]. In particular, [13] introduced commuter heterogeneity in terms of their (linear) cost coefficients, in addition to the required arrival time. [11] provides existence and uniqueness results for the multiclass model with nonlinear costs, under fairly general conditions. We note that our fluid model can be considered a special case of these models with the desired arrival times all set to zero though the bottleneck model does not have a predetermined opening time. However, the explicit expressions presented here for the equilibrium as well as the analysis of the PoA and the ways to reduce it are new.

The organization of this paper is as follows: In Section 2 we describe our model. We start with a brief description of the stochastic queueing system and its fluid scale analysis, followed by a description of the arrival game for the fluid model. In Section 3 we focus on the single-class case, for which the results are particularly simple, and show that unique Nash equilibrium corresponds to a uniform arrival profile over a finite interval. We generalize the results to the multi-class settings in Section 4 where we consider finite number of classes. In both Sections 3 and 4, we also compute and bound the price of anarchy for the derived equilibrium. We briefly discuss the generalization to a continuum of classes in Section 5. In Section 6 we discuss some ways to reduce the price of anarchy. In Section 7, we present numerical results for a simple experiment that suggest that the equilibrium arrival profile that is valid in the fluid regime may be close to equilibrium in the finite- $N$  queue, for  $N$  reasonably large. Finally, we end with a brief conclusion in Section 8.

## 2 Model Description

This section introduced the fluid model that we analyze in this paper. We start with a brief description of the underlying stochastic queueing system, discuss the fluid limit of this model (with fixed arrival patterns), and then describe the game model that we consider in

the rest of the paper.

## 2.1 The Stochastic Queueing System

Consider a queueing system that caters to a finite number  $N$  of customers, which are served on a first-come first-served basis.  $N$  may be random, with a finite mean  $E(N)$ . The required service times of these customers form an i.i.d. sequence  $(V_j : 1 \leq j \leq N)$  that may have a general distribution with rate  $\mu = 1/E(V_j)$ . Each customer  $j$  independently picks his arrival time, as a sample from a probability distribution with CDF<sup>1</sup>  $F_j(\cdot)$ . Service starts at time  $t = 0$ , and commences until all customers are served. Customers may arrive and queue up both before and after  $t = 0$ . For simplicity, assume at the outset that each  $F_i$  is supported on a finite time interval.

Suppose that customers wish to be served as early as possible, while minimizing their waiting time. We capture these (possibly conflicting) goals through a linear cost function. Let

$$c_j(w_j, \tau_j) = \alpha_j w_j + \beta_j \tau_j$$

denote  $j$ 's cost function, where  $w_j$  is his waiting time in the queue,  $\tau_j$  his service completion time, and  $\alpha_j > 0$ ,  $\beta_j > 0$  are the respective cost sensitivities. Given the collection  $\mathbf{F} = \{F_j\}$  of arrival time distributions for all customers, both  $w_j$  and  $\tau_j$  become well-defined random variables, and we may consider the expected cost  $C_{\mathbf{F}}^j = E_{\mathbf{F}}(\alpha_j w_j + \beta_j \tau_j)$ , where  $E_{\mathbf{F}}$  is the expectation induced by  $\mathbf{F}$ .

As usual, we say that the collection  $\mathbf{F} = \{F_j\}$  of arrival time distributions is a **Nash equilibrium point** (NEP) for this problem if no user  $j$  can reduce his expected cost  $C_{\mathbf{F}}^j$  by unilaterally modifying his arrival time distribution  $F_j$ . Our goal is to characterize these NEPs and study their properties. This will be done within an approximate fluid model, which greatly facilitates the analysis and leads to explicit solutions. We thus turn to consider the fluid approximation of this system.

**Remark 1** The cost parameters  $(\alpha_j, \beta_j)$  of each customer define his *type* (or *class*). Customer classes will be defined more precisely within the fluid model. The above description assumes that the customer types are deterministic (or, equivalently, that the number of

---

<sup>1</sup>Throughout the paper, we describe probability measures (and, more generally, positive measures) on the real line by their cumulative distribution function (CDF). Thus,  $F(t)$  corresponds to the measure  $m_F$  with  $m_F\{(-\infty, t]\} = F(t)$ . We shall use the term  $F$ -measure (of a given set) to refer to the  $m_F$ -measure of that set.

customers of each type is fixed). Alternatively, the type of each customer may be randomly selected according to a known probability distribution over a given set of types. The arrival distribution of each customer may naturally depend on his type (which is considered private information). The expectation operator  $E_{\mathbf{F}}$  should then account also for averaging over the type distribution of all (other) customers.

## 2.2 The Fluid Limit

This section motivates the fluid model that is the main subject of his paper, by considering the fluid-scale limit of the above-described queueing system as the number of arrivals becomes large. Our discussion here is limited to the case where the arrival time distributions of all customers are given. Thus, we do not consider the formal relation between the finite-customer game and its fluid limit, which requires a separate and more refined treatment.

Consider a sequence of queueing systems indexed by  $n \geq 1$ , defined on a common probability space, and let  $N^n$  be the population size (number of customers) in the  $n$ -th system. Assume that  $\lim_{n \rightarrow \infty} \frac{N^n}{n} = \Lambda > 0$  (with probability 1). Let  $F_i^n(\cdot)$  denote the arrival time distribution for customer  $i$  in the  $n$ -th system. The service parameters are as described above. In particular, the service time distribution does not depend on  $n$ , and has rate  $\mu$ .

Let  $F^n(t) = \sum_{i=1}^{N^n} F_i^n(t)$  denote the *aggregate* arrival profile in the  $n$ -th system. Suppose that the collection  $\{F_i^n\}$  is given so that

$$\frac{1}{n} F^n(nt) \rightarrow F(t) \tag{1}$$

as  $n \rightarrow \infty$ , uniformly on compact sets (u.o.c.), where  $F(\cdot)$  is the *fluid arrival profile*. It follows that  $F(\cdot)$  is the CDF of a positive measure on the real line with total mass  $\Lambda$ . More specifically,  $F$  is non-decreasing, right-continuous, with  $F(-\infty) = 0$  and  $F(\infty) = \Lambda$ .

Note that the time axis is scaled by a factor of  $n$  in (1). This accounts for the increase in the overall service time requirement of all  $N^n$  customers in the  $n$ -th system. Importantly, under this time scaling the service time of a single customer diminishes to zero as  $n$  increases.

We further observe that the same fluid arrival profile  $F$  can arise from different choices of individual arrival distributions, ranging from i.i.d. arrival times to deterministic ones. The following simple example illustrates this point.

**Example 1** Suppose  $F$  is a uniform distribution on  $[-T, T]$  with mass  $\Lambda$ , namely  $F(t) = \Lambda \frac{t+T}{2T}$  on that interval. Then,  $F$  can arise out of each of the following possibilities.

- a. *IID arrivals*: Each  $F_i^n$  corresponds to a uniform distribution on  $[-nT, nT]$  for some  $T > 0$ , namely  $F_i^n(t) = \frac{t+nT}{2nT}$  on  $[-nT, nT]$ . Then,  $\frac{1}{n}F^n(nt) = \frac{N^n}{n} \left( \frac{t+T}{2T} \right)$  and this converges to  $F(t)$ , almost surely, u.o.c. as  $n \rightarrow \infty$ .
- b. *Deterministic arrivals*:  $F_i^n$  corresponds to the deterministic arrival time  $t_i = (2i-n)T$ . Equivalently,  $F_i^n(t) = \mathbf{1}\{t \geq (2i-n)T\}$ , where  $\mathbf{1}\{\cdot\}$  denotes an indicator function.

We next consider the queue-length process and its fluid limit. Let  $Q^n(t)$  denote the queue length at time  $t$  (including the customer in service), namely,  $Q^n(t)$  is the cumulative number of arrivals minus service completions up to and including time  $t$ , and let

$$\bar{Q}^n(t) = \frac{Q^n(nt)}{n}$$

denote its scaled version. To specify the fluid limit of  $Q^n$ , let  $S(t) = \mu t \mathbf{1}\{t \geq 0\}$  denote the fluid-scale potential service process (recall that service starts at  $t = 0$ ). Also define  $X(t) = F(t) - S(t)$ , and

$$Q(t) = X(t) + \sup_{0 \leq s \leq t} [-X(s)]^+. \quad (2)$$

Then, by standard results (see, for example, [1], Theorem 6.5 and its proof) it follows that as  $n \rightarrow \infty$ , the following process-level convergence of the scaled queueing process:

$$\bar{Q}^n(\cdot) \rightarrow Q(\cdot)$$

holds, almost-surely, u.o.c. This process-level convergence result evidently relies on the functional strong law of large numbers.

The limit queue process  $Q(t)$  corresponds to a fluid system with deterministic input and output streams of fluid. The cumulative arrival process is given by  $F(t)$ , and the service rate is  $\mu(t) = \mu \mathbf{1}\{t \geq 0\}$ . This fluid model will be the subject of our subsequent analysis.

## 2.3 The Multiclass Fluid Model

We proceed to describe the concert arrival game for the fluid model with a finite number of customer classes. The customer population is represented by the set  $[0, \Lambda]$ , where  $\Lambda$  stands for the total workload, and each customer corresponds to a single point in this interval.

These infinitesimal customers arrive at a service facility with potential service rate  $\mu$  (in terms of fluid units per unit time), that activates at time  $t = 0$ . Thus, all customers may be served within  $T_f = \Lambda/\mu$  time units. All customers join a single queue, and are served in the order of their arrival. If a non-zero mass of customers arrives simultaneously (represented by a jump in  $F(t)$ ), then their queueing order is determined randomly and with symmetric probabilities.

Customers may belong to different classes, which differ in terms of their cost parameters. Let  $\mathcal{I} = \{1, 2, \dots, I\}$  denote the set of customer classes. For each class  $i \in \mathcal{I}$ , let  $\Lambda_i$  denote the total workload carried by its members. Thus  $\sum_i \Lambda_i = \Lambda$ , and serving all class  $i$  customers requires  $\Lambda_i/\mu$  time units. The cost function for a class  $i$  customer is given by

$$C_i(w, \tau) = \alpha_i w + \beta_i \tau$$

where  $w$  is this customer's waiting time in the queue,  $\tau \geq 0$  his service completion time, and  $\alpha_i > 0$ ,  $\beta_i > 0$  are the respective cost sensitivities to the waiting time and service completion time.

Consider a customer who arrives at time  $t$  and is placed at the end of a queue of size  $q$ . His waiting time will be  $w = q/\mu + \max\{0, -t\}$  so that he completes his service and leaves the system at  $\tau = t + w = q/\mu + \max\{0, t\}$ . (Note that the service time of individual customers is null since customers are infinitesimal.)

Let  $F_i$  denote the *class- $i$  arrival profile*. It is the CDF of a positive measure on the real line with total mass  $\Lambda_i$ . Thus,  $F_i(-\infty) = 0$ ,  $F_i(\infty) = \Lambda_i$  and  $F_i(t)$  is right-continuous and non-decreasing in  $t$ . An *arrival profile* is the collection  $\{F_i\}$  of arrival profiles, one for each classes. The sum  $F(t) = \sum_i F_i(t)$  denotes the *aggregate arrival profile*. As discussed in Section 2.2, an arrival profile  $F_i$  should be interpreted as a deterministic summary of the arrival decisions of the individual customers, which may themselves be deterministic or stochastic. The following restriction applies to each  $F_i$ .

**Remark 2** To avoid lingering over some mathematical subtleties, we shall assume at the outset that the measure represented by  $F_i$  has no singular continuous component, and is therefore the sum of an absolutely continuous component and a discrete component (see [15], Pg. 108-113, for instance).

Given the aggregate arrival profile  $F = \sum_i F_i$ , the queue-size process  $Q(t)$  is uniquely defined by equation (2). Therefore, the expected waiting time  $W(t)$  of a potential arrival at time  $t$  is well defined as well. Specifically, if  $Q(t)$  is continuous at  $t$ , then the waiting



time is deterministic and given by  $W(t) = Q(t)/\mu + \max\{0, -t\}$ . If  $Q(t)$  has a jump at  $t$  (due to an upward jump in the arrival profile  $F$ ), then the position of an arriving customer would be uniformly distributed in  $[Q(t-), Q(t+)]$  with average  $\bar{Q}(t) = \frac{1}{2}(Q(t-), Q(t+))$ , so that the expected waiting time is  $W(t) = \bar{Q}(t)/\mu + \max\{0, -t\}$ . Let  $W_F(t)$  denote the expected waiting time that corresponds to a given arrival profile  $F$ .

The expected cost of a class  $i$  customer that arrives at  $t$  is now given by

$$C_F^i(t) = \alpha_i W_F(t) + \beta_i(t + W_F(t)). \quad (3)$$

More generally, the expected cost incurred by a class  $i$  customer who selects his arrival by sampling from probability distribution  $G$  is

$$\mathcal{C}_F^i(G) = \int_{-\infty}^{\infty} (\alpha_i W_F(t) + \beta_i(t + W_F(t))) dG(t).$$

We proceed to define the Nash equilibrium for the induced game. A *multi-strategy* for this game is a collection  $\{G_s(\cdot), s \in [0, \Lambda]\}$  of probability distributions on the real line, one for each customer  $s$ , represented by their CDFs.

**Definition 1** A multi-strategy  $\{G_s(\cdot), s \in [0, \Lambda]\}$  is a Nash equilibrium point if

(i)  $F(t) = \int_0^\Lambda G_s(t) ds$  is well defined for each  $t$ , and

(ii) For any customer  $s \in [0, \Lambda]$  of class  $i$ ,

$$\mathcal{C}_F^i(G_s) \leq \mathcal{C}_F^i(\tilde{G}), \quad \text{for every CDF } \tilde{G}.$$

That is, no customer  $s$  can improve his cost by modifying his own arrival time distribution. Note that this definition makes use of the fact that the action of a single (infinitesimal) customer does not affect the arrival profile  $F(t)$ . This property is shared by the class of non atomic anonymous games (cf. [16]), to which the present model belongs.

The specific consideration of each customer in the last definition is too detailed for our purpose. A more useful definition may be given in terms of the class arrival profiles.

**Definition 2** An arrival profile  $\{F_i, i \in \mathcal{I}\}$  is an **equilibrium profile** if, for each class  $i$ , there exists a set  $\mathcal{T}_i$  of  $F_i$ -measure  $\Lambda_i$  on which  $C_F^i(t)$  is minimal, namely,

$$C_F^i(\tau) \leq C_F^i(t) \quad \text{for all } \tau \in \mathcal{T}_i \text{ and } -\infty < t < \infty.$$

Essentially, this definition requires the cost  $C_F^i(t)$  to be minimal on the support of  $F_i$ .

The two definitions may be seen to be compatible in the following sense:

(i) First, given an equilibrium profile  $\{F_i, i \in \mathcal{I}\}$ , a compatible equilibrium multi-strategy  $\{G_s(\cdot), s \in [0, \Lambda]\}$  may be obtained (for example) by letting  $G_s = F_i/\Lambda_i$  for each customer  $s$  of class  $i$ . Thus, all customers of a given class  $i$  are assigned identical arrival distributions, which adds up to the given arrival profile  $F_i$  for that class. This immediately implies that  $F(t) \triangleq \int_0^\Lambda G_s(t)ds = \sum_i F_i(t)$ , and property (ii) of Definition 1 now follows since, for  $G_s = F_i/\Lambda_i$ , we get by Definition 2 that  $\mathcal{C}_F^i(G_s) = \min_t C_F^i(t)$ , while the latter is clearly not larger than  $\mathcal{C}_F^i(\tilde{G})$  for any CDF  $\tilde{G}$ .

(ii) Conversely, an equilibrium multi-strategy  $\{G_s(\cdot), s \in [0, \Lambda]\}$  induces a unique arrival profile for each class, given by  $F_i(t) = \int_0^\Lambda G_s(t)\mathbf{1}\{s \in S_i\}ds$ , where  $S_i$  is the set of class  $i$  customers. Now,  $\{F_i\}$  is an *equilibrium* profile. Indeed, by Definition 1(ii) it follows that, for each  $s \in S_i$ ,  $\mathcal{C}_F^i(G_s) = \min_t C_F^i(t)$ , hence there must exist a set of times  $\mathcal{T}_s$  of  $G_s$ -measure 1 on which  $C_F^i(t)$  attains that minimal value. Therefore  $C_F^i(t)$  is minimal also on the union  $\mathcal{T}_i = \bigcup_{s \in S_i} \mathcal{T}_s$ , while the  $F_i$ -measure of  $\mathcal{T}_i$  is  $\Lambda_i$ , since the  $G_s$  measure of  $\mathcal{T}_i$  is 1 for each  $s \in S_i$  (as  $1 \geq G_s(\mathcal{T}_i) \geq G_s(\mathcal{T}_s) = 1$ ). Thus, the requirements of Definition 2 are satisfied.

### 3 Analysis of the Single-Class Model

To bring out salient features of the analysis, we first consider the single-class case. Here all customers share the same cost parameters, and we may drop the class index  $i$  from the notation. The results in this case are particularly simple: The equilibrium arrival profile turns out to be a uniform distribution, and the price of anarchy exactly equals 2.

The following lemma will be useful in simplifying the expression for the cost function under equilibrium conditions. Some notation is introduced first. Recall that  $T_f = \Lambda/\mu$ , and let

$$t^* = \inf\{t \geq 0 : F(t) < \mu t\}.$$

This is the first time beyond 0 at which the server becomes starved.

**Lemma 1** *For any equilibrium arrival profile  $F$ ,*

- (i)  $t^* = T_f$  (i.e., the server works at full rate till the last customer is served).
- (ii) There are no point masses in  $F$ , so that  $F(t)$  is absolutely continuous in  $t$ .

(iii) For  $t \leq T_f$ ,

$$W_F(t) = F(t)/\mu - t. \quad (4)$$

As is apparent from the proof below, Condition (ii) in Lemma 1 is applicable even to the finite  $n$  queueing system (not just its fluid limit).

**Proof of Lemma 1:** (i) Clearly,  $t^* \leq T_f$ , since all customers are served by  $T_f$  at full service rate. Suppose that  $t^* < T_f$ . Then  $F$  cannot be an equilibrium arrival profile. To see this, note that  $Q(t^*) = 0$  by definition of  $t^*$ , so that  $W(t^*) = 0$ . Furthermore, since  $t^* < T_f$ , a positive mass of customers have not been served yet, and since  $Q(t^*) = 0$  these customers have not arrived by  $t^*$ , so that  $F(t^*) < \Lambda$ . Thus, those customers that arrive after  $t^*$  can improve their cost by arriving at  $t^*$  instead and getting served immediately. This implies that  $F$  cannot be an equilibrium profile.

(ii) Suppose that  $F$  has a point mass of size  $\lambda > 0$  at some  $t = t_1$ . Then, a customer that arrives at  $t_1$  sees, on average, half  $(\lambda/2)$  of the customers that arrive at  $t_1$  before her. However, by arriving at  $t_1 - \epsilon$  with  $\epsilon > 0$ , such a customer would arrive ahead of this bunch, thereby reducing its waiting time by  $\lambda/2\mu - \epsilon$  at least, and leaving earlier. Clearly, for  $\epsilon$  small enough this means that arriving at  $t_1$  is not optimal for such a customer. It follows that  $F$  has no point masses, namely no discrete component. Since  $F$  has no continuous singular component by assumption, it follows that  $F$  is absolutely continuous.

(iii) We have just established that  $F$  has no point masses. This implies that an arrival at  $t$  will see the entire queue  $Q(t)$  before him. For  $t < 0$ ,  $Q(t) = F(t)$ , and the equality in (4) follows since  $-t$  is the customer wait before the server becomes active, and  $F(t)/\mu$  is the remaining queueing delay once the server becomes active. For  $0 \leq t \leq T_f$ , (4) follows from part (i) of this Lemma as service proceeds at full rate in the interval  $[0, T_f]$ , which implies that  $Q(t) = F(t) - \mu t$ , while  $W(t) = Q(t)/\mu$ .  $\square$

It follows from Lemma 1 that under the equilibrium arrival profile  $F$ , the cost  $C_F(t)$  at any time  $t \leq T_f$  equals

$$C_F(t) = (\alpha + \beta)F(t)/\mu - \alpha t. \quad (5)$$

Let  $T_0 = -\frac{\Lambda\beta}{\mu\alpha}$ . The cost in (5) becomes independent of  $t$  for  $t \in [T_0, T_f]$  if we select  $F = F^*$  where  $F^*(t) = 0$  for  $t \leq T_0$ ,  $F^*(t) = \Lambda$  for  $t \geq T_f$ , and

$$F^*(t) = \Lambda \frac{t - T_0}{T_f - T_0}, \quad t \in [T_0, T_f].$$

In that case, (5) gives  $C_F(t) = \beta\Lambda/\mu = \beta T_f$  for  $t \in [T_0, T_f]$ .

**Theorem 1**  $F^*$  is the unique equilibrium arrival profile with  $T_0 = -\frac{\Lambda\beta}{\mu\alpha}$  and  $T_f = \frac{\Lambda}{\mu}$ .

**Proof:** We first verify that  $F^*$  is an equilibrium profile. First, as noted above,  $C_{F^*}(t) = \beta\Lambda/\mu \triangleq c_0$  for  $t \in [T_0, T_f]$ . For  $t > T_f$  we have  $W(t) = 0$ , hence  $C_F(t) = \beta t > \beta T_f = c_0$ . For  $t < T_0$ , an arrival at  $t$  is first in queue and gets served at 0, hence  $C_{F^*}(t) = \alpha(-t) > -\alpha T_0 = c_0$ . Thus,  $C_{F^*}(t)$  is minimal on the interval  $[T_0, T_f]$ , which has  $F^*$ -measure  $\Lambda$ . Thus,  $F^*$  is an equilibrium arrival profile by Definition 2.

We next show that  $F^*$  is the unique equilibrium. Let  $F$  be any equilibrium arrival profile. By Definition 2, there exists a set  $\mathcal{T}$  of  $F$ -measure  $\Lambda$  on which  $C_F(t)$  equals some constant  $c_1$ , while  $C_F(t) \geq c_1$  elsewhere. From Lemma 1, we know that all customers are served by  $T_f$  so that  $F(T_f) = \Lambda$ . Therefore, we can restrict the set  $\mathcal{T}$  to  $(-\infty, T_f]$ . Moreover, as  $C_F(t)$  is continuous by Lemma 1, we can replace  $\mathcal{T}$  with its closure without changing the above properties. To summarize,  $\mathcal{T}$  can be taken to be a closed set which is bounded above by  $T_f$ .

Let  $t_1$  be the maximal point in  $\mathcal{T}$ . As just noted,  $t_1 \leq T_f$ . We claim that  $t_1 = T_f$ . Indeed, if  $t_1 < T_f$ , then an arrival at time  $t_1$  is the last to arrive and thus gets served last at  $T_f$ , so that  $C_F(t_1) > \beta T_f = C_f(T_f)$ , which is a contradiction to  $t_1 \in \mathcal{T}$ . Therefore  $t_1 = T_f$ , implying that  $T_f \in \mathcal{T}$ .

Now, by definition of  $\mathcal{T}$ ,  $T_f \in \mathcal{T}$  implies that  $C_F(t) = C_F(T_f) = \beta T_f$  for every  $t \in \mathcal{T}$ . Note that this cost is identical to the cost computed for  $F^*$  on  $[T_0, T_f]$ . But since (5) holds at any equilibrium, it follows that  $F(t) = F^*(t)$  for  $t \in \mathcal{T} \cap [T_0, T_f]$ . But this implies that  $F(t) = F^*(t)$  for  $t \in [T_0, T_f]$ , since  $F^*$  is strictly increasing on that interval while  $F(t)$  is continuous and cannot increase outside the set  $\mathcal{T}$  (as  $\mathcal{T}$  has  $F$ -measure  $\Lambda$ ). Finally, noting that  $F^*(T_0) = 0$  and  $F^*(T_f) = \Lambda$ ,  $F$  is completely defined and equals  $F^*$ .  $\square$

**Remark 3** Observe that the equilibrium cost  $C_F(t) = \beta T_f = \beta\Lambda/\mu$  is independent of  $\alpha$ . To understand that, note that for the last arriving customer at  $t = T_f$ , the waiting time is zero and total cost is just the lateness cost  $\beta T_f$ , which also has to be the cost at other time instants  $t \in [T_0, T_f]$  at equilibrium.

**Remark 4** The equilibrium queue size increases linearly for  $t \leq 0$  according to  $Q(t) = F^*(t) = \frac{\mu\alpha}{\alpha+\beta}(t - T_0)$ , and decreases linearly for  $t \geq 0$  according to  $Q(t) = F^*(t) - \mu t = \frac{\Lambda\beta}{\alpha+\beta}(T_1 - t)$ . The maximal queue size is obtained at time zero and equals  $Q(0) = \Lambda\frac{\beta}{\alpha+\beta}$ . Interestingly, the latter is independent of the service rate  $\mu$ .

We next evaluate the price of anarchy (PoA) for the single class model. Recall that the

social cost  $J_{\text{soc}}$  is the sum of costs over all customers. For a given arrival profile  $F$ , we obtain by (3) (with the class index dropped),

$$J_{\text{soc}}(F) = \int C_F(t) dF(t) = \int (\alpha W_F(t) + \beta(t + W_F(t))) dF(t). \quad (6)$$

The PoA quantifies the efficiency loss due to selfish decision making by individuals, as the maximum ratio of the social cost at any equilibrium ( $J_{\text{eq}}$ ) to the optimal social cost ( $J_{\text{opt}}$ ). The PoA is then an upper bound on the above ratio for any equilibrium, and equal to this ratio when the equilibrium is unique. Since, by Theorem 1, the equilibrium arrival profile is unique, we simply define PoA as

$$\text{PoA} = \frac{J_{\text{eq}}}{J_{\text{opt}}}.$$

**Proposition 1 (PoA for the single-class model)** *Recall that  $T_f = \Lambda/\mu$ . Then*

$$(i) \quad J_{\text{opt}} = \frac{1}{2}\beta\Lambda T_f.$$

$$(ii) \quad J_{\text{eq}} = \beta\Lambda T_f.$$

$$(iii) \quad \text{Consequently, PoA} = 2.$$

**Proof:** (i) At the socially optimal solution, the arrival instants of all customers are to be selected to minimize the social cost. Since the fluid model is deterministic, the arrival time of every customer can be set to the instant his service is due to start, which eliminates all queueing delay and is therefore optimal. Thus,  $W(t) \equiv 0$ . It is also evident that starving the server before all work is done cannot be optimal, so that the server must work at full rate  $\mu$  from  $t = 0$  to  $T_f = \Lambda/\mu$ . Putting these two observations together implies that the uniform arrival profile  $F(t) = \Lambda t/T_f$  for  $0 \leq t \leq T_f$  is optimal. Therefore, by (6),

$$J_{\text{opt}} = \int \beta t dF(t) = \frac{1}{2}\beta\Lambda T_f.$$

(This expression becomes obvious once we observe that the mean arrival time is  $T_f/2$ .)

(ii) Recall that the cost for each customer at the unique equilibrium profile  $F^*$  is constant and equal to  $\beta T_f$ . Therefore, the social cost is  $\Lambda\beta T_f$ .  $\square$

Thus, for the single class model, the social cost at equilibrium is always one half the optimal cost, for any choice of cost parameters and service rate.

We close this section by pointing out an important extension to the basic cost model.

**Remark 5 (When order of service matters.)** The cost function considered so far includes two components: the delayed service cost and the waiting cost in the queue. In many settings of interest, such as queueing for a better seat, it is not the time at which service is obtained that is important, but rather the number of customers that obtain service before us. Fortunately, this leads to only minor changes in our fluid model. To see this, note that this change corresponds to replacing the cost function  $C(t) = \alpha W(t) + \beta(t + W(t))$  from (3) with

$$\hat{C}(t) = \alpha W(t) + \beta F(t). \quad (7)$$

For this new cost, we can repeat the argument in Lemma 1 to deduce that  $t^* = T_f$  in equilibrium, and therefore  $W(t) = F(t)/\mu - t$ . Thus, the cost (7) equals

$$\hat{C}(t) = \frac{F(t)}{\mu}(\alpha + \hat{\beta}) - \alpha t,$$

where  $\hat{\beta} = \beta\mu$ . Comparing with (5), it is evident that the two cost functions coincide once  $\beta$  is replaced by  $\hat{\beta}$ . Thus, our previous results hold for the modified cost function as well after making this substitution. In particular, the PoA remains 2.

## 4 The Multiclass Problem

We now turn to the multiclass fluid model, where customers can be heterogeneous in terms of their cost parameters. As described in Section 2.3, we divide the customer population into a finite number of classes, each characterized by distinct parameters. In the next section we briefly consider the multiclass model with a continuum of classes.

### 4.1 The Equilibrium Profile

We proceed to identify explicitly the equilibrium arrival profile. To that end, define the cost ratio parameters

$$m_i = \frac{\alpha_i}{\alpha_i + \beta_i}, \quad i = 1, \dots, I.$$

Let us re-order the class indices in increasing order of  $m_i$ , so that  $m_i \leq m_{i+1}$ . We will assume for simplicity that all the cost ratio parameters  $m_i$  are distinct. When this is not the case, one can simply unify customer classes that have identical  $m_i$ 's, and all the results of this section essentially hold.

**Theorem 2** Suppose  $m_1 < m_2 < \dots < m_I$ . Then, the equilibrium profile  $\{F_i\}$  exists, is unique, and specified as follows: Let  $T_0 < T_1 < \dots < T_I$  be an increasing sequence of time instants defined by

$$T_I = \Lambda/\mu, \quad T_{i-1} = T_i - \frac{\Lambda_i}{\mu m_i}, \quad i = 0, 1, \dots, I. \quad (8)$$

Then,  $F_i$  corresponds to a uniform distribution on  $[T_{i-1}, T_i]$  with density  $\mu m_i$ , namely

$$F'_i(t) = \mu m_i \mathbf{1}\{T_{i-1} \leq t < T_i\}. \quad (9)$$

We proceed to prove this result. To begin with, we observe that Lemma 1 and its proof remain unchanged in the multiclass case. Thus, under any equilibrium profile  $\{F_i\}$ , the server operates at its full rate  $\mu$  from time 0 till the last customer is served. Hence all customers are served by time  $T_f = \Lambda/\mu$ . Furthermore, a customer that joins the queue at time  $t$  will leave it at time  $\tau = F(t)/\mu$ . Therefore, the cost function for a class  $i$  arrival at  $t$  is given by

$$\begin{aligned} C_i(t) &= \alpha_i(\tau - t) + \beta_i\tau = (\alpha_i + \beta_i)\tau - \alpha_i t \\ &= (\alpha_i + \beta_i) \frac{F(t)}{\mu} - \alpha_i t. \end{aligned} \quad (10)$$

The next Lemma establishes the relationship between the arrival times of the different classes at equilibrium.

**Lemma 2** Let  $\{F_i\}$  be an equilibrium profile.

(i) If an interval  $(t_1, t_2)$  belongs to the support of  $F_i(t)$ , then

$$F'_i(t) = \mu m_i \quad \text{for } t \in (t_1, t_2).$$

(ii) Let  $i$  and  $j$  be two class indices so that  $m_i < m_j$ . Then all arrivals of class  $i$  occur before those of class  $j$ .

The following lemma is useful for proving Lemma 2.

**Lemma 3** Let  $\{F_i\}$  be an equilibrium profile, and denote  $F = \sum_i F_i$ . Then, there are no gaps in the aggregate arrival profile, i.e.,  $F(t_2) - F(t_1) > 0$  for all  $t_2 > t_1$  such that  $0 < F(t_1) < \Lambda$ .

**Proof:** Suppose, to the contrary, that there are no arrivals on  $(t_1, t_2)$ . By our assumptions on  $t_1$  there are some arrivals both before and after this interval. Since the server operates at full rate over  $(t_1, t_2)$ , it follows that the last customer to enter before  $t_1$  will not get served before  $t_2$ . Therefore, by arriving just before  $t_2$ , this customer will reduce her waiting time while leaving at the same time as before, thereby improving her cost. Thus, this arrival profile cannot be an equilibrium profile.  $\square$

**Proof of Lemma 2:** (i) By the equilibrium definition, it follows that  $C_i(t)$  is constant on  $(t_1, t_2)$ . From Lemma 1 it easily follows that each  $F_i$  is absolutely continuous so it admits a density that we denote by  $F'_i(t)$ .

Noting (10), it follows by differentiation that on that interval,

$$F'_i(t) = \mu \frac{\alpha_i}{\alpha_i + \beta_i} = \mu m_i.$$

(ii) Suppose there are classes  $i$  and  $j$  with  $m_i < m_j$  such that some class  $j$  arrivals arrive in some interval  $(t_1, t_2)$  just before class  $i$  arrivals in some interval  $(t_2, t_3)$  with  $t_1 < t_2 < t_3$ . That there will be non-zero arrivals in each of these two intervals is given by Lemma 3. Let us compare the cost incurred by a class  $j$  arrival on these two intervals. For  $t \in (t_1, t_2)$ ,  $C_j(t)$  is constant (by definition of the equilibrium) and equals  $C_j(t_2)$  (by continuity). Now, from item (i) we know that  $F'(t) = \mu m_i$  on  $(t_2, t_3)$ , hence on that interval,

$$\begin{aligned} C'_j(t) &= \frac{d}{dt} \left( (\alpha_j + \beta_j) \frac{F(t)}{\mu} - \alpha_j t \right) \\ &= (\alpha_j + \beta_j) \frac{F'(t)}{\mu} - \alpha_j = (\alpha_j + \beta_j) m_i - \alpha_j \\ &= (\alpha_j + \beta_j)(m_i - m_j) < 0. \end{aligned}$$

This implies that the cost  $C_j(t)$  is strictly smaller on  $(t_2, t_3)$  than on  $(t_1, t_2)$ , which shows that the latter interval cannot be in the support of  $F_j$  at equilibrium, contrary to our assumption.  $\square$

**Proof of Theorem 2:** To establish Theorem 2, we first show that an equilibrium profile must have the indicated form. From Lemma 2(ii) it follows that the arrivals of the different classes are ordered in increasing order of their  $m_i$  parameters. Now, from Lemma 3 it follows that the arrivals of each class  $i$  are supported on a single interval  $[\tau_i, T_i]$ , and that these intervals are contiguous so that  $\tau_i = T_{i-1}$ . From Lemma 2(i) we see that the arrival profile of each class  $i$  on its interval  $[T_{i-1}, T_i]$  is uniform with rate  $\mu m_i$ . Computing the overall arrival volume on that interval gives  $\mu m_i(T_i - T_{i-1}) = \Lambda_i$ , which implies the recursive relation in (8). Finally,  $T_I = \Lambda/\mu$  follows from Lemma 3, as already indicated.



It is now a simple matter to verify that the indicated arrival profile is indeed an equilibrium profile. Clearly, the cost  $C_i(t)$  is constant on  $[T_{i-1}, T_i]$  by construction. Moreover, arguing as in the proof of Lemma 2, it is readily verified that  $C'_i(t) > 0$  for  $t > T_i$  and  $C'_i(t) < 0$  for  $t < T_{i-1}$ , thereby establishing that the cost  $C_i(t)$  is indeed minimized on the support  $[T_{i-1}, T_i]$  of  $F_i$ .  $\square$

We end this subsection with a few observations regarding the equilibrium profile. The aggregate arrival profile  $F(t) = \sum_i F_i(t)$  can be expressed more explicitly as follows.  $F(t)$  is piecewise linear, with slope  $\mu m_i$  on  $[T_{i-1}, T_i]$ . The times  $T_i$  are given by

$$T_i = \Lambda/\mu - \sum_{j=i+1}^I \frac{\Lambda_j}{\mu m_j}. \quad (11)$$

At these times,

$$F(T_i) = \Lambda - \sum_{j=i+1}^I \Lambda_j = \sum_{j=1}^i \Lambda_j \quad (12)$$

with linear interpolation on  $[T_{i-1}, T_i]$  at slope  $\mu m_i$  (see Figure 1). Note that  $T_0 < 0$  (since  $m_i < 1$ ), so that arrivals start before  $t = 0$  as in the single class case. Further, the aggregate arrival profile is *convex* for  $t \leq T_I$ , meaning that *the arrival rate is increasing in time*, reaching its peak towards the end of the service period. Still, the queue length is strictly decreasing beyond  $t = 0$  (which again follows since  $m_i < 1$ .) Finally, arrivals are ordered in increasing order of  $m_i = \frac{\alpha_i}{\alpha_i + \beta_i}$ , or equivalently in increasing order of  $\frac{\alpha_i}{\beta_i}$  which indicates the relative cost they attribute to waiting over being late.

## 4.2 Price of Anarchy

We now compute the Price of Anarchy (PoA) for the multiclass model. We shall first compute the social cost at equilibrium,  $J_{\text{eq}}$ , followed by the optimal social cost  $J_{\text{opt}}$ .

### 4.2.1 Computing $J_{\text{eq}}$

The social cost is defined as the sum of all costs of all customers, at the given arrival profile. Consider the equilibrium arrival profile that was computed in the previous subsection. Since the equilibrium cost  $C_i(t)$  is the same for all members of each class, say  $C_i$ , we obtain

$$J_{\text{eq}} = \sum_i \Lambda_i C_i. \quad (13)$$

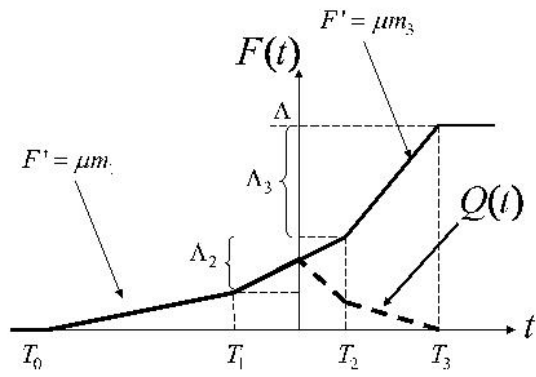


Figure 1: The cumulative distribution of the aggregate arrival profile in equilibrium

The cost  $C_i$  may be computed in any point in  $[T_{i-1}, T_i]$ . Picking  $T_i$ , we get

$$C_i = C_i(T_i) = (\alpha_i + \beta_i) \frac{F(T_i)}{\mu} - \alpha_i T_i.$$

It will be convenient to express this as

$$\mu C_i = \beta_i F(T_i) + \alpha_i (F(T_i) - \mu T_i).$$

Substituting  $T_i$  and  $F(T_i)$  from (11) and (12), we get

$$\mu C_i = \beta_i \sum_{j=1}^i \Lambda_j + \alpha_i \left( \Lambda - \sum_{j=i+1}^n \Lambda_j - \Lambda + \sum_{j=i+1}^n \Lambda_j \frac{\alpha_j + \beta_j}{\alpha_j} \right) \quad (14)$$

$$= \beta_i \sum_{j=1}^i \Lambda_j + \alpha_i \sum_{j=i+1}^n \Lambda_j \frac{\beta_j}{\alpha_j}. \quad (15)$$

**Remark 6** We can observe three distinct components in this expression. The right most sum is the influence of later arrivals (customer classes with  $j > i$ ) on class  $i$ . The influence of customer classes that arrive earlier ( $j < i$ ) is summarized by the preceding term, summed up to  $i - 1$ . The remaining term  $\beta_i \Lambda_i$  expresses the effect of competition within class  $i$  customers.

Substituting the last expression in (13), we obtain

$$J_{\text{eq}} = \sum_i \Lambda_i C_i = \sum_i \frac{\Lambda_i}{\mu} \left( \beta_i \sum_{j=1}^i \Lambda_j + \alpha_i \sum_{j=i+1}^n \Lambda_j \frac{\beta_j}{\alpha_j} \right). \quad (16)$$

A more symmetric form for  $J_{\text{eq}}$  can be obtained. Note that the ratio  $\frac{\beta_i}{\alpha_i}$  is decreasing in  $i$  (since the opposite holds for  $m_i$  by assumption). Therefore,

$$\begin{aligned} J_{\text{eq}} &= \frac{1}{\mu} \sum_{i=1}^I \Lambda_i \Lambda_j \alpha_i \left( \sum_{j=1}^i \frac{\beta_i}{\alpha_i} + \sum_{j=i+1}^I \frac{\beta_j}{\alpha_j} \right) \\ &= \frac{1}{\mu} \sum_{i,j=1}^I \Lambda_i \Lambda_j \alpha_i \min \left\{ \frac{\beta_i}{\alpha_i}, \frac{\beta_j}{\alpha_j} \right\}. \end{aligned} \quad (17)$$

We note that the latter expression is independent of ordering of the classes.

### 4.2.2 The Optimal Social Cost

The optimal social cost  $J_{\text{opt}}$  is obtained by optimizing the arrival times and server allocation for all customers. Here there would be no queues, as each customer can arrive exactly when her turn to be served arrives. It may then be seen through a simple interchange argument that the optimal ordering of arrivals between classes is in decreasing order of  $\beta_i$ . Let  $\sigma(i)$  be an index permutation so that  $\beta_{\sigma(1)} \geq \dots \geq \beta_{\sigma(n)}$ . Then, class  $\sigma(i)$  customers arrive uniformly with rate  $\mu$  between  $\tau_{i-1}$  and  $\tau_i$ , where  $\tau_i = \mu^{-1} \sum_{j=1}^i \Lambda_{\sigma(j)}$ . The overall cost for this class becomes

$$J_i = \Lambda_{\sigma(i)} \beta_{\sigma(i)} \frac{T_{i-1} + T_i}{2} = \frac{\Lambda_{\sigma(i)}}{\mu} \beta_{\sigma(i)} \left( \sum_{j=1}^{i-1} \Lambda_{\sigma(j)} + \frac{1}{2} \Lambda_{\sigma(i)} \right)$$

so that

$$J_{\text{opt}} = \sum_i J_i = \sum_{i=1}^I \frac{\Lambda_{\sigma(i)}}{\mu} \beta_{\sigma(i)} \left( \sum_{j=1}^{i-1} \Lambda_{\sigma(j)} + \frac{1}{2} \Lambda_{\sigma(i)} \right). \quad (18)$$

Again, we can express  $J_{\text{opt}}$  in a more symmetric form. Indeed,

$$J_{\text{opt}} = \frac{1}{2\mu} \sum_{i=1}^I \Lambda_{\sigma(i)} \left( \sum_{j=1}^{i-1} \Lambda_{\sigma(j)} \beta_{\sigma(i)} + \Lambda_{\sigma(i)} \beta_{\sigma(i)} + \sum_{j=i+1}^I \Lambda_{\sigma(j)} \beta_{\sigma(j)} \right) \quad (19)$$

$$= \frac{1}{2\mu} \sum_{i,j=1}^I \Lambda_{\sigma(i)} \Lambda_{\sigma(j)} \min\{\beta_{\sigma(i)}, \beta_{\sigma(j)}\} \quad (20)$$

where the first equality follows by splitting the last sum in (18) into two equal terms and changing order of summation in one of them, and the second equality follows since  $\beta_{\sigma(i)}$  is decreasing in  $i$ . Now, it may be seen that the last expression does not depend on the permutation  $\sigma$ , hence we can remove the permutation and finally obtain

$$J_{\text{opt}} = \frac{1}{2\mu} \sum_{i,j=1}^I \Lambda_i \Lambda_j \min\{\beta_i, \beta_j\}. \quad (21)$$

We note the following simple bounds on  $J_{\text{eq}}$  and  $J_{\text{opt}}$  that readily follow from (17) and (21):

$$J_{\text{eq}} \leq \frac{1}{\mu} \sum_{i,j=1}^I \Lambda_i \Lambda_j \beta_i \leq \frac{1}{\mu} \beta_{\max} \Lambda^2 \quad (22)$$

$$J_{\text{opt}} \geq \frac{1}{2\mu} \beta_{\min} \sum_{i,j=1}^I \Lambda_i \Lambda_j = \frac{1}{2\mu} \beta_{\min} \Lambda^2 \quad (23)$$

where  $\Lambda = \sum_{i=1}^I \Lambda_i$ ,  $\beta_{\min} = \min_i(\beta_i)$ , and  $\beta_{\max} = \max_i(\beta_i)$ . We next proceed to derive additional bounds on the *ratio* of  $J_{\text{eq}}$  and  $J_{\text{opt}}$ .

### 4.2.3 PoA Bounds

The equations derived above imply the following explicit expression for the PoA:

$$\text{PoA} \triangleq \frac{J_{\text{eq}}}{J_{\text{opt}}} = 2 \frac{\sum_{i,j=1}^I \Lambda_i \Lambda_j \alpha_i \min\{\frac{\beta_i}{\alpha_i}, \frac{\beta_j}{\alpha_j}\}}{\sum_{i,j=1}^I \Lambda_i \Lambda_j \min\{\beta_i, \beta_j\}}. \quad (24)$$

As we will see below, the PoA ranges around its single-class value of 2. We proceed to derive some bounds on this value. Essentially, we will be interested in bounds that depend only on the ranges of the cost parameters ( $\alpha_i$  and  $\beta_i$ ) but not on the relative size ( $\Lambda_i$ ) of the customer classes. We start with some special cases, where only one parameter varies across classes.

#### Proposition 2

(i) Identical wait sensitivities. *Suppose  $\alpha_i \equiv \alpha_0$ : the wait sensitivities are identical for all customer classes. Then*

$$\text{PoA} = 2.$$

(ii) Identical lateness sensitivities. *Suppose  $\beta_i \equiv \beta_0$ : the lateness sensitivities are identical for all classes. Then*

$$\text{PoA} \leq 2, \quad (25)$$

and

$$\text{PoA} \geq 2 - (1 - I^{-1}) \left( 1 - \frac{\alpha_{\min}}{\alpha_{\max}} \right) \geq 1 + \frac{\alpha_{\min}}{\alpha_{\max}}, \quad (26)$$

where  $\alpha_{\max} = \max_i \alpha_i$ ,  $\alpha_{\min} = \min_i \alpha_i$ , and  $I$  is the number of classes.

**Proof:** Item (i) is immediate from (24). As for (ii), from the same equation we obtain the upper bound

$$\text{PoA} = 2 \frac{\sum_{i,j=1}^I \Lambda_i \Lambda_j \min\{1, \frac{\alpha_i}{\alpha_j}\}}{\sum_{i,j=1}^I \Lambda_i \Lambda_j} \leq 2. \quad (27)$$

On the other hand, proceeding from the last expression and recalling that  $\alpha_i \leq \alpha_j$  for  $i < j$ ,

$$\text{PoA} = 2 - \frac{2 \sum_{i,j=1}^I \Lambda_i \Lambda_j (1 - \min\{1, \frac{\alpha_i}{\alpha_j}\})}{\sum_{i,j=1}^I \Lambda_i \Lambda_j} \quad (28)$$

$$= 2 - \frac{2 \sum_{i < j} \Lambda_i \Lambda_j (1 - \frac{\alpha_i}{\alpha_j})}{\sum_{i,j=1}^I \Lambda_i \Lambda_j} \quad (29)$$

$$\geq 2 - \frac{\max_{i,j} (1 - \frac{\alpha_i}{\alpha_j}) \cdot 2 \sum_{i < j} \Lambda_i \Lambda_j}{\sum_{i,j=1}^I \Lambda_i \Lambda_j} \quad (30)$$

$$= 2 - \left(1 - \frac{\alpha_{\min}}{\alpha_{\max}}\right) \frac{\sum_{i \neq j} \Lambda_i \Lambda_j}{(\sum_i \Lambda_i)^2}. \quad (31)$$

It is easily verified that the last fraction is maximized when all the  $\Lambda_i$ 's are equal, and in that case it equals  $(I - 1)/I$ . Hence follows the lower bound in (26).  $\square$

Item (i) of the last proposition is evidently an exact extension of the PoA result for the single-class case, giving the same value of 2. Regarding (ii), we first note the upper bound of 2 is strict unless all the  $\alpha_i$ 's are equal as well. Thus, in this case, diversity in the waiting sensitivities of the customers actually improves the PoA compared to the single class case. As for the lower bound, for two user classes ( $I = 2$ ) with  $\alpha_1 < \alpha_2$  it reads

$$\text{PoA} \geq 1.5 + 0.5 \frac{\alpha_1}{\alpha_2}.$$

We observe that this bound is tight, and is achieved when  $\Lambda_1 = \Lambda_2$ .

We now turn to consider the general case, when both sets of cost parameters may vary across customer classes. The following set of bounds is obtained simply by bounding separately the ratios of each pair of corresponding terms in the numerator and denominator of (24).

**Proposition 3** *Let  $H_{\max} = \max_{i,j} H(i, j)$  and  $H_{\min} = \min_{i,j} H(i, j)$ , where*

$$H(i, j) = \frac{(\alpha_i + \alpha_j) \min\{\frac{\beta_i}{\alpha_i}, \frac{\beta_j}{\alpha_j}\}}{2 \min\{\beta_i, \beta_j\}}.$$

*Then*

$$2H_{\min} \leq \text{PoA} \leq 2H_{\max}. \quad (32)$$

Consequently,

$$\text{PoA} \leq 1 + \frac{\alpha_{\max}}{\alpha_{\min}}, \quad (33)$$

$$\text{PoA} \leq 1 + \frac{\beta_{\max}}{\beta_{\min}}, \quad (34)$$

$$\text{PoA} \geq (1 + \frac{\alpha_{\min}}{\alpha_{\max}}) \frac{\beta_{\min}}{\beta_{\max}}. \quad (35)$$

**Proof:** The bounds in (32) follow immediately after noting that, by collecting terms, (24) may be written as:

$$\text{PoA} = 2 \frac{\sum_i \beta_i \Lambda_i^2 + \sum_{i < j} \Lambda_i \Lambda_j (\alpha_i + \alpha_j) \min\{\frac{\beta_i}{\alpha_i}, \frac{\beta_j}{\alpha_j}\}}{\sum_i \beta_i \Lambda_i^2 + \sum_{i < j} \Lambda_i \Lambda_j 2 \min\{\beta_i, \beta_j\}}$$

so that  $H(i, j)$  is the ratio of the coefficients of the  $i, j$  terms. The remaining bounds follow by appropriately bounding  $H(i, j)$ . As for (33), assuming that  $\beta_i \leq \beta_j$  we get

$$2H(i, j) = \frac{(\alpha_i + \alpha_j) \min\{\frac{\beta_i}{\alpha_i}, \frac{\beta_j}{\alpha_j}\}}{\beta_i} \leq \frac{(\alpha_i + \alpha_j) \frac{\beta_i}{\alpha_i}}{\beta_i} = 1 + \frac{\alpha_j}{\alpha_i} \leq 1 + \frac{\alpha_{\max}}{\alpha_{\min}}$$

(the case  $\beta_i > \beta_j$  is symmetric since  $H(i, j) = H(j, i)$ ). As for (34), supposing that  $\frac{\beta_i}{\alpha_i} \leq \frac{\beta_j}{\alpha_j}$ , we get

$$2H(i, j) = \frac{\alpha_i \frac{\beta_i}{\alpha_i} + \min\{\frac{\alpha_j \beta_i}{\alpha_i}, \beta_j\}}{\min\{\beta_i, \beta_j\}} \leq \frac{\beta_i + \beta_j}{\min\{\beta_i, \beta_j\}} \leq 1 + \frac{\beta_{\max}}{\beta_{\min}}.$$

Finally, to establish (35), consider again that  $\beta_i \leq \beta_j$  so that

$$2H(i, j) = (\alpha_i + \alpha_j) \min\{\frac{1}{\alpha_i}, \frac{1}{\alpha_j} \frac{\beta_j}{\beta_i}\} = \min\{1 + \frac{\alpha_j}{\alpha_i}, (1 + \frac{\alpha_i}{\alpha_j}) \frac{\beta_j}{\beta_i}\} \geq (1 + \frac{\alpha_{\min}}{\alpha_{\max}}) \frac{\beta_{\min}}{\beta_{\max}}.$$

□

Equation (34) provides an upper bound on the PoA in terms of the  $\beta$  parameters only. In fact, a tighter bound of this form may be derived through somewhat refined analysis. This bound also points to the “worst case” conditions in terms of the PoA when the  $(\beta_i)$  parameters are given.

**Proposition 4**  $\text{PoA} \leq 1 + \sqrt{\frac{\beta_{\max}}{\beta_{\min}}}.$

Proof of Proposition is somewhat lengthy and is given in the appendix.

We note that the bound of the last proposition is tight, in the sense that for any set of  $\beta_i$ 's, the bound is satisfied with equality for some  $(\alpha_i, \Lambda_i)$  parameters. Indeed, as implied by the proof, setting the  $\beta_i$ 's in increasing order, equality is obtained for  $\Lambda_2 = \dots = \Lambda_{I-1} = 0$ ,  $\Lambda_1/\Lambda_I = \sqrt{\beta_I/\beta_1}$ , and  $\alpha_I/\alpha_1 = \beta_I/\beta_1$  (cf. (24)).

## 5 The Continuous Parameter Model

We next consider our model with a continuous set of customer classes, rather than discrete. It may be argued that this model is more realistic, which comes at the expense of larger computational (and possibly technical) difficulty. Our treatment here will be brief and informal, and we will essentially rely on the discrete-parameter results to infer the form of the equilibrium arrival profile in the present case.

Let  $q \in I$  denote here the *continuous* class parameter. We can identify  $q$  with the two cost parameters  $(\alpha_q, \beta_q) \in \mathfrak{R}_+^2$ . Let  $g_1(q) \geq 0$  be a density function on  $I$ , with total mass  $\int g_1(q) dq = \Lambda$ . Thus,  $g_1(q)$  denotes the density of arrivals of class  $q$ . We assume that there are no point masses in the cost parameter distribution, so that  $g_1$  is finite.

Let  $m_q = \alpha_q / (\alpha_q + \beta_q) \in [0, 1]$  denote the cost ratio parameter for class  $q$  customers. Since the equilibrium arrival profile is completely characterized by this parameter, it will be useful to define its density. Thus, let  $g(m) \geq 0$  denote a density function on  $[0, 1]$ , which is obtained from  $g_1$  as

$$g(m) = \int \mathbf{1}\{m_q = m\} g_1(q) dq.$$

We assume that  $g(m)$  is finite as well. Obviously,  $\int g(m) dm = \Lambda$ . Further, let

$$G(m) = \int_0^m g(\eta) d\eta$$

denote the (absolutely continuous) cumulative distribution function of  $g$ . For simplicity, we will assume that  $g$  has finite support (i.e.,  $m$  is bounded).

As in the discrete parameter case, let  $F(t)$  describe the aggregate arrival profile of the customer population. The equilibrium arrival profile is defined as before. Looking at the continuous model as the limit of the discrete one, with the number of classes going to infinity, we may infer the following analogous properties of the equilibrium arrival profile (see Lemmas 2 and 3, Theorem 2 and Figure 1).

1. The server operates at full rate  $\mu$  till the last customer is served. Thus, the last customer is served at  $T_f = \Lambda/\mu$ .
2. Arrivals occur in increasing order of  $m$ . That is, customers of class  $q_1$  arrive before those of class  $q_2$  if  $m_{q_1} < m_{q_2}$ .
3. If arrivals at time  $t$  have cost ratio parameter  $m(t)$ , then

$$F'(t) = \mu m(t). \tag{36}$$



It follows that all customers with  $m \leq m(t)$  arrive up to time  $t$ , hence

$$F(t) = G(m(t)).$$

We proceed to derive differential equations for  $m(t)$  and  $F(t)$ . Differentiating the last equation gives

$$F'(t) = g(m(t))m'(t)$$

and together with (36) we get

$$m'(t) = \mu \frac{m(t)}{g(m(t))} \quad , \quad t \leq T_f. \quad (37)$$

The boundary condition for this equation is obtained by noting that the last arrivals occur at  $T_f = \Lambda/\mu$  and have maximal  $m$ . Thus, letting  $m_{\max}$  denote the maximal point in the support of  $g(m)$ ,

$$m(T_f) = m_{\max}.$$

$m(t)$  may now be computed from the differential equation with a boundary condition. The equilibrium arrival profile  $F(t)$  may then be computed using  $F(t) = G(m(t))$ .

We note that a direct equation for  $F(t)$  follows by combining (36) with (37), yielding

$$F''(t) = \frac{\mu F'(t)}{g(\mu^{-1}F'(t))}$$

with terminal conditions  $F'(T_f) = \mu m_{\max}$  and  $F(T_f) = \Lambda$ . It is clearly seen that  $F''(t) \geq 0$  over  $t \leq T_f$ , hence  $F(t)$  is convex there.

It is easy to verify that the arrival profile thus defined is indeed an equilibrium profile. Recall that the cost function for a class  $q$  arrival is given by (see equation (10)):

$$C_q(t) = (\alpha_q + \beta_q) \frac{F(t)}{\mu} - \alpha_q t = (\alpha_q + \beta_q) \left( \frac{F(t)}{\mu} - m_q t \right).$$

It further follows by construction and (36) that customers with parameter  $m_q$  arrive at time  $t_q$  defined by  $F'(t_q) = \mu m_q$ . We will show that  $t_q$  minimized  $C_q$ . Differentiating, we get

$$C'_q(t) = (\alpha_q + \beta_q) \left( \frac{F'(t)}{\mu} - m_q \right).$$

Therefore,  $C'_q(t) = 0$  at  $t = t_q$ . Furthermore,

$$C''_q(t) = (\alpha_q + \beta_q) \frac{F''(t)}{\mu}.$$

But as observed above  $F(t)$  is convex on  $t \leq T_f$  and hence so is  $C_i$ . Thus,  $t_q$  is a minimizer there. It is also clear that the cost  $C_q$  is increasing for  $t$  beyond  $T_f$ , hence  $t_q$  is a global minimizer of  $C_i$ .

We turn to an example that illustrates the required computations in the simple case of uniformly distributed cost parameters.

**Example:** Let the cost ratio parameter  $m$  of the customer population be uniformly distributed on some interval  $[m_0, m_1]$ , namely

$$g(m) = g_0 \mathbf{1}\{m_0 \leq m \leq m_1\}, \quad g_0 = \frac{\Lambda}{m_1 - m_0}.$$

Then, by (37),

$$m'(t) = \frac{\mu}{g_0} m(t), \quad t \leq T_f; \quad m(T_f) = m_1,$$

with the solution

$$m(t) = m_1 e^{\frac{\mu}{g_0}(t-T_f)}, \quad T_0 \leq t \leq T_f.$$

Here  $T_0$  must satisfy  $m(T_0) = m_0$ , so that  $T_0 = T_f - \frac{g_0}{\mu} \log(\frac{m_1}{m_0})$ . The equilibrium arrival density  $F'$  is given by

$$F'(t) = g(m(t))m'(t) = g_0 m'(t) = \mu m_1 e^{\frac{\mu}{g_0}(t-T_f)}, \quad T_0 \leq t \leq T_f.$$

Evidently, the arrival distribution at equilibrium turns out to be an exponentially increasing function. Finally, the cumulative arrival distribution  $F(t)$  may be obtained by integrating  $F'$  and using  $F(T_f) = \Lambda$ , yielding

$$F(t) = g_0 m_1 (e^{\frac{\mu}{g_0}(t-T_f)} - 1) + \Lambda, \quad T_0 \leq t \leq T_f.$$

□

To close this section, we observe that the PoA bounds from Section 4.2, which depend only on the range of the cost parameters  $\alpha$  and  $\beta$ , should hold without modification in the present continuous-parameter model as well.

## 6 Reducing the Price of Anarchy

In this section, we discuss some ways in which price of anarchy may be reduced. For simplicity, we consider the single class customer setting, although the generalization to multi-class is straightforward. A key message of this section is that the fluid model is

sufficiently tractable to provide elegant and intuitive answers to many natural methods for reducing PoA.

We first consider the case where PoA may be reduced by segmenting the population so that certain proportions are served only after specified thresholds. This has applications in numerous settings. For instance, when large number of candidates are to be interviewed by an organization, often they are segmented and are asked to report at different time segments. Visa and immigration centers often through online scheduling assign separate time windows to customers coming in for an appointment. We also discuss performance degradation that may occur if sub-optimal choices are made. This is of obvious practical importance as it is difficult to make optimal decisions given inherent model uncertainty.

We then briefly discuss how improved PoA may be obtained by assigning varying priorities to different segments of population. One popular application that closely approximates this is in airplane boarding where economy customers are assigned different priority based on their seat location. Finally, we discuss how better price of anarchy may be obtained through differential pricing by charging a tariff to customers that are served early. When done optimally, this can be quite effective in controlling PoA. We also discuss performance degradation that may occur with suboptimal pricing.

In the above three settings, it is easy to quantify the reduction in price of anarchy as a function of the population segments created. In particular, we see that when done optimally, the price of anarchy equals  $1 + 1/n$  where  $n$  denotes the number of population segments. Without loss of generality we take  $\Lambda = 1$  in this section.

## 6.1 Service Delayed for Some Customers

To convey the main points simply, we focus primarily on the case where the population is divided into two segments. Specifically, consider the case where  $(1 - a)$  proportion of the population is allowed to be served only after time  $\hat{\tau} > 0$ . Call this the second population. The first population corresponds to the proportion  $a$  that is allowed to be served at any time  $t \geq 0$ . We allow the second population to queue up before time  $\hat{\tau}$ , so that after time  $\hat{\tau}$  they join the end of the queue of population 1 customers at the service facility (if any) and are served after them. Within the same population the service is always in the order of customer arrival. After time  $\hat{\tau}$ , customers from both the populations join at the end of the existing queue at the service facility and are served in the order they arrive.

We first consider the case  $\hat{\tau} = \frac{a}{\mu}$ . This is a critical point as population 1 completes its

service requirements at time  $\frac{a}{\mu}$ . We do not discuss the case  $\hat{\tau} > \frac{a}{\mu}$  separately as in this case the queue is empty in the interval  $\hat{\tau} - \frac{a}{\mu}$  and the analysis is trivial. The case  $\hat{\tau} < \frac{a}{\mu}$  is analyzed in some detail as it provides interesting insights on how variedly customers may behave as  $\hat{\tau}$  decreases. The key observation is that there is a certain *phase* change in customer behavior at  $\hat{\tau} = \hat{\tau}^*$ , where

$$\hat{\tau}^* = \frac{a}{\mu} - (1 - a) \frac{\beta}{\alpha\mu}. \quad (38)$$

While for  $\hat{\tau}^* < \hat{\tau} < \frac{a}{\mu}$ , there exists a unique equilibrium solution and the aggregate arrival profile varies with  $\hat{\tau}$ , for  $\hat{\tau} < \hat{\tau}^*$  there may be multiple equilibria, the aggregate arrival profile is independent of  $\hat{\tau}$  and is identical to the unconstrained case.

### 6.1.1 $\hat{\tau} = \frac{a}{\mu}$

Under this scheme, the unique equilibrium is easily seen to correspond to both the populations blissfully unaware of the other, the first population arrives as if the second does not exist and the server facility opens at time 0, the second population arrives as if the server facility opens at time  $a/\mu$  and queues up appropriately before time  $a/\mu$ . Specifically, the first population has  $a$  proportion of customers that arrive uniformly between the interval  $[-\frac{\beta a}{\alpha\mu}, \frac{a}{\mu}]$  and the second population has  $(1 - a)$  proportion of customers that arrive uniformly between  $[\hat{\tau}^*, \frac{1}{\mu}]$ . Both arrive at rate  $\frac{\alpha\mu}{(\alpha+\beta)}$  in their respective arrival intervals so that the cost incurred by arrivals in each population is constant independent of the arrival times. See Figure 2 for an illustration.

Then, a customer from the first population has no incentive to arrive outside the interval  $[-\frac{\beta a}{\alpha\mu}, \frac{a}{\mu}]$  where the cost would be higher. Similarly, the customer in second population arriving at any time in the interval  $[\hat{\tau}^*, \frac{1}{\mu}]$  has a constant cost and a higher cost outside this interval.

The cost incurred by a customer in the first population equals  $\beta a/\mu$ , and that by a customer in second population equals  $\beta/\mu$ .

The overall cost, since population 1 has proportion  $a$  and population 2 has proportion  $1 - a$ , equals

$$\frac{\beta}{\mu}(a^2 + (1 - a)).$$

The social optimal corresponds to zero waiting and the associated overall cost equals  $\frac{\beta}{2\mu}$ .

The PoA then equals

$$2(a^2 + (1 - a)).$$

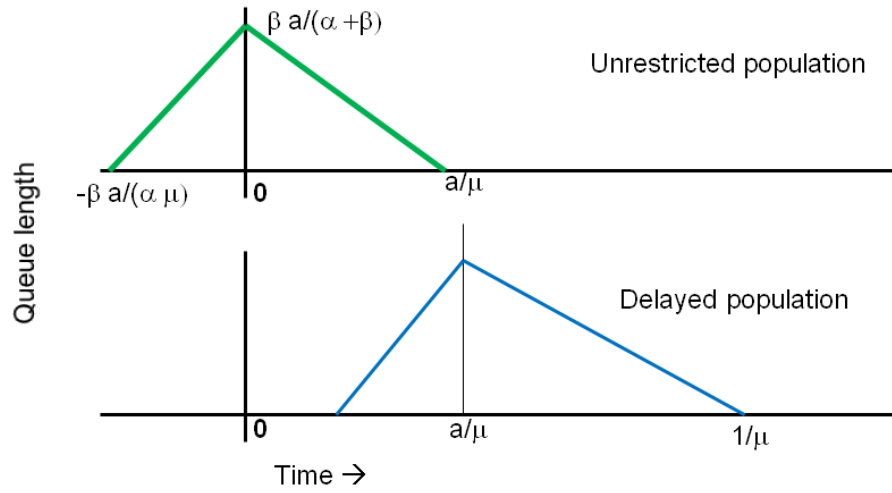


Figure 2: Equilibrium queue length profile for the two populations. Population 1 comprises  $a$  proportion and is served in the interval  $[0, a/\mu]$ . Population 2 comprises  $(1 - a)$  proportion and is allowed service after time  $a/\mu$ , although it starts queueing from time  $\hat{\tau}^*$  onwards.

This is minimized at  $a = 1/2$  where the PoA equals  $3/2$ .

Hence, to achieve optimal PoA, when the population is segmented into two parts, it is best to schedule half the population to come half the total serving time later.

### 6.1.2 $\hat{\tau} < \frac{a}{\mu}$

The following proposition summarizes this setting for different values of  $\hat{\tau}$ . Let  $m$  denote the ratio  $\frac{\alpha}{(\alpha+\beta)}$ .

#### Proposition 5

1. For  $\hat{\tau}^* < \hat{\tau} < \frac{a}{\mu}$ , there exists a unique equilibrium where the first population arrives uniformly over

$$[-\frac{\beta a}{\alpha \mu} - (\frac{a}{\mu} - \hat{\tau}), \hat{\tau}]$$

at rate  $\mu m$  and the second population arrives uniformly over  $[\hat{\tau}^*, \frac{1}{\mu}]$  at rate  $\mu m$ . The PoA increases linearly from  $2(a^2 + (1-a))$  to 2 as  $\hat{\tau}$  decreases from  $\frac{a}{\mu}$  to  $\hat{\tau}^*$ .

2. For  $\hat{\tau} \leq \hat{\tau}^*$ , in equilibrium, population 1 arrives uniformly over

$$[-\frac{\beta}{\alpha \mu}, \hat{\tau}^*]$$

at rate  $\mu m$  and population 2 arrives uniformly over  $[\hat{\tau}^*, \frac{1}{\mu}]$  at rate  $\mu m$ . Here, multiple equilibria may exist, and in each equilibrium, each customer in either class incurs a cost  $\frac{\beta}{\mu}$ . Hence, PoA equals 2.

**Proof:** Let

$$c_1 = \alpha(\frac{a}{\mu} - \hat{\tau}) + \beta \frac{a}{\mu}$$

and  $c_2 = \beta/\mu$ .

First consider  $\hat{\tau}^* < \hat{\tau} < \frac{a}{\mu}$ . We first argue that the specified arrival profile of the two populations is in equilibrium. To see this note that each customer in population 1 incurs a constant cost equal to that incurred by the last customer of this population arriving at time  $\hat{\tau}$  and served at time  $\frac{a}{\mu}$ , i.e.,  $c_1$ .

The cost incurred by each customer in population 2 equals  $c_2 > c_1$ . Therefore, customers in population 1 have no incentive to come after time  $\hat{\tau}$ . They clearly have no incentive to

come before  $-\frac{\beta a}{\alpha \mu} - (\frac{a}{\mu} - \hat{\tau})$ . Similarly, population 2 has no incentive to come outside the specified intervals.

To see that this equilibrium is unique, note that as before, under any equilibrium, the server will serve at a full rate till time  $1/\mu$ . Clearly, the last customer to be served in equilibrium will arrive at time  $1/\mu$  and incur the cost  $c_2$ . She cannot be from population 1, as then she has the option of arriving at time  $\hat{\tau}$  and be served by at most time  $a/\mu$ . That is, her cost in equilibrium must be bounded from above by  $c_1 < c_2$ .

Hence, population 2 customer is the last one to arrive and in equilibrium the cost incurred by each customer of population 2 equals  $c_2$ . Clearly, population 1 cannot arrive after time  $\hat{\tau}$  and incur cost less than  $c_2$  as then a customer from population 2 can replicate this to lower her cost. Hence, since each customer in population 1 has a constant cost, this population arrives uniformly in an interval  $[\tau - \frac{a(\alpha+\beta)}{\mu\alpha}, \tau]$  at rate  $\mu m$  for some  $\tau \leq \hat{\tau}$ . Again, if  $\tau < \hat{\tau}$ , the last customer of this population can improve her cost by coming at  $\hat{\tau}$ , so  $\tau = \hat{\tau}$ . In particular, cost incurred by population 1 customer equals  $c_1$ , and they are served uninterruptedly till time  $a/\mu$ . From population 2's viewpoint, then, in equilibrium the queue opens at time  $a/\mu$ , and hence in equilibrium it must follow the profile specified in the proposition.

It is easily seen that PoA increases linearly from  $2(a^2 + (1-a))$  to 2 as  $\hat{\tau}$  decreases from  $\frac{a}{\mu}$  to  $\hat{\tau}^*$ .

Now consider the case  $\hat{\tau} \leq \hat{\tau}^*$ . First note that under the strategy specified in this proposition, each customer from both the populations incurs cost  $c_2$ , and cannot improve this by arriving at another time.

To see that all equilibria must have cost  $c_2$ , first note that in equilibrium the cost incurred by either population cannot be more than  $c_2$ , the cost incurred by the last customer arriving at  $1/\mu$ .

Now suppose that  $\hat{\tau} < \tau^*$  so that  $c_1 > c_2$ . If all of population 1 arrives by time  $\hat{\tau}$ , the cost incurred by its last customer (who will be served at time  $a/\mu$  and will need to wait at least  $a/\mu - \hat{\tau}$ ) is greater than  $c_1$ . Hence, this cannot be in equilibrium and some customers must arrive after time  $\hat{\tau}$ . These must have the same cost as population 2 customers in equilibrium. Thus, equilibrium cost for each customer must equal  $c_2$ .

Finally, consider  $\hat{\tau} = \tau^*$  so that  $c_1 = c_2$ . Then, even if all customers of population 1 arrive by  $\hat{\tau}$ , their cost must be constant and equal to that of their last customer that has to arrive

at  $\tau^*$  in equilibrium. So, there cost must equal  $c_2$ . Therefore, in all equilibria, the cost incurred by each customer equals  $c_2$ . Hence, PoA equals 2. □

**Remark 7** Note that in the latter case multiple equilibria with same cost exist in the sense that any strategy where  $\frac{\hat{\tau} + \frac{\beta}{\alpha\mu}}{\hat{\tau}^* + \frac{\beta}{\alpha\mu}}$  arrive uniformly at rate  $\mu m$  in the interval

$$[-\frac{\beta}{\alpha\mu}, \hat{\tau}]$$

and the remaining population 1 and population 2 customers arrive uniformly at rate  $\mu m$  in the interval  $[\hat{\tau}, \frac{1}{\mu}]$  is also an equilibrium strategy as the cost incurred by each customer in either population is  $\frac{\beta}{\mu}$ .

### 6.1.3 Generalization to multiple thresholds

Note that when we set  $\hat{\tau}$  optimally at  $\frac{1}{2\mu}$  for  $a = 1/2$ , we obtained the optimal PoA of  $3/2$ . This generalizes so that if we restrict  $\frac{m}{n}$  proportion of people to come after time  $\frac{n-m}{n\mu}$  time for  $m = 1, 2, \dots, n-1$  then the equilibrium cost for customers getting served in a slot  $(\frac{m}{n\mu}, \frac{m+1}{n\mu})$  equals

$$\frac{(m+1)\beta}{n\mu}.$$

Again, this is  $1/n$  proportion of the population so that the total cost equals

$$\frac{\beta}{n\mu}(\frac{1}{n} + \frac{2}{n} + \dots + 1)$$

or  $\frac{\beta(n+1)}{2n\mu}$  so that the PoA equals  $(n+1)/n$  and converges to 1 as  $n \rightarrow 1$ .

## 6.2 Priority Queueing

Another way to achieve PoA equal to  $\frac{n+1}{n}$  as in Subsection 6.1.3 is through dividing the population into  $n$  separate segments and assigning different priorities to them. Specifically, suppose that the population is divided into  $n$  segments with  $(a_i : i \leq n)$  denoting the respective proportions (the cost function is identical for each segment). The population segment with lower index is given priority over the segment with higher index. Then, in equilibrium customers arrive in disjoint intervals, customers of segment 1 arrive first uniformly in the interval  $[-\frac{\beta a_1}{\alpha\mu}, \frac{a_1}{\mu}]$  and are served by the server in the interval  $[0, \frac{a_1}{\mu}]$ . Similarly,



customers of segment  $j \geq 2$  arrive uniformly in the interval  $[\sum_{i=1}^{j-1} \frac{a_j}{\mu} - \frac{\beta a_i}{\alpha \mu}, \sum_{i=1}^j \frac{a_i}{\mu}]$  and are served in the interval  $[\sum_{i=1}^{j-1} \frac{a_i}{\mu}, \sum_{i=1}^j \frac{a_i}{\mu}]$ .

The cost incurred by segment  $i$  equals  $\beta \sum_{i=1}^j \frac{a_i}{\mu}$  so that overall price of anarchy equals

$$2 \left[ \sum_{j=1}^n a_j \left( \sum_{i=1}^j a_i \right) \right].$$

Through simple optimization, it can be seen that this is minimized by setting  $a_j = \frac{1}{n}$  for each  $j$  so that PoA equals  $\frac{n+1}{n}$  as in Subsection 6.1.3.

### 6.3 Reduction in PoA through Charging Tariffs

Recall that in Section 6.1 in the two population setting, we obtained the best PoA when we divided the populations in equal parts and allowed the second population to come after time  $\frac{1}{2\mu}$ . Then, the cost to each customer in the first population was  $\frac{1}{2\mu}$  less than that of customers in the second population. This suggests a procedure for implementing discriminatory pricing.

For brevity, we restrict our discussion to the case where customers joining the service facility queue by time  $\frac{1}{2\mu}$  have to pay a constant tariff  $p$  while the customers joining the service facility queue after this time pay no tariff. We refer to the former as population 1 and latter as population 2. We assume here that demand of one unit is fixed and is unaffected by the pricing strategy of the service provider. Again, we allow population 2 to queue up before time  $\frac{1}{2\mu}$  separately and join at the end of service facility queue at time  $\frac{1}{2\mu}$ . In this case, they are served after population 1 customers at the service facility queue at that time, if any, and in their order of arrival amongst population 2. We further assume that the tariff collected is returned to the society so this does not enter into the price of anarchy calculations. We now discuss different scenarios depending upon the value of  $p$ . The proofs of Propositions 6 and 7 are not central to our analysis and are given in the appendix.

#### 6.3.1 $p = \frac{\beta}{2\mu}$

In this scenario, the first population arrives uniformly between  $[-\frac{\beta}{2\alpha\mu}, \frac{1}{2\mu}]$  at rate  $\mu m$ , and the other between  $[\frac{1}{2\mu} - \frac{\beta}{2\alpha\mu}, \frac{1}{\mu}]$  at the same rate. The cost incurred by both the populations is  $\frac{\beta}{\mu}$ : For the first population it is  $\frac{\beta}{2\mu}$  from waiting and time to service and another  $\frac{\beta}{2\mu}$  from the tariff for coming early.

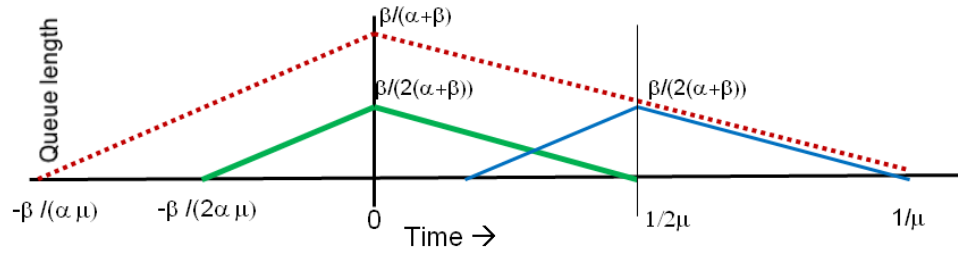


Figure 3: The dotted line denotes the queue profile before differential pricing. After differential pricing the darkened line denotes the queue profile of population 1 that pays  $\beta/(2\mu)$  more than population 2 whose queue profile is shown using the lighter line. The cost to customer joining either of the two populations equals  $\beta/\mu$ .

Thus, a customer is indifferent to coming as part of population 1 or 2. The revenue collected by the service provider from tariffs equals  $\frac{\beta}{4\mu}$ . The PoA, as before, equals 3/2. See Figure 3 for an illustration of this scenario.

### 6.3.2 $p = (1 + c)\frac{\beta}{2\mu}$ , $c > 0$

In this scenario, the following proposition summarizes the results for different values of  $c$ .

#### Proposition 6

1. In unique equilibrium,  $(\frac{1}{2} - \frac{c}{4})$  proportion of customers arrive as population 1, for  $c \leq 2$ , at rate  $\mu m$ , uniformly over

$$\left[ -\frac{\beta}{\alpha\mu}(\frac{1}{2} - \frac{c}{4}), \frac{1}{\mu}(\frac{1}{2} - \frac{c}{4}) \right],$$

and  $(\frac{1}{2} + \frac{c}{4})$  proportion arrive as population 2 at rate  $\mu m$  uniformly over

$$\left[ \frac{1}{2\mu} - \frac{\beta}{2\alpha\mu}(1 + \frac{c}{2}), \frac{1}{\mu} + \frac{c}{4\mu} \right].$$

For  $c \geq 2$ , all customers come as population 2 as for  $c = 2$ .

2. Furthermore, for  $c \leq 2$ , PoA equals

$$= \frac{3}{2} + \frac{c(1+c)}{4}. \quad (39)$$

For  $c > 2$  it equals 3.

See Figure 4 for an illustration.

### 6.3.3 $p = (1 - c)\frac{\beta}{2\mu}$ , $0 < c < 1$

The following proposition summarizes the results for different values of  $c$  with the above price  $p$ .

#### Proposition 7

1. For  $0 \leq c \leq 1$ , in unique equilibrium, proportion  $\frac{1}{2} + \frac{\beta c}{2(\alpha + \beta)}$  of customers arrive as population 1 at rate  $\mu m$ , uniformly over

$$\left[ -\frac{\beta}{2\alpha\mu}(1 + c), \frac{1}{2\mu} \right].$$

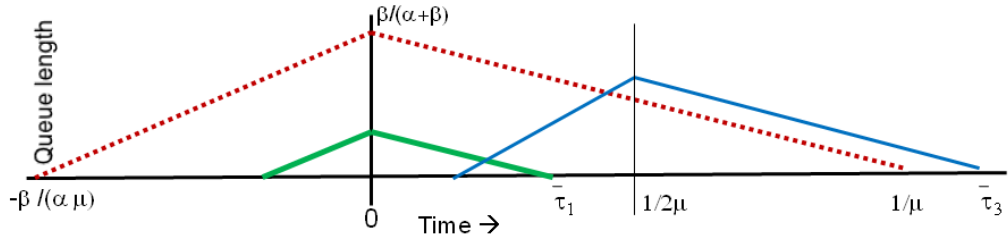


Figure 4: The dotted line denotes the queue profile before differential pricing. The darkened line denotes the queue profile of population 1 that pays  $\beta(1+c)/(2\mu)$  more than population 2 whose queue profile is shown using the lighter line. The population 1 is served till  $\tilde{\tau}_1 = \frac{1}{2\mu} - \frac{c}{4\mu}$  and population 2 is served till  $\tilde{\tau}_3 = \frac{1}{\mu} + \frac{c}{4\mu}$ . The cost to customer joining either of the two populations equals  $\frac{\beta}{\mu}(1 + \frac{c}{4})$ .

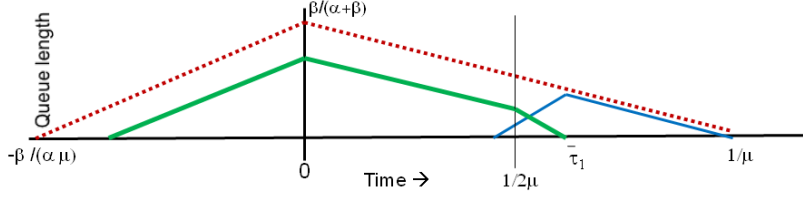


Figure 5: The dotted line denotes the queue profile before differential pricing. The darkened line denotes the queue profile of population 1 that pays  $\beta(1-c)/(2\mu)$  more than population 2 whose queue profile is shown using the lighter line. The population 1 is served till  $\check{\tau}_1 = \frac{1}{2\mu} + \frac{\beta c}{2\mu(\alpha+\beta)}$ . The cost to customer joining either of the two populations equals  $\frac{\beta}{\mu}$ .

In addition, proportion  $\frac{1}{2} - \frac{\beta c}{2(\alpha+\beta)}$  of customers arrive as population 2 at rate  $\mu m$ , uniformly over

$$\left[ \frac{1}{2\mu} - \frac{\beta}{2\alpha\mu}(1-c), \frac{1}{\mu} \right].$$

2. Furthermore,

$$PoA = \frac{3}{2} + \frac{c(\alpha + \beta c)}{2(\alpha + \beta)}. \quad (40)$$

This equals  $3/2$  at  $c = 0$  and  $2$  at  $c = 1$ .

See Figure 5 for an illustration.

Note that for tariff  $0 \leq p \leq \frac{\beta}{2\mu}$ , the cost to each customer remains fixed at  $\frac{\beta}{\mu}$  while this had increased for  $p > \frac{\beta}{2\mu}$ . It is easily seen that by having  $n-1$  separate tariffs so that customers served in the interval  $(\frac{i}{n\mu}, \frac{i+1}{n\mu})$  for  $(i = 0, 1, 2, \dots, n-1)$  are charged amount  $\frac{\beta}{\mu} \frac{n-i-1}{n}$ , we can achieve PoA equal to  $\frac{n+1}{n}$  as in Subsection 6.1.3.

## 7 Numerical Experiments

In our analysis in a single class customer setting, we derived the unique equilibrium arrival profile for an asymptotically limiting fluid regime where the number of customers increased to infinity. We refer to this as the asymptotic equilibrium arrival profile. When the number of customers is finite, the associated equilibrium arrival profile may be more sophisticated and determining it may be a subject for interesting future research. In this section we numerically test the efficacy of the asymptotic equilibrium profile in the fixed  $N$  customer setting for a simple example to get a sense of its closeness to equilibrium in finite- $N$  queue, as  $N$  increases. We consider the case where there are  $N$  single class customers with linear costs that follow two variants of the asymptotic equilibrium strategy: In Case I, the customers select their arrival times by sampling from a uniform distribution over their support. In Case II, the customers arrive at deterministic evenly spaced intervals. As pointed out in Section 2, both cases represent a finite-sample approximation to the uniform fluid distribution. To further contrast the two cases, we assume that customer service times are exponentially distributed in Case I, while they are uniformly distributed with lower variance in Case II. We then, in both the cases, plot the expected cost incurred by a tagged customer as a function of her arrival time for increasing values of  $N$ . We observe that the resulting cost (suitably normalized) converges to a constant as  $N$  increases. This convergence is faster in Case II where the system is less noisy. This suggests that for reasonable values of  $N$ , the asymptotic equilibrium arrival profile may be close to an actual equilibrium arrival profile, although as mentioned earlier, further research is needed to establish this.

**Case I:** We set the linear cost coefficients  $\alpha = 2$  and  $\beta = 1$ . The customer service times are exponentially distributed with rate  $\mu = 1$ . Each arrival selects her arrival time as uniformly distributed in the interval  $N \times [-\frac{\beta}{\alpha\mu}, \frac{1}{\mu}]$ . Customers are served on a first come first serve basis. We use simulation to estimate the expected waiting time and hence the expected cost of the tagged customer that arrives at times  $N \times [-\frac{\beta}{\alpha\mu}, 0, \frac{0.5}{\mu}, \frac{0.8}{\mu}, \frac{0.95}{\mu}, \frac{1}{\mu}]$ . The cost of the customer is normalized by dividing by  $N$ . Figure 6 shows the normalized expected cost for the tagged customer as a function of her normalized arrival time (arrival time divided by  $N$ ) for  $N = 10, 50, 100, 500, 1000$  and  $10000$ . Ten thousand independent simulation replications are conducted to estimate the expected waiting time in each configuration. Typically, the 95% confidence width of the resulting estimator is within 0.5% of the value of the estimator. When,  $N = 10,000$ , and the customer arrives at times  $N \times \frac{0.95}{\mu}$  or at  $N \times \frac{1}{\mu}$ , this ratio was below 3%, again for 10,000 replications.

Note that the normalized expected cost of the tagged customer trivially equals 1 for her arrival time between  $N \times [-\frac{\beta}{\alpha\mu}, 0]$ . As the graph shows, this cost is higher than 1 and is increasing as the arrival time increases to  $\frac{N}{\mu}$ . However, for large  $N$  (for instance,  $N = 1,000$ ) this cost more-or-less stabilizes to 1.

Intuitively, this can be understood by recalling the well known Lindley's recursion

$$W_{n+1} = \max(W_n + S_n - I_{n+1}, 0), \quad (41)$$

where  $W_n$  denotes the waiting time of customer  $n$  in a first come first serve queue,  $S_n$  denotes this customer's service time and  $I_{n+1}$  denotes the inter-arrival time between customer  $n$  and  $n+1$ . In our model all customers that arrive before time zero wait till time zero when the system initiates service. Lindley's recursion is then valid for all customers that arrive after time zero.

Note that, if in our simulations, we set

$$W_{n+1} = W_n + S_n - I_{n+1}, \quad (42)$$

for all arrivals after time zero, then it is easily seen that the resultant normalized expected cost will be 1 for an arrival at any time during  $N \times [0, \frac{1}{\mu}]$ . However, the expected waiting time increases (and hence the expected cost increases) due to the relation (41) assigning higher value to a waiting time compared to (42) whenever an arrival finds an empty queue.

The difference between the two expected costs (one computing waiting time using Lindley's recursion, other using linear recursion) may be small when the probability of the queue emptying between time zero and the time of tagged customer's arrival is small. This probability is obviously small for tagged customer's arrival time close to zero (as there are many customers waiting for service at time zero) and increases as this arrival time gets closer to  $N/\mu$ . It can easily be shown that for a given  $\epsilon \in (0, 1)$ , as  $N$  becomes large, the probability of the queue becoming empty in the interval  $[0, \frac{N(1-\epsilon)}{\mu}]$  goes to zero, and hence the normalized cost stabilizes to 1 with increasing  $N$ .

Note that for finite  $N$ , under a symmetric equilibrium strategy, the tagged customer must see constant cost at all times along the support of other customers arrival distribution. Figure 6 suggests that to achieve this, customers must put relatively less weight towards the end of their support compared to asymptotic equilibrium strategy.

**Case II:** Here, the customers arrive at deterministic equi-spaced time intervals - Customer  $i$  for  $i = 1, 2, \dots, N$  arrives at

$$\left( -\frac{N\beta}{\alpha\mu} + \frac{1}{2\mu} \frac{(\alpha + \beta)}{\alpha} + \frac{(i-1)}{\mu} \frac{(\alpha + \beta)}{\alpha} \right).$$

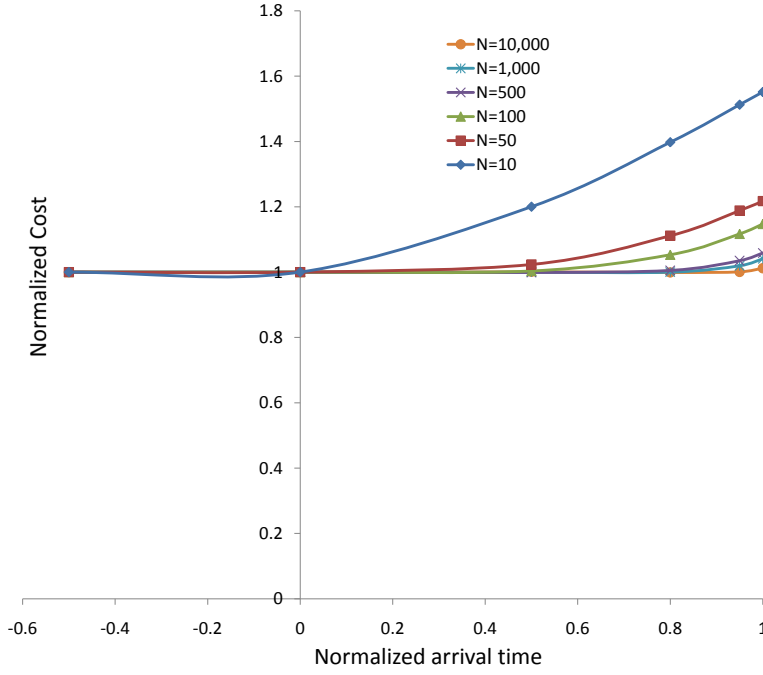


Figure 6: We consider  $N$  single class customers with  $\alpha = 2, \beta = 1$ , service times exponentially distributed with rate  $\mu = 1$ . Customers arrival times are uniformly distributed between  $N \times [-\frac{\beta}{\alpha\mu}, \frac{1}{\mu}]$ . The graph shows the expected cost of a customer arriving to this queue at different times. Cost and time are normalized by dividing by  $N$ .



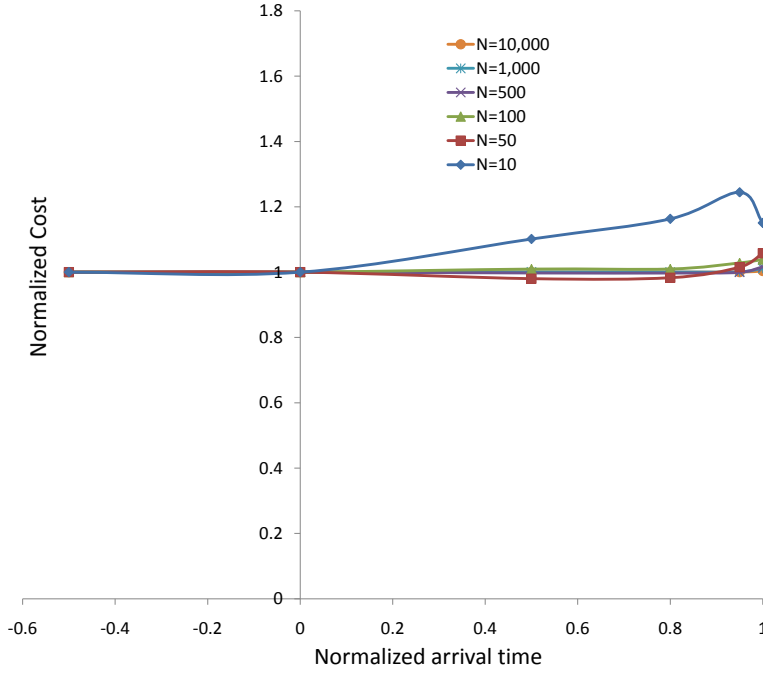


Figure 7: We consider  $N$  single class customers with  $\alpha = 2, \beta = 1$ , service times are uniformly distributed between  $[1/2, 3/2]$ . Customers arrive at deterministic equally spaced intervals. The graph shows the expected cost of a customer arriving to this queue at different times. Cost and time are normalized by dividing by  $N$ .

All parameter values are as in Case I. The service times are assumed to be uniformly distributed between  $[1/2, 3/2]$  ( so their variance equals  $1/12$  as compared to variance of 1 in Case I). This may be more realistic in many applications (such as concert or cafeteria queues) where the service times show little variability. Figure 7 shows the normalized expected cost for the tagged customer as a function of her normalized arrival time as in Case I. As expected, the convergence to 1 is much faster in this case. Note that for small values of  $N$ , the normalized cost may actually be less than 1 for tagged customers arrival at times that are just before the next arrival in the deterministic arrival grid.

## 8 Conclusion

In this paper we considered the queueing problem that may arise in settings such as concert and movie theaters, cafeterias, DMV offices, Black Friday shopping queues, etc., where a large number of customers may queue up before a facility that opens for service at a particular time. The customers strategically select their arrival time distributions to trade-off waiting time in queue with costs due to late arrival. We developed a queueing framework for this problem for which we identified the fluid limit. We observed that the fluid limit allows a great deal of tractability in analyzing the strategic arrival problem faced by each customer. We identified the unique arrival profile for each customer class in equilibrium, and showed that the price of anarchy equals 2 in the single-class model while it varies around this value in the multiclass case. We further discussed structural changes in the queueing discipline and simple pricing schemes that can be used to reduce the price of anarchy. We also demonstrated through a simple numerical example that the proposed asymptotic equilibrium arrival profiles may be close to equilibrium in the finite- $N$  queue, for  $N$  reasonably large.

As part of future work, we plan to study the equilibrium properties of the fluid model under more general cost functions as well as study the model introduced here under the diffusion limit. Extension to multi-server queueing networks would also be of interest in many applications particularly communication networks. We hope that this analysis motivates further research in strategic analysis of queueing systems.

**Acknowledgements:** *Authors would like to thank the reviewers for their careful, helpful and detailed comments that helped improve the manuscript.*

## 9 Appendix: Some Proofs

**Proof of Proposition 4:** From (21) and (22), we have

$$\text{PoA} \leq \frac{2 \sum_{i,j=1}^I \Lambda_i \Lambda_j \beta_i}{\sum_{i,j=1}^I \Lambda_i \Lambda_j \min\{\beta_i, \beta_j\}}. \quad (43)$$

We proceed to bound the last expression, by computing its maximum over  $\mathbf{\Lambda} \geq 0$ , where  $\mathbf{\Lambda} = (\Lambda_1, \dots, \Lambda_I)$ .

Let us reorder the class indices so that  $\beta_1 < \beta_2 < \dots < \beta_I$  (in case there are equal coefficients we can collapse them into a single class). Denote the right-hand side of (43) by  $F(\mathbf{\Lambda})$ , and let  $N(\mathbf{\Lambda})$  and  $D(\mathbf{\Lambda})$  denote the nominator and denominator of that expression. We will show that the maximum of  $F$  is attained when  $\Lambda_2 = \dots = \Lambda_{I-1} = 0$  and  $\Lambda_1/\Lambda_I = \sqrt{\beta_I}/\sqrt{b_1}$ . The required bound is then the value of  $F$  at this point.

A maximizer  $\mathbf{\Lambda}$  of  $F$  must satisfy

$$F'_k(\mathbf{\Lambda}) \triangleq \frac{\partial F(\mathbf{\Lambda})}{\partial \lambda_k} \leq 0, \quad k = 1, \dots, I$$

with equality if  $\Lambda_k > 0$ . Since  $F = N/D$ , we get

$$\frac{N'_k D - D'_k N}{D^2} \leq 0$$

or equivalently

$$\frac{N'_k}{D'_k} \leq \frac{N}{D} \quad (44)$$

with equality if  $\Lambda_k > 0$ .

Consider three consecutive coordinates  $(\Lambda_{k-1}, \Lambda_k, \Lambda_{k+1})$  of the maximizer  $\mathbf{\Lambda}$ , with  $2 \leq k < I - 1$ , and suppose that  $\Lambda_k > 0$ . We will show that this is impossible. Since the right hand side of (44) is independent of  $k$ , we get at this point

$$\frac{N'_{k-1}}{D'_{k-1}} \leq \frac{N'_k}{D'_k} \geq \frac{N'_{k+1}}{D'_{k+1}}$$

which implies that

$$\frac{N'_k - N'_{k-1}}{D'_k - D'_{k-1}} \geq \frac{N'_{k+1} - N'_k}{D'_{k+1} - D'_k}. \quad (45)$$

Now, direct computation of the relevant derivatives gives

$$\begin{aligned} N'_k &\triangleq \frac{\partial N(\mathbf{\Lambda})}{\partial \Lambda_k} = \sum_i \beta_i \Lambda_i + \beta_k \sum_i \Lambda_i \\ D'_k &\triangleq \frac{\partial D(\mathbf{\Lambda})}{\partial \Lambda_k} = 2 \sum_i \Lambda_i \min\{\beta_i, \beta_k\} \end{aligned}$$

so that

$$\frac{N'_k - N'_{k-1}}{D'_k - D'_{k-1}} = \frac{(\beta_k - \beta_{k-1}) \sum_i \Lambda_i}{2(\beta_k - \beta_{k-1}) \sum_{i \geq k} \Lambda_i} = \frac{\sum_i \Lambda_i}{2 \sum_{i \geq k} \Lambda_i}$$

and similarly

$$\frac{N'_{k+1} - N'_k}{D'_{k+1} - D'_k} = \frac{\sum_i \Lambda_i}{2 \sum_{i \geq k+1} \Lambda_i}.$$

Comparing the last two expressions, it is evident that (45) can hold only if  $\Lambda_k = 0$ .

It follows that any maximizer  $\mathbf{\Lambda}$  of  $F$  must have  $\Lambda_2 = \dots \Lambda_{I-1} = 0$ . To determine  $\Lambda_1$  and  $\Lambda_I$ , observe that  $F$  now reduces to

$$\begin{aligned} F(\mathbf{\Lambda}) &= 2 \frac{\beta_1 \Lambda_1^2 + (\beta_1 + \beta_2) \Lambda_1 \Lambda_2 + \beta_2 \Lambda_2^2}{\beta_1 \Lambda_1^2 + 2\beta_1 \Lambda_1 \Lambda_2 + \beta_2 \Lambda_2^2} \\ &= 2 + \frac{2(\beta_2 - \beta_1) \Lambda_1 \Lambda_2}{\beta_1 \Lambda_1^2 + 2\beta_1 \Lambda_1 \Lambda_2 + \beta_2 \Lambda_2^2} = 2 + \frac{2(\beta_2 - \beta_1)}{\beta_1 \lambda + 2\beta_1 + \beta_2 / \lambda} \end{aligned}$$

where  $\lambda \triangleq \Lambda_1 / \Lambda_2$ . Minimizing the last denominator over  $\lambda$  (which is equivalent to maximizing  $F$ ) gives  $\beta_1 - \beta_2 / \lambda^2 = 0$ , or  $\lambda = \sqrt{\beta_2 / \beta_1}$ . Substituting this maximizing value back in  $F$  gives the upper bound  $F(\mathbf{\Lambda}) = 1 + \sqrt{\beta_I / \beta_1}$ . Recalling that the  $\beta_i$ 's were arranged in increasing order, this establishes the claimed upper bound.  $\square$

**Proof of Proposition 6:** To see this result note that in an equilibrium both the populations will arrive in disjoint intervals at rate  $\mu m$ . Suppose that the population 1 arrives uniformly between  $[\tilde{\tau}_0, \tilde{\tau}_1]$ , where  $\tilde{\tau}_1 \leq \frac{1}{2\mu}$  and  $\tilde{\tau}_0 < 0$  (recall that service begins at time zero). The second population arrives uniformly between times  $[\tilde{\tau}_2, \tilde{\tau}_3]$  for  $\tilde{\tau}_3 > \frac{1}{\mu}$  and  $\tilde{\tau}_2 < \frac{1}{2\mu}$  (recall that service for this population begins at  $\frac{1}{2\mu}$ ). Since the cost incurred by the two populations is the same, we have

$$(1 + c) \frac{\beta}{2\mu} + \beta \tilde{\tau}_1 = \beta \tilde{\tau}_3.$$

Since the total service allocated is for time  $\frac{1}{\mu}$ , we have

$$\tilde{\tau}_1 + (\tilde{\tau}_3 - \frac{1}{2\mu}) = \frac{1}{\mu}.$$

It follows that

$$\begin{aligned} \tilde{\tau}_3 &= \frac{1}{\mu} + \frac{c}{4\mu}, \\ \tilde{\tau}_1 &= \frac{1}{2\mu} - \frac{c}{4\mu}, \end{aligned}$$

and the cost incurred by each customer equals  $\frac{\beta}{\mu}(1 + \frac{c}{4})$ . The proportion of customers coming in as population 1 equals  $\mu\tilde{\tau}_1 = (\frac{1}{2} - \frac{c}{4})$ .  $\tilde{\tau}_0$  and  $\tilde{\tau}_2$  can be easily seen to be as specified in the proposition since the arrival rates for each population are known.

To compute PoA, note that the revenue from population 1 equals

$$\beta\tilde{\tau}_1(\frac{1}{2} - \frac{c}{4}) = \frac{\beta}{\mu}(\frac{1}{2} - \frac{c}{4})^2.$$

The revenue from population 2 equals

$$\beta\tilde{\tau}_3(\frac{1}{2} + \frac{c}{4}) = \frac{\beta}{\mu}(1 + \frac{c}{4})(\frac{1}{2} + \frac{c}{4}),$$

so that (39) follows. □

**Proof of Proposition 7:** It can be argued as in the proof of Proposition 5 that population 1 arrives uniformly between  $[\check{\tau}_0, \frac{1}{2\mu}]$ , at rate  $\mu m$  and is served till time  $\check{\tau}_1$ , for some  $\check{\tau}_1 > \frac{1}{2\mu}$  and  $\check{\tau}_0 < 0$ . Also, population 2 arrives uniformly between  $[\check{\tau}_2, \frac{1}{\mu}]$  at rate  $\mu m$  for some  $\check{\tau}_2$ . Note that the cost incurred by the two populations is the same and that specifies  $\check{\tau}_1$  through the equation

$$(1 - c)\frac{\beta}{2\mu} + \alpha(\check{\tau}_1 - \frac{1}{2\mu}) + \beta\check{\tau}_1 = \frac{\beta}{\mu},$$

or

$$\check{\tau}_1 = \frac{1}{2\mu} + \frac{\beta c}{2\mu(\alpha + \beta)}.$$

The proportion of customers coming in as population 1 equals

$$\mu\check{\tau}_1 = \frac{1}{2} + \frac{\beta c}{2(\alpha + \beta)}.$$

It is then easily inferred that  $\check{\tau}_0 = -\frac{\beta}{2\alpha\mu}(1 + c)$ . Similarly, since  $\frac{1}{2} - \frac{\beta c}{2(\alpha + \beta)}$  proportion of population is of type 2, we can evaluate that  $\check{\tau}_2 = \frac{1}{2\mu} - \frac{\beta}{2\alpha\mu}(1 - c)$ .

We now compute the PoA. Note that each customer in population 1 incurs a cost

$$\alpha(\check{\tau}_1 - \frac{1}{2\mu}) + \beta\check{\tau}_1 = \frac{\beta}{2\mu}(1 + c).$$

Each customer in population 2 incurs cost  $\frac{\beta}{\mu}$ . From this we can determine the total cost and after algebraic manipulations conclude that (40) holds. □

## References

- [1] H. Chen and D. Yao, *Fundamentals of queueing Networks*, Springer-Verlag, 2001.
- [2] G. Gilboa-Freedman, R. Hassin and Y. Kerner, “Price of anarchy in the Markovian single server queue”, preprint, 2009.
- [3] A. Glazer and R. Hassin, “ $M/M/1$ : On the equilibrium distribution of customer arrivals”, *European J. of Oper. Research* 13:146-150, 1983.
- [4] R. Hassin and M. Haviv, *To Queue or Not to Queue*, Kluwer Academic Publishers, 2003.
- [5] R. Hassin and Y. Kleiner, “Equilibrium and optimal arrival patterns to a server with opening and closing times”, preprint, 2009.
- [6] M. Haviv and T. Roughgarden, “The Price of Anarchy in an Exponential Multi-Server”, *Operations Research Letters* 35(4):421-426, 2007.
- [7] C. A. Jr. and R. Sherman, “Waiting Line Auctions”, *Journal of Political Economy* 90 (2), 280-294
- [8] M.A. Lariviere and J. A. van Mieghem, “Strategically seeking service: How competition can generate Poisson arrivals”, *Manufacturing and Service Operations Management* 6(1):23-40, 2004.
- [9] P. Lederer and L. Li, “Pricing, Production, Scheduling, and Delivery-time Scheduling”, *Operations Research* 45(3):407-420, 1997.
- [10] S. Juneja and R. Jain, “The Concert/Cafeteria Queuing Problem: A Game of Arrivals”, ICST/ACM Fourth International Conference on Performance Evaluation Methodologies and Tools - ValueTools 2009. Received the Best Paper Award.
- [11] R. Lindley, “Existence, uniqueness, and trip cost function properties of user equilibrium in the bottleneck model with multiple user classes”, *Transportation Science* 38(3):293-314, 2004.
- [12] P. Naor, “The regulation of queue size by levying tolls”, *Econometrica* 37(1):15-24, 1969.

- [13] G.F. Newell, “The morning commute for nonidentical travellers”, *Transportation Science* 21(2), 74-88, May 1987.
- [14] A. Rapoport, W.E. Stein, J.E. Parco and D.A. Seale, “Strategic Play in Single-Server Queues with Endogenously Determined Arrival Times, *Journal of Economic Behavior and Organization* 55:67-91, 2004.
- [15] H.L. Royden, *Real Analysis*, 3rd ed., Prentice Hall, 1988.
- [16] D. Schmeidler, “Equilibrium points of nonatomic games”, *J. Stat. Physics* 7(4):295-300, 1973.
- [17] J. Van Mieghem, “Price and service discrimination in queueing systems: incentive compatibility of  $Gc\mu$  scheduling”, *Management Science*, 46(9):1249-1267, 2000.
- [18] W.S. Vickrey, “Congestion theory and transport investment”, *The American Economic Review* 59:251-260, 1969.
- [19] S. Wang and L. Zhu, “A dynamic queueing model,” *The Chinese Journal of Economic Theory* 1:14-35, 2004.
- [20] W. Whitt, *Stochastic Process Limits*, Springer-Verlag, 2002.