Saint Petersburg State University
Department of mathematical game theory and statistical decisions

Kasimova Yana Aleksandrovna

Master's thesis

Optimal information disclosure in multi-server queueing games

Specialization 01.04.02
Applied Mathematics and Informatics
Master's Program Game Theory and Operations Research

Research advisor,
Cand. Sc. (Phys.-Math),
assistant
Kostyunin S. Yu.

Saint-Petersburg
2018

Contents

Introduction

Queueing Theory is considered as a branch of applied probability theory. Its applications are in different fields, e.g. different services, communication networks, computer systems, machine plants and so forth. It occupies a considerable place in our life. We are constantly confronted with queues in our everyday life and every time we decide to join the queue or leave it. For this area there exists a huge body of publications.

The subject of queueing theory can be described as follows: consider a center which served a customers and a population of customers, which at some times enter the service center in order to obtain service. It is often the case that the service center can only serve a limited number of customers. If a new customer arrives and the service is full, he enters a queue and waits until the service facility becomes available. So we can identify three main elements of a service center: a population of customers, the service facility and the queue. As a simple example of a service center consider an car washing: many drivers come to the car wash, but car washing can only serve one customer at a time, A newly arriving and friendly drivers proceeds directly to the end of the queue, if the service facility (the car wash) is busy. This corresponds to a FIFO service (first in, first out).

Some examples of the use of queueing theory in networking are the dimensioning of buffers in routers or multiplexers, determining the number of trunks in a central office in POTS, calculating end-to-end throughput in networks and so forth.

Also service center can be presented as several separate servers.For instance, owner of our car washing has become rich, and he can buy a new machine for car washing, in such way we get two machines for washing so it means we have several servers in our model. By using information about multi-servers model and its main characteristics we consider case when our queue model has k servers.

Queueing Theory tries to answer questions like e.g. the mean waiting time in the queue, the mean sojourn time (waiting time in the queue plus service times), mean utilization of the service facility, distribution of the number of customers in the queue, distribution of the number of customers in the system and so forth. These questions are mainly investigated in a stochastic scenario, where e.g. the interarrival times of the customers or the service times are assumed to be random.

The server that serves customers in the queue always seeks to maximize its revenue.

To maximize its revenue the server in various ways is trying to attract customers to the queue. One of these methods is the server's information policy, that is, to hide or not hide information about the queue, or to hide from a certain moment.

The study of the information policy of the server begins with the article Naor [1], he has influenced on literature of the impact of queue length disclosure on consumer welfare and producer's revenue.

In Naor's model, homogeneous customers see the current queue length upon their arrival at any state of the system and make decisions to join the queue or balk. Thus, the service provider discloses full queue length information, and such setting would be referred to as an observable queue in the literature to come.

Later information policy was considered in various aspects. Edelson and Hilderbrand [2] studied the unobservable version of Naor's model where customers do not have any queue length information upon arriving to the service system.

After these papers, the observable and unobservable queues were studied when there are heterogeneous service values(e.g., Larsen [3]), heterogenous time costs (e.g., Afe'che and Mendelson [4]) and so on.

Readers who are interested in the comprehensive review by Hassin and Haviv [5] and the references in Hassin [6]. In particular, Hassin [7], Chen and Frank [8] and Shone et al. [9] compared the outcomes between an observable queue and an unobservable queue, and identified conditions when one setting is superior than the other.

We consider state-dependent information policy, which include the abovementioned observable and unobservable cases as two extreme events.

Hassin and Roet-Green [10] considered a model where arriving customers to an unobservable queue can acquire the queue length information for a fee, and Hu et al. [11] studied a scenario where a portion of the customer population are fully aware of the queue length information while others are fully not. In both of these papers (plus Simhon et al. [12] which we will discuss in details later), the system is observable to some customers while unobservable to others.

To give value to our model, consider customers calling a customer-support phone number or engaged in some service live chat. The service provider can announce the current queue length information to customers who arrive at certain states but not at other states (hereafter, referred to as "informed" and "uninformed" states, respectively). Customers make strategic decisions to join the queue or balk with the information given by the service provider. If the goal of the server is to maximize revenue/throughput, then intuitively it makes sense for the service provider to inform customers of the queue length when the queue is short (i.e., it is less than some threshold D) and hide the queue length information when the queue is long.

However, Simhon et al. [12] proved that such a policy is never optimal for the service provider, and the optimal policy is either an observable queue or an unobservable queue. We note that in Simhon et al. [12] even though a customer who arrives at an uninformed state is not informed of the current queue length, she is assumed to be aware of the threshold disclosure policy of the server and the value of D (e.g., she is an experienced customer who uses the service frequently). In other words, when an arriving customer is not informed of the current queue length, she can infer that the current queue length must equal or exceed D and uses such information to make a decision to join the queue or balk. This assumption is reasonable since customers can be based on experience, because they often use server services. We try to find out which policy is optimal for server, and also consider policy which optimal for social welfare.

There are exists one more case, when customers who arrive at the uninformed states are not aware of the disclosure policy used by the service provider, so these uninformed customers behave as if they had arrived at an unobservable queue. This assumption is also reasonable, since customers are either first-time or non-frequent service users who do not possess much knowledge of the service provider's routine information disclosure decisions (e.g., calling a customer service hot-line once in a while).

Consider a similar but different settings in this work.

We review abovementioned two cases, when customer are aware about disclosure policy, as was described by Simhon [12], and when customer are not aware about information policy, as was presented by Shiliang [13] but our model will be provided by k servers for customer service.

So that to find the optimal information disclosure policy we consider game between clients, and get equilibrium behavior of clients in the system. Queue was considered as a game by Hassin [6]. Customers in service systems act independently in order to maximize their welfare. Yet, each customer's optimal behavior is affected by acts taken by the system managers and by the other customers. And the intresting thing that the result is an aggregate "equilibrium" pattern of behavior which may not be optimal from the point of view of society as a whole.

Our work treats to the large volume of literature that studies the influence of information disclosure (i.e., queue length, real-time waiting time, service rate, service quality, etc.) to arriving customers. For example, Guo and Zipkin [14] examine M/M/1 queues with three levels of information: no information, partial information (the queue length), and full information (the exact sojourn time). In Economou and Kanta [15] and Guo and Zipkin [16], some partitioned queue information (such as range of queue lengths) is available to the customers to make their decisions, join or balk. Armony et al. [17] examined a setting where the server announces the delay information of the last customer to enter service, and it was benchmarked to the fixed delay model where long-run average delay information is announced. In Allon et al. [18], non-verifiable queue length information is provided to the customers. Cui and Veeraraghavan [19] investigated the influence of service rate revelation when customers do not have the information. Kremer and Debo [20] used lab experiments to examine the influence of waiting time on customers' acquiring behavior when the service quality is known to some consumers but unknown to others. Yu et al. [21] provided empirical evidence on the influence of delay announcements on customer behavior. Nevertheless, in all of these papers, the information disclosure policy is not state-dependent.

We consider first case, when customers are aware about information disclosure policy in first chapter, then we investigate a scenario, when customers are not aware about information discosure policy in second chapter, after that we will make a conclusion about information disclosure policy in third chapter.

1 Optimal information disclosure policies in a multiple-server queue

1.1 Model settings

Consider a multiple-server queue (M/M/k). Potential clients arrive to the system according to a Poisson process with rate λ, upon their arrival customers make a decision to either join the queue or balk, here reneging is not allowed, it means that if customer join the queue he cannot balk this queue after joining. Joining customers are served follow first come first out discipline. Service times of the customers are independent and exponentially distributed with rate μ. All joining customers incur a linear waiting cost of C per unit of time while they standing in the system, it means that they pay waiting cost while they in the queue and when they served in the server. Without loss generality, consider waiting cost C equals 1.

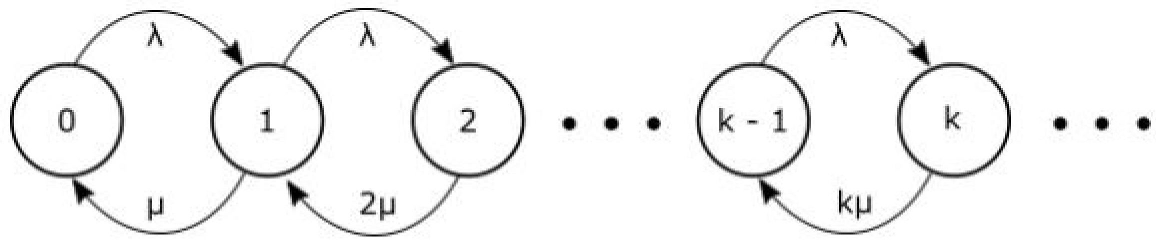

Figure 1.1: multiple-server queue (M/M/k).

Upon completion of the service, each of them receives reward R which is net of some constant price of the service P.

Let $R > \frac{C}{\mu}$ because otherwise all customers will balk from the queue, since it is not profitable for them.

Also we consider k servers in the system, if customer arrive to the system and one of the servers is idle, then this server get new arrived customer, if all servers are busy, then which first will become idle, thus get new arrived customer.

The state of the system, $i \geq 0$, is the total number of customers(including the customer in the queue and at the server). The servers implement state-dependent queue length disclosure policies. It means that decision of provider conceal or reveal information about queue length to a new client depends on current queue length. Such a policy can be defined as $X = (x_0, x_1, \dots, x_i, \dots)$,, here we use terminology from Shiliang [13], where

$$x_i = \begin{cases} 1, & \text{the server reveals the current queue length at state } i \text{ ;} \\ 0, & \text{the server conceals the current queue length at state } i. \end{cases} \tag{1}$$

For example, only concealing odd states can be represented by policy $X^{odd} = (1, 0, 1, 0, 1, \dots)$.

Write a disclosure information policy for an integer threshold $D \in (0, 1, 2, \dots, \infty)$ as X^D where

$$x_i = \begin{cases} 1, & \text{if } 0 \leq i < D; \\ 0, & \text{if } i \geq D \end{cases} \tag{2}$$

X^D define that the servers should disclosure the current length of the queue if length of queue less then D and conceal the queue length, if length equals or more than D.

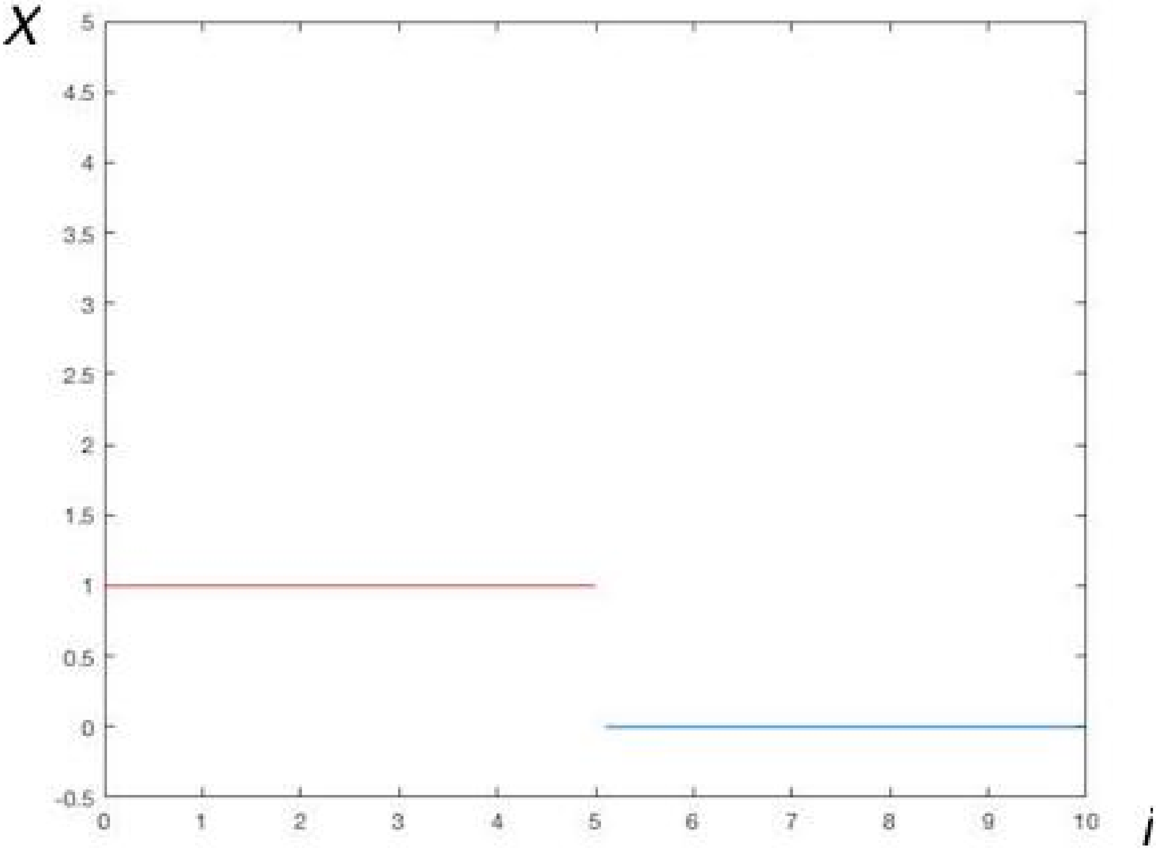

Figure 1.2: X^D where $D = 5$.

If servers never conceal information, then we have the observable queue system, it means that we can always find out the queue length, let denote this disclosure information policy as X^∞ according definition of disclosure policies of servers. In contrast, there are exist the information disclosure policy, which is always conceal information about the queue length, i.e. $x_i = 0 \quad \forall i$, which described by Edelson and Hiderband [2]. This threshold policy is equivalent to the unobservable queue, it means that customers never can know about queue length. Let denote this disclosure information policy as X^0.

For notational convenience, when $D \in (-1, -2, \dots)$ define X^D as follows

$$x_i = \begin{cases} 1, & \text{if } i = -D; \\ 0, & \text{if } i \neq -D \end{cases} \tag{3}$$

1.2 Client's Decisions

Let the servers implements some policy $X = (x_0, ..., x_i, ...)$. Consider a client, who arrives at some state $i \geq 0$. If $x_i = 1$, the client observes the current queue length i, and he joins the system if his expected benefit more than zero, and balks if it is not. Expected benefit can define as follows:

$$R - (\frac{1}{\mu} + \frac{i-k+1}{k\mu})C \tag{4}$$

where R reward for service, C cost of staying in the system per unit time, and $\frac{1}{\mu} + \frac{i-k+1}{k\mu}$ is waiting time for i client.

This is observable model described in Naor [1]. Let L denote Naor's balking threshold, so

$$L = \lfloor (1 + \mu(\frac{R}{C} - \frac{1}{\mu}))k \rfloor \tag{5}$$

here we have brackets, which round of the real number to the nearest smaller whole. Then the client decision upon arrival at state i when $x_i = 1$ can be characterized as

$$\begin{cases} i < L, & \text{the client joins the queue ;} \\ i \geq L, & \text{the client balks} \end{cases} \tag{6}$$

If $x_i = 0$, the arriving client does not know about the current queue length i. As a consequence, client behaves as if the system were totally unobservable, which is the model described in Edelson and Hilderbrand [2]. To be precise, in this case, when client does not know about queue length

we assume that every client who does not observe the current queue length has a belief that the servers's disclosure policy is X^0. This is reasonable assumption when clients do not shop the service frequently. In contrast in Simhon et al. [12] assume that if $x_i = 0$ and the servers have X^D information disclosure policy then the arriving client is aware that current queue length more then D. So we have two cases when arriving clients are not aware about the current queue length at all and when they are aware that queue length more than some threshold, which used by the servers in their information disclosure policy. Consider these two cases, in second chapter when they have some knowledgments about queue length, with condition that $D > L$, because if it is not then clients always balk the queue and in third chapter when clients do not know about queue length.
In both cases we have, that if client upon arrival does not know about the current queue length then he joins to the queue with probability q and balks the queue with probability $1 - q$, let denote q as joining probability.

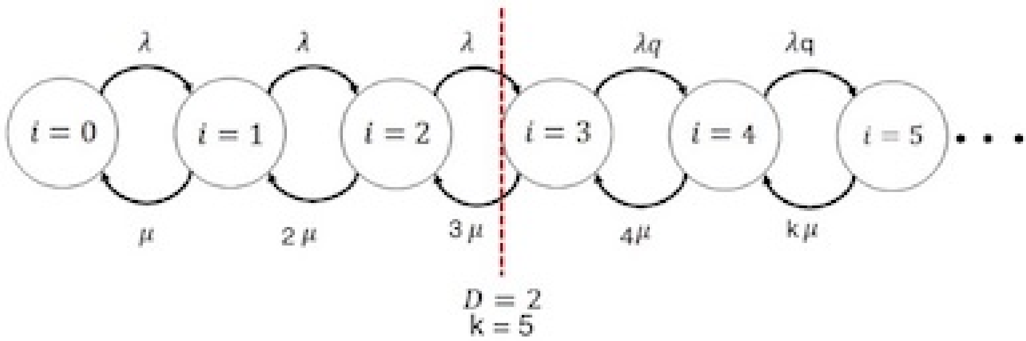

Figure 1.3: M/M/k with $D = 2$, $k = 5$ threshold.

1.3 Threshold Policies

We consider the set of threshold policies considered in Simhon et al. [12], it means, that servers have X^D for $D \geq 0$(including X^0 and X^∞) information disclosure policy. We obtain the revenue-maximizing policy from this consideration set.

Note that when $D > L$, the underlying queue system under policy X^D is M/M/k/L(it means no state higher that L will be ever reached). It leads to the identical system(governed by X^∞) in Naor [1]. Therefore, it is sufficient to search for the optimal policies in the reduced set $(X^0, X^1, ..., X^L, X^\infty)$.

1.4 Optimization problem

It the next chapters we consider information disclosure policy so that to achieve the main goal to obtain the revenue-maximizing policy. This policy to provide the maximum payoff for servers. It aims to maximize the utilization of the queue. Utilization of the queue can be defined as follows:

$$U = \frac{E[N_s]}{k}, \tag{7}$$

where N_s expected number of served clients, X the set of considered information disclosure policies. If we know steady-state probability $\pi_0, \pi_1, ...$, then

$$E[N_s] = \sum_{i=0}^{k-1} i\pi_i + k\sum_{i=k}^{\infty} \pi_i \tag{8}$$

The last term is Erlang C formula, we can rewrite as

$$k\sum_{i=k}^{\infty} \pi_i = k\sum_{n=k}^{\infty} \pi_n = \frac{A^k}{k!}\frac{k^2}{(k-A)}\pi_0 \tag{9}$$

where $A = \frac{\lambda}{\mu}$. This formula represents the proportion of time that all k servers are busy, it means that this formula shows probability that client will be wait in the queue.

Also, maximum of utilization of the queue it means maximum λ_{eff}, which is equivalent to minimum of the idle stationary probability of the queue, denoted by π_0. Then optimization problem for the provider is the following:

$$\min_X \pi_0 \tag{10}$$

In this work we consider optimization problem of minimizing the idle stationary probability, but we are going to consider utilization so that find optimal information disclosure policy in near future.

In the next two chapters we consider two cases, in the second it will be case, when clients are aware about information disclosure policy and in the third chapter consider when clients are not aware about policy, after that we compare these two cases in conclusion.

2 Optimal information disclosure policies of provider for experienced clients

2.1 Equilibrium analysis of threshold policies

Consider first case, when we assume that the uninformed clients are aware of the policy used by the provider (this information can be obtained by trials or via exogenous sources).An uninformed customer makes her decision based on her expected sojourn time, denoted by W. Thus, an uninformed customer can evaluate her sojourn time given the system parameters and the provider policy.

Since an uninformed client knows that the queue length is greater than D, his expected sojourn time, defined $W(D,q)$, depends on the threshold D and on the decision of the other uninformed clients, and abovementioned joining probability denoted by q, since q is probability then $q \in [0,1]$, as all client are identical, hence we only consider symmetric strategies. We denote an equilibrium solution by q.

If $D \geq L-1$, uninformed clients will never join the queue because their expected waiting times, conditioned on the fact that the queue length is at least L, is necessarily higher than the reward R (recall that $L = \lfloor (1+\mu(\frac{R}{C} - \frac{1}{\mu}))k \rfloor)$. Thus, the unique equilibrium is $q = 0$ and a threshold policy with a threshold $D \geq L-1$ is equivalent to the observable model.

Henceforth, we only consider threshold policies with $D \in 0, 1, .., L2$ .Thus, at equilibrium, all informed clients join, while all uninformed customers join with probability q. The evolution of the queue length process forms a Markov chain with transition rate λ from state i to state $i+1$ if $i \leq D$ and transition rate λq otherwise. The transition rate from i to $i-1$ is μ for all $i > 0$.

Consider steady-state probability for our model and let denote $A = \frac{\lambda}{\mu}$

$$\pi_n = \frac{A^n \pi_0}{n!}, \qquad n < k < D \tag{11}$$

$$\pi_n = \frac{A^n \pi_0}{k! k^{n-k},} \qquad D+1 \geq n \geq k \tag{12}$$

$$\pi_n = \frac{A^n \pi_0 q^{n-D-1}}{k! k^{n-k},} \qquad n > D+1 \tag{13}$$

The idle stationary probability(i.e., the probability that the queue is empty) as a function of the threshold D and the joining probability q is given by

$$\pi_0 = \Big(\sum_{n=0}^{k-1} \frac{A^n}{n!} + \sum_{n=k}^{D+1} \frac{A^n}{k! k^{n-k}} + \sum_{n=D+2}^{\infty} \frac{A^n q^{n-D-1}}{k! k^{n-k}}\Big)^{-1} = \tag{14}$$
$$= \Big(\sum_{n=0}^{k-1} \frac{A^n}{n!} + \sum_{n=k}^{D+1} \frac{A^n}{k! k^{n-k}} + \frac{A^{D+2} q}{k!} \frac{k}{(k - Aq)}\Big)^{-1}$$

We define $\pi(i|i > D)$ to be the probability that the queue length is i, given that the queue length is higher than D. Using conditional probability rules, we obtain

$$\pi(i|i > D) = \frac{\pi_i}{\sum_{j=D+1}^{\infty} \pi_j}. \tag{15}$$

The expected waiting time for an uninformed new client is

$$W(D, q) = \sum_{i=D+1}^{\infty} (\frac{1}{\mu} + \frac{i-k+1}{k\mu}) \pi(i|i > D) \tag{16}$$

here, in brackets is expected time in the system for new client, if the system already contains i clients. As a result we get waiting time

$$W(D, q) = \frac{\sum_{i=D+1}^{\infty} \pi_i (\frac{1}{\mu} + \frac{i-k+1}{k\mu})}{\sum_{i=D+1}^{\infty} \pi_i} = \tag{17}$$
$$= \frac{1}{\mu} + \frac{D+2-k}{k\mu} + \frac{\lambda q}{(k\mu - \lambda q) k\mu}$$

As we know waiting time we can determine the equilibrium strategy q^* which determines if an uninformed client joins the queue or not. First of all we should show the existence and uniqueness of the equilibrium.

Lemma 2.1. A game with a threshold information policy, which was described above has a unique equilibrium

Proof. Joining when the queue length is smaller than L and balking otherwise is a dominant strategy for the informed clients, hence unique. We next consider the behavior of the uninformed clients.

Consider profit for each uninformed client, who just arrived to the system

$$R - CW(D, q)$$

recall that in this model we set up $C = 1$, then profit for client will be $R-W(D, q)$, where q is strategy of another clients, who arrived before current client.

Consider three possible cases

- $R - W(D, q) > 0$ then client choose to join to the queue, because service give him positive profit, then $q = 1$,
- $R-W(D, q) < 0$ client balks, because profit from service will be negative for client, then $q = 0$,
- $R-W(D, q) = 0$ this is situation when the client does not care to deviate or join the queue, and this scenario forms a mixed equilibrium, because if client wants to deviate from the strategy that the previous clients have chosen, he will not benefit more than zero in any case.

From Eq.(17), we notice that the expected payoff of an uninformed client that joins (i.e., $W(D, q)$) is decreasing with q, while the expected payoff of balking R does not depend on q. So payoff function for client $R-W(D, q)$ is a monotone non-increasing function in q, and we have *avoid the crowd*(ATC) property(see Hassin [6] which guarantees that the equilibrium is unique. It means that the higher is the threshold adopted by others, the lower is the threshold giving the best response for a given client

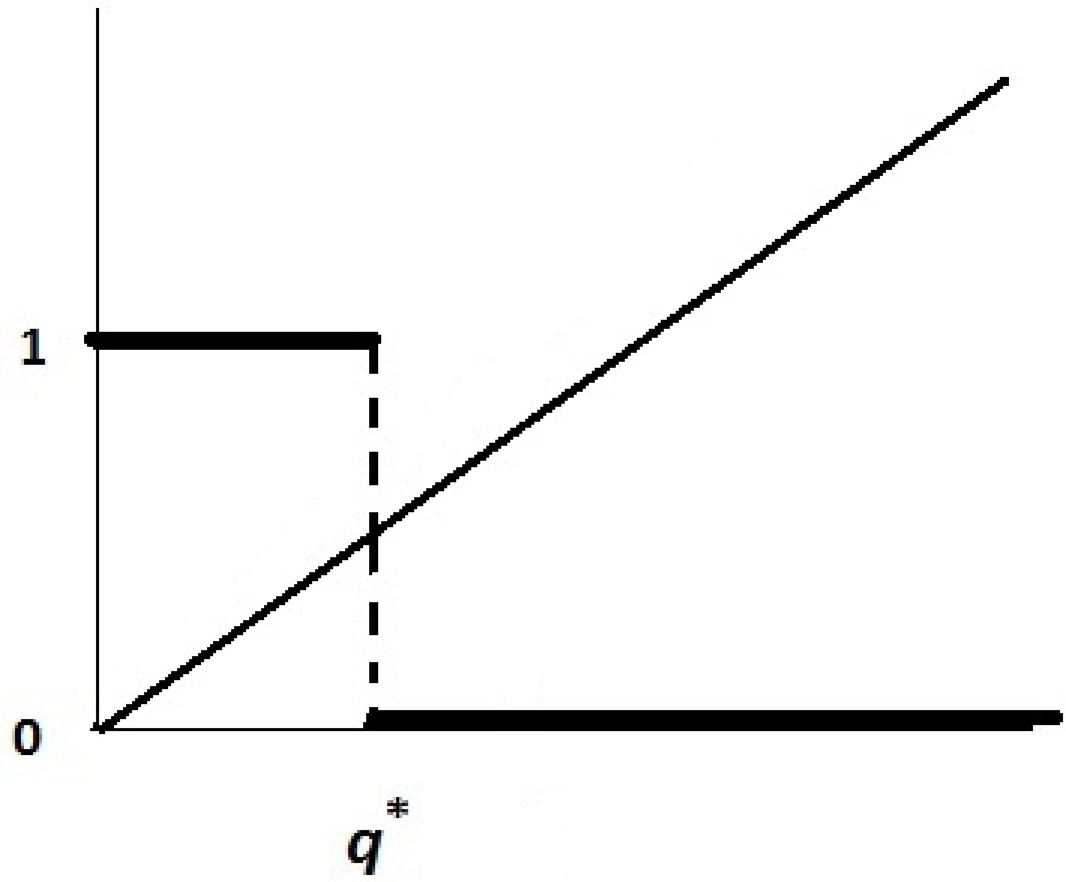

Figure 2.1: Avoid the crowd.

Here we see on Fig.2.1. that we have only one intersection and it proves that we have a unique equilibrium q^*.■

We start the equilibrium analysis with the simple case $\frac{\lambda}{k\mu} < 1$ and reward R more than waiting time multiplied by C per unit of time. In this case, if used policy $\pi(X^0)$, then all clients join the queue because the reward more than costs of staying in the queue. This outcome is obviously optimal for the provider and no other policy can outperform it. In any other case, an equilibrium where all clients join does not exist, define at equilibrium $q < 1$.

Henceforth, we focus on that case which implies that either the system is overloaded, that is $\frac{\lambda}{k\mu} > 1$, or it is underloaded with bounded reward, that is $\frac{\lambda}{k\mu} < 1$ and $R < W * C$

The unique equilibrium might be a pure equilibrium with $q^* = 0$ or a mixed equilibrium with $q^* \in (0, 1)$. In this scenario of a pure equilibrium, an uninformed client is aware about the queue length is exactly $D + 1$(as no uninformed clients joins the queue), and hence his expected waiting time is $\frac{D+2}{k\mu}$. In order for $q = 0$ to be an equilibrium, an uninformed customer should not be better off by joining, which implies that

$$D + 2 \geq kR\mu \geq \lfloor kR\mu \rfloor = L$$

As we only consider threshold policies with $D \leq L - 2$, we deduce that that a pure equilibrium exists only if $kR\mu = D + 2$

Consider the case of having a mixed equilibrium. We obtain the fraction of uninformed clients that join the queue, by using the property that at a mixed equilibrium, each player must be indifferent between the actions of joining and not joining the queue. Hence, at equilibrium $W(D, q) = R$. We obtain our joining probability

$$q^* = \frac{k\mu(kR\mu - D - 2)}{(kR\mu - D - 1)\lambda} \tag{18}$$

Since a pure equilibrium with $q^* = 0$ only exists when $(1 + \mu(\frac{R}{C} - \frac{1}{\mu}))k = kR\mu = D + 2$, the equation above contains both the case of a mixed equilibrium and the case of a pure equilibrium.

So, we receive the clients' equilibrium strategy under an information disclosure policy, which adopted by the provider. To recap this result we can introduce by theorem.

Theorem 2.1. If a provider uses the information disclosure policy X^D, all informed clients join the queue if the queue length is strictly smaller than L. uninformed clients join the queue with probability $q^*(D)$, where

$$q^*(D) = \begin{cases} 0, & \text{if D} \geq L-1, \\ \frac{k\mu(kR\mu - D - 2)}{(kR\mu - D - 1)\lambda}, & \text{if } otherwise. \end{cases} \tag{19}$$

2.2 Comparison of information disclosure policies

With this results about joining probability $q^*(D)$ we can define the optimal information disclosure policy that the provider should pursue so that optimize its revenue. We derive and compare the idle stationary probabilities, at equilibrium, for three types of information disclosure policies: X^D, X^0, X^∞. We ignore the case of $D \geq L - 1$ as it is equivalent to the policy X^∞

For deriving the idle stationary probability when a threshold policy is X^D is used, we substitute joining probability, which we get in previous theorem

$$\pi_0 = \Big(\sum_{n=0}^{k-1} \frac{A^n}{n!} + \sum_{n=k}^{D+1} \frac{A^n}{k!k^{n-k}} + \sum_{n=D+2}^{\infty} \frac{A^n q^{*n-D-1}}{k!k^{n-k}}\Big)^{-1} = \tag{20}$$

$$= \Big(\sum_{n=0}^{k-1} \frac{A^n}{n!} + \sum_{n=k}^{D+1} \frac{A^n}{k!k^{n-k}} + \sum_{n=D+2}^{\infty} \frac{A^n k\mu(kR\mu - D - 2)^{n-D-1}}{(kR\mu - D - 1)\lambda^{n-D-1} k!k^{n-k}}\Big)^{-1}$$

recall that $A = \frac{\lambda}{\mu}$

When sharing queue length information with all customers, we have, at equilibrium, an M/M/k/L queue, it means that we have queue with finite capacity L. The idle stationary probability is

$$\pi_0 = \Big(\sum_{n=0}^{k-1} \frac{A^n}{n!} + \sum_{n=k}^{D+1} \frac{A^n}{k!k^{n-k}} + \sum_{n=D+2}^{L} \frac{A^n}{k!k^{n-k}}\Big)^{-1} \tag{21}$$

To compare between the performances of the different policies, we distinguish between the case of an overloaded queue, it means that $\frac{A}{k} > 1$ and when we have underloaded queue, it means that $\frac{A}{k} < 1$. Note that the strategic behavior of the clients guarantees that the effective load is always smaller that one.

So we show that in an overloaded system, for any value of R and $D \in (0, 1, ..., L-2)$ we always have $X^D(q^*) > X^\infty$.

Theorem 2.2. Considering the set of deterministic threshold based information disclosure policies. To maximize its revenue, a service provider must use the full information disclosure policy X^∞ if we have overloaded system

Proof. Consider case when our queue is overloaded, then compare π_0 when we have information disclosure policy X^D and π_0 when we have X^∞

$$\pi_0 = \Big(\sum_{n=0}^{k-1} \frac{A^n}{n!} + \sum_{n=k}^{D+1} \frac{A^n}{k!k^{n-k}} + \sum_{n=D+2}^{\infty} \frac{A^n k\mu (kR\mu - D - 2)^{n-D-1}}{(kR\mu - D - 1)\lambda^{n-D-1} k!k^{n-k}}\Big)^{-1} =$$

And

$$\pi_0 = \Big(\sum_{n=0}^{k-1} \frac{A^n}{n!} + \sum_{n=k}^{D+1} \frac{A^n}{k!k^{n-k}} + \sum_{n=D+2}^{L} \frac{A^n}{k!k^{n-k}}\Big)^{-1}$$

we should consider only the last terms, and show that the last term of π_0 when we have information disclosure policy X^D less than last term of pi_0 when we have X^∞. We should recall that $q^* = 0$ if $D \geq L-1$ so if we have $D \geq L-1$ than

$$\pi_0(X^\infty)$$

less than $\pi_0(X^D)$, therefore consider $D < L-1$

$$\sum_{n=D+2}^{\infty} \frac{A^n k\mu (kR\mu - D - 2)^{n-D-1}}{(kR\mu - D - 1)\lambda^{n-D-1} k!k^{n-k}}$$

and

$$\sum_{n=D+2}^{L} \frac{A^n}{k!k^{n-k}}$$

consider

$$\sum_{n=D+2}^{\infty} \frac{A^n k\mu(kR\mu - D - 2)^{n-D-1}}{(kR\mu - D - 1)\lambda^{n-D-1} k!k^{n-k}} =$$

$$= \sum_{n=D+2}^{\infty} \frac{(kR\mu - D - 2)}{(kR\mu - D - 1)}^{n-D-1} \frac{\lambda}{k\mu}^{D+1} \frac{1}{k!k^{n-k}} =$$

$$= \frac{\lambda}{k\mu}^{D+1} \frac{1}{k!k^{D+1-k}} \sum_{n=D+2}^{\infty} \frac{(kR\mu - D - 2)}{(kR\mu - D - 1)}^{n-D-1} \frac{1}{k^{n-D-1}}$$

we see that

$$\frac{(kR\mu - D - 2)}{(kR\mu - D - 1)k} < 1$$

so we have geometric progression with an infinite sum of the denominator is less than 1, so we can get

$$\pi_0(X^D) = \frac{\lambda}{k\mu}^{D+1} \frac{1}{k!k^{D+1-k}} \frac{k(kR\mu - D - 1) - (kR\mu - D - 2)}{k(kR\mu - D - 1)}$$

and via geometric progression get $\pi_0(X^\infty)$

$$\pi_0(X^\infty) = \frac{\lambda}{k\mu}^{D+1} \frac{1}{k!k^{D+1-k}} \frac{1 - \frac{\lambda}{k^2\mu}^{L-D-1}}{1 - \lambda^{L-D-1}}$$

reduce multipliers in $\pi_0(X^D)$ and $\pi_0(X^\infty)$ we should compare only these multipliers

$$\frac{k(kR\mu - D - 1) - (kR\mu - D - 2)}{k(kR\mu - D - 1)}$$

and

$$\frac{1-\frac{\lambda}{k^2\mu}^{L-D-1}}{1-\lambda^{L-D-1}}$$

as we see first multiplier less than 1, but the second multiplier more than 1, so as a result we get that

$$\pi_0(X^D) > \pi_0(X^\infty)$$

but if $\frac{\lambda}{k\mu} < 1$ this result will be incorrect for not so small D.■

In other words, sometimes it is not optimal to use a partial information policy, based on the queue length, which is somewhat counterintuitive.
Now consider information disclosure policy X^0 it means that we have unobservable queue. The effective arrival rate is λq^{**}. So, as we have equilibrium state, then we can find q^{**} from equation

$$W(q^{**})C = R \tag{22}$$

recall that $C = 1$ and that waiting time for unobservable M/M/k equals

$$W(q^{**}) = \frac{C_k(A)}{k\mu - \lambda} + \frac{1}{\mu} \tag{23}$$

where $A = \frac{\lambda q^{**}}{\mu}$, $C_k(A)$ Erlang C-formula. So we get result

$$\frac{\frac{A^k}{k!}\frac{k}{(k-A)}}{\left(\sum_{n=0}^{k-1}\frac{A^n}{n!} + \frac{A^k}{k!)}\frac{k}{(k-A)}\right)(k\mu - \lambda)} + \frac{1}{\mu} = R \tag{24}$$

From this equation via numerical calculations, we get q^{**} when we have two servers and substitute this q^{**} in equation for π_0 for unobservable queue, which define as

$$\pi_0 = \left(\sum_{n=0}^{k-1}\frac{\frac{\lambda q}{\mu}^n}{n!} + \sum_{n=k}^{\infty}\frac{\frac{\lambda q}{\mu}^n}{k!k^{n-k}}\right)^{-1} \tag{25}$$

Simhon [12] considered the case when the provider has one server, now, we consider scenario when the provider has two servers. It follows from the fact that calculations for common scenario, when any $k > 0$ and $k \in Z$ very complicated, therefore we restrict ourselves to the case which is a particular case when $k = 2$

Theorem 2.3. Considering the set of deterministic threshold based information disclosure policies. To maximize its revenue, a service provider must use the full conceal information disclosure policy X^0 if we have under-loaded system, it means that $\frac{\lambda}{k\mu} < 1$

Proof. We should prove that

$$\pi_0(X^0) = \Big(\sum_{n=0}^{k-1} \frac{\frac{\lambda q}{\mu}^n}{n!} + \sum_{n=k}^{\infty} \frac{\frac{\lambda q}{\mu}^n}{k!k^{n-k}}\Big)^{-1}$$

less that

$$\pi_0(X^D) = \Big(\sum_{n=0}^{k-1} \frac{\frac{\lambda q}{\mu}^n}{n!} + \sum_{n=k}^{D+1} \frac{\frac{\lambda q}{\mu}^n}{k!k^{n-k}} + \sum_{n=D+2}^{\infty} \frac{\frac{\lambda q}{\mu}^n q^{*n-D-1}}{k!k^{n-k}}\Big)^{-1}$$

when $\frac{\lambda}{k\mu} < 1$.

we have that $k = 2$ and using geometric progression, we get following results

$$\pi_0(X^0) = \Big(\sum_{n=0}^{1} \frac{\frac{\lambda q}{\mu}^n}{n!} + \sum_{n=2}^{\infty} \frac{\frac{\lambda q}{\mu}^n}{2!2^{n-2}}\Big)^{-1} =$$

$$= \Big(\frac{2}{1-\frac{\lambda q}{2\mu}} - 1\Big)^{-1} =$$

$$= \Big(\frac{4\mu}{2\mu - \lambda q} - 1\Big)^{-1}$$

and

$$\pi_0(X^D) = \Big(\sum_{n=0}^{k-1} \frac{\frac{\lambda q}{\mu}^n}{n!} + \sum_{n=k}^{D+1} \frac{\frac{\lambda q}{\mu}^n}{k!k^{n-k}} + \sum_{n=D+2}^{\infty} \frac{\frac{\lambda q}{\mu}^n q^{*n-D-1}}{k!k^{n-k}}\Big)^{-1} =$$

$$= \Big(-1 + \frac{2(1-\frac{\lambda}{2\mu})^{D+1}}{1-\frac{\lambda}{2\mu}} + \frac{2q^{-D-1}\frac{\lambda q}{2\mu}^{D+2}}{1-\frac{\lambda q}{2\mu}}\Big)^{-1} =$$

$$= \Big(-1 + \frac{4\mu(1-\frac{\lambda}{2\mu})^{D+1}}{2\mu-\lambda} + \frac{\lambda^{D+2}q}{(2\mu-\lambda q)2\mu^{D+1}}\Big)^{-1}$$

note, that when $D \to \infty$ then last term of $\pi_0(X^D)$ equals zero, thus we should compare two values

$$\frac{4\mu}{2\mu - \lambda q}$$

and

$$\frac{4\mu(1-\frac{\lambda}{2\mu})^{D+1}}{2\mu-\lambda}$$

as $q < 1$ and $\frac{\lambda}{2\mu} < 1$ then $4\mu > 4\mu(1-\frac{\lambda}{2\mu})^{D+1}$ and $2\mu - \lambda q < 2\mu - \lambda$, thus we have that

$$\frac{4\mu}{2\mu - \lambda q} > \frac{4\mu(1-\frac{\lambda}{2\mu})^{D+1}}{2\mu-\lambda}$$

It means that

$$\pi_0(X^0) < \pi_0(X^D)$$

■

So as a result we get that, X^L will not be optimal under the setting of models, because any client who does not monitor the queue length can infer that the current length of queue must be greater or equal to L. These clients will refuse to join the queue and the provider thus lose profit.

2.3 Comparing X^∞ and X^D policies in underloaded and overloaded systems

Consider examples and compare $\pi_0(X^D)$ and $\pi_0(X^\infty)$

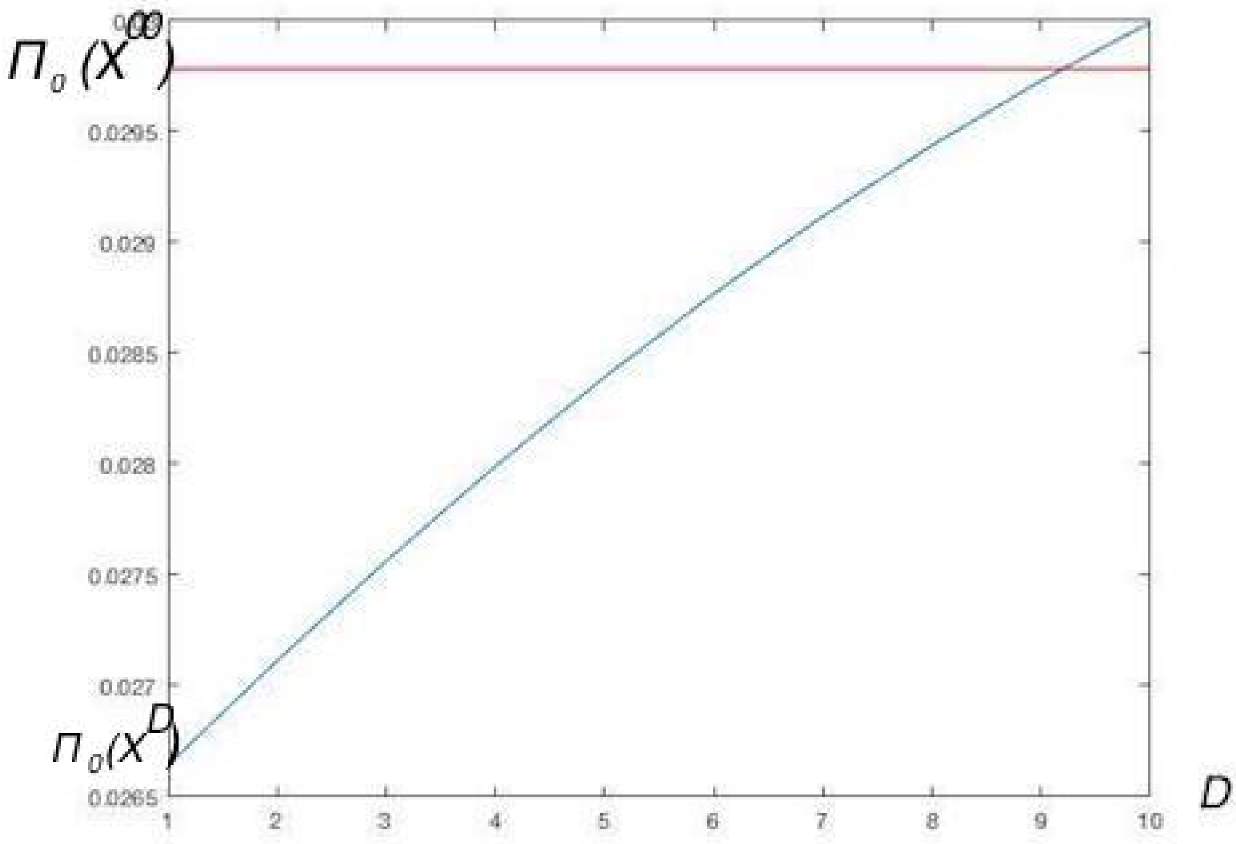

Figure 2.2: $\lambda = 9, 8$, $\mu = 5$, $k = 2$ and$R = 2$.

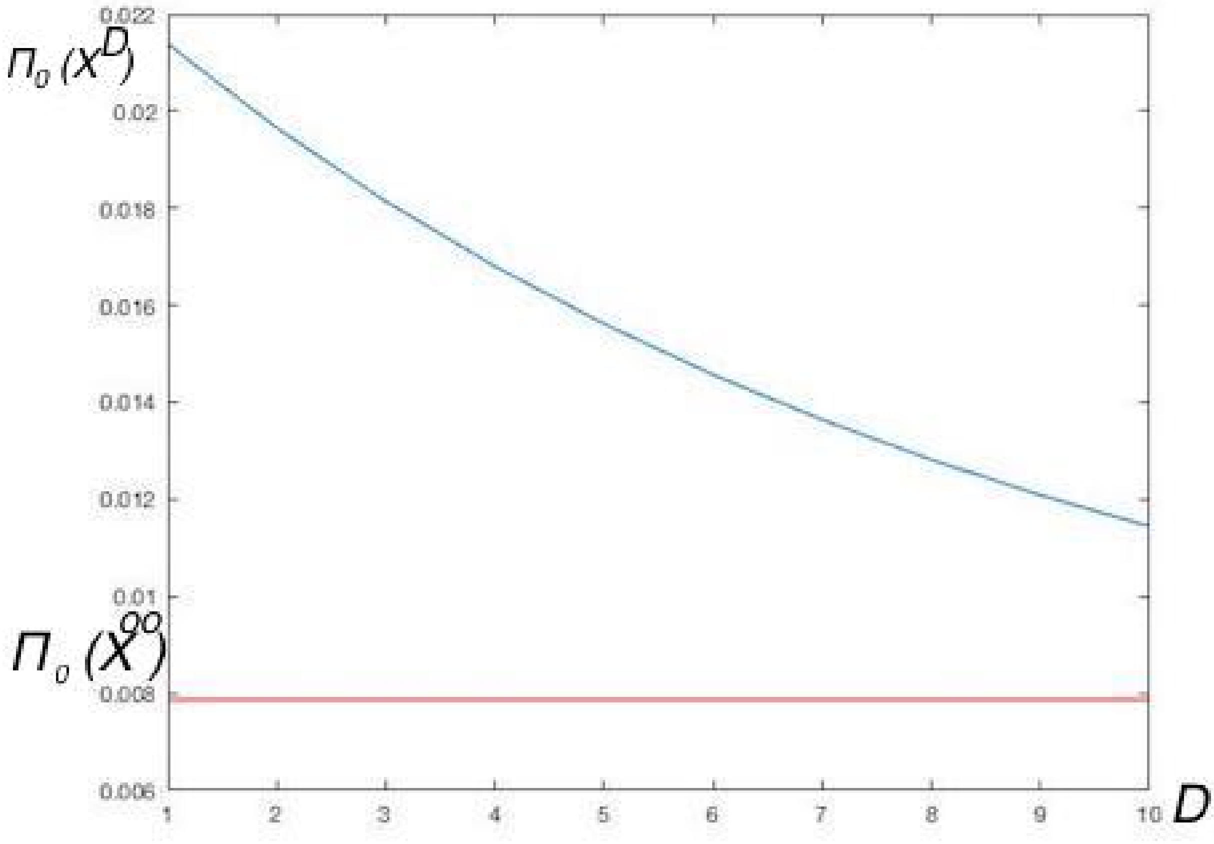

Figure 2.3: $\lambda = 11$, $\mu = 5$, $k = 2$ and$R = 2$.

Fig.2.2. shows the stationary idle probability when different threshold values are used in an underloaded queue with $\lambda = 9, 8$, $\mu = 5$, $k = 2$ and $R = 2$.In this case $L = 20$. Thus, we consider threshold policies from 0 to 10. We can see from the figure that the idle stationary probability of any threshold policy is always greater than X^0. For some threshold values, it is also greater than X^∞.

In Fig.2.3. we consider an overloaded queue with $\lambda = 11$, $\mu = 5$, $k = 2$ and $R = 2$. In this case, the idle stationary probability of any threshold policy from 0 to 10 is bounded between X^0 and X^∞. This time, the policy that minimizes the idle probability is to always inform customers about the queue length, i.e. X^∞.

2.4 Comparing X^0 and X^D policies in underloaded and overloaded systems

So we have two servers, it means that $k = 2$, let reward will be $R = 2$, and $\lambda = 9, 8$, $\mu = 5$, calculate for this case $\pi_0(D)$ and $\pi_0(X^0)$

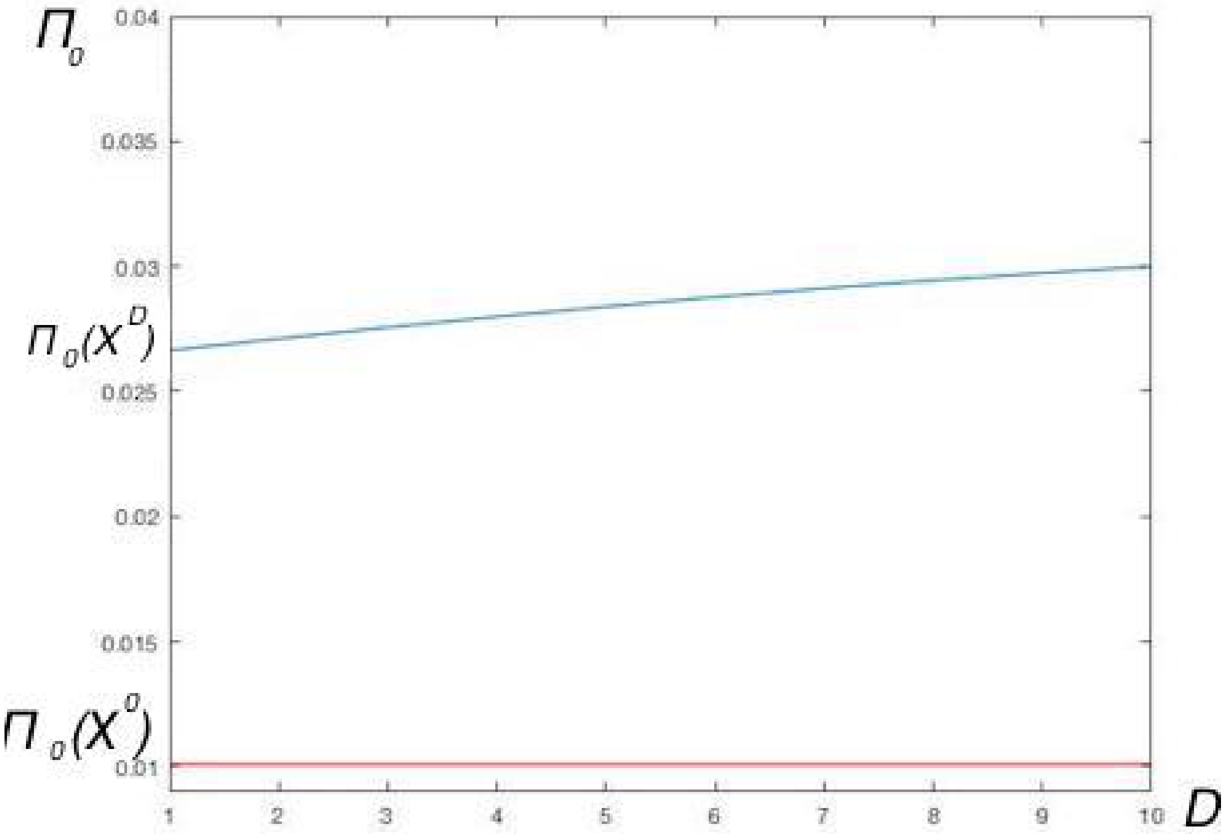

Figure 2.4: $\lambda = 9, 8$, $\mu = 5$, $k = 2$ and$R = 2$.

As we see on Fig.2.4. $\pi_0(X^0)$ provide minimum of idle stationary probability, at the same time $\pi_0(D)$ increase with increasing D. Consider another case, when $\frac{\lambda}{k\mu} > 1$, let $\lambda = 11$, $\mu = 5$, $k = 2$ and $R = 2$

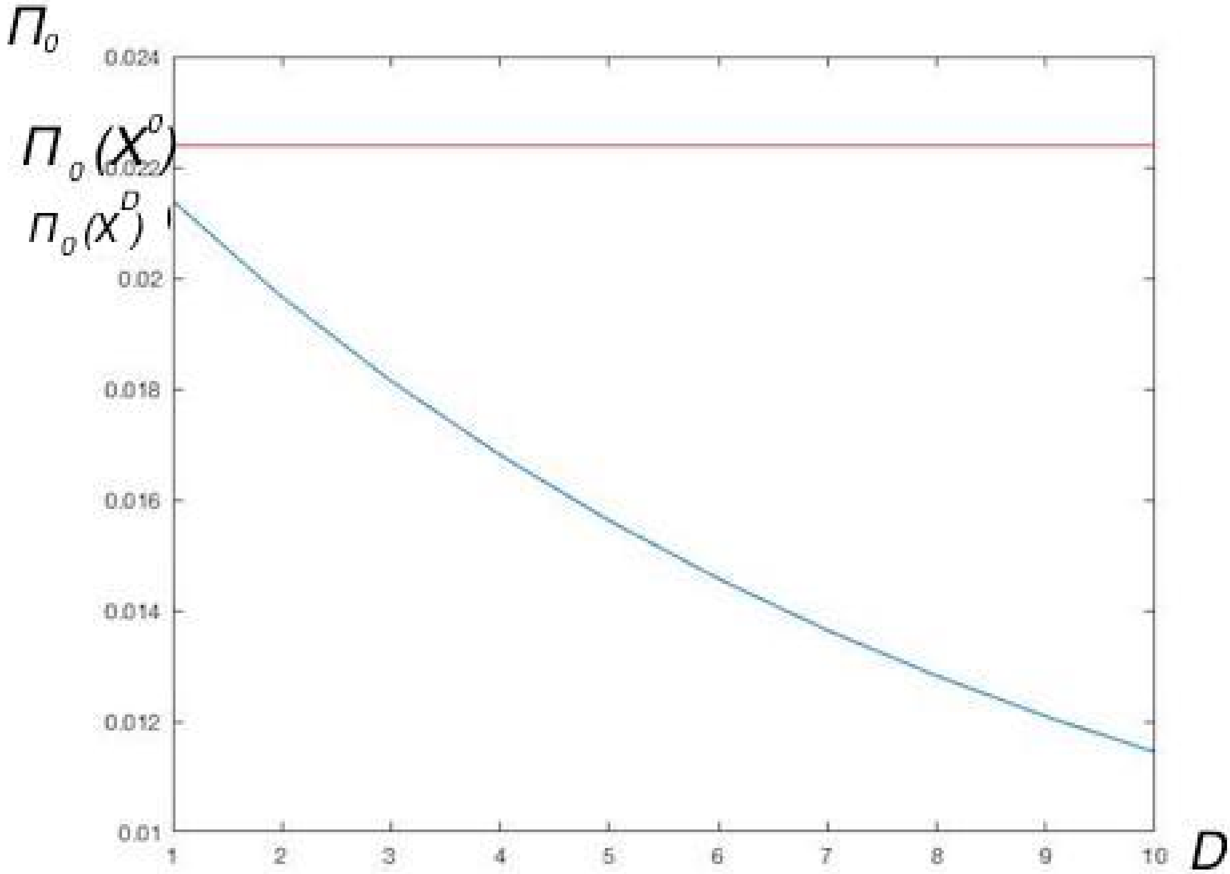

Figure 2.5: $\lambda = 11$, $\mu = 5$, $k = 2$ and$R = 2$.

On Fig.2.5. we see that more optimal for this case information disclosure policy X^D

3 Optimal information disclosure policies of provider for inexperienced clients

3.1 Maximizing revenue information disclosure policy

Now, consider another case, when we have inexperienced clients, it means that once the provider begins to conceal information, clients will not be aware about threshold D, so after the provider will hide information about queue length, the queue will become unobservable at all. Recall that we have a multiple-server queue (M/M/k). Potential clients arrive to the system according to a Poisson process with rate λ. Joining customers are served FIFO discipline. Service times of the customers are independent and exponentially distributed with rate μ. All joining customers incur a linear waiting cost of C per unit of time while they standing in the system. Without loss generality, consider waiting cost C equals 1. And after service, client get reward R.

Client, which does not be aware about queue length joins to the queue with probability q and balks the queue with probability $1-q$. But this probability q will differ from probability, which we found in previous chapter, because now, clients do not know about threshold D.

Under this conditions, we obtain the following result

Theorem 3.1. X^L is a optimal information disclosure policy in the system with inexperienced clients among set of information disclosure policies $(X^0, X^1, ..., X^L, X^\infty)$.

Proof. Consider any arbitrary policy $X = (x_0, x_1, ..., x_i, ...)$. When $i < L$, the effective arrival rate at state i of the underlying queue is λ if $x_i = 1$, and λq if $x_i = 0$. So the effective arrival rate at any state $i \in (0, 1, ..., L-1)$ can be written as

$$\lambda_i = \lambda((1-q)x_i + q)$$

Define $l = min(i|i \geq L, x_i = 1)$. Then no state higher than k will be ever reached. Moreover, the effective arrival rate at states $(L, L+1, ..., l-1)$ is λq. It follows that the effective arrival rate, denoted by λ_{eff} is introduced by

$$\lambda_{eff} = \begin{cases} \lambda_i & i < L, \\ \lambda q & L \leq i < l. \end{cases} \tag{26}$$

Given that the transition rate form state i to state $i-1$ is μ when $i < k$ and $k\mu$ in other cases. So calculate the steady-state probabilities $(\pi_i(X))$

$$\begin{cases} \frac{\lambda}{\mu}^n \frac{\prod_{i=0}^{n}((1-q)x_i+q)}{n!} & 0 \leq n \leq k-1, \\ \frac{\lambda}{\mu}^n \frac{\prod_{i=0}^{n}((1-q)x_i+q)}{k!k^{n-k}} & k \leq n \leq L, \\ \frac{\lambda}{\mu}^n \frac{q^{n-D-1}}{k!k^{n-k}} & L < n \leq l. \end{cases} \tag{27}$$

From $\sum_{i=0}^{l} \pi_i(X) = 1$, we have

$$\pi_0(X) + \sum_{n=1}^{k-1} \pi_i(X) + \sum_{n=k}^{L} \pi_i(X) + \sum_{n=L+1}^{l} \pi_i(X) = 1$$

then we get

$$\pi_0(X) = \Big(\sum_{n=0}^{k-1} \frac{\lambda}{\mu}^n \frac{\prod_{i=0}^{n}((1-q)x_i+q)}{n!} + \sum_{n=k}^{L} \frac{\lambda}{\mu}^n \frac{\prod_{i=0}^{n}((1-q)x_i+q)}{k!k^{n-k}} + \tag{28}$$
$$+ \sum_{n=L+1}^{l} \frac{\lambda}{\mu}^n \frac{q^{n-D-1}}{k!k^{n-k}} \Big)^{-1}$$

Consider two cases.

Let $R = CW$, where C costs of staying, and W is waiting time in the system, then our $q < 1$ and therefore the $\pi_0(X)$ reaches maximum when

$x_i = 1$ $i \in (0, ..., L-1)$ and when $l \to \infty$. Thus X^L corresponds to the optimal policy.■

Now, consider that $R > CW$, then $q = 1$ and then the maximum value of $\pi_0(X)$ is obtained as long as $l \to \infty$, which means $x_i = 0$ when $i \geq L$. As x_i can be either 0 or 1 when $i < L$ there are 2^L many different optimal policies although all of them lead to the same queuing dynamics. So, X^L is one of the optimal policies.

Theorem 2.1 discloses that to maximize revenue, the server should reveal the queue length information to the customers when the queue is shorter than Naor's balking threshold L, but conceal the queue length when the queue is longer than L. Define this policy as X^L It seems obvious that X^L becomes the optimal information disclosure policy for the provider. When the current queue length is less than the threshold L, the provider reveal the queue length to make sure that any arriving client will join the queue. On the other hand, when the current queue length is L or greater, any arriving client will not join the queue if they knew the true about queue length, therefore it is in the provider's best interest to disclose the information.

3.2 Comparing information disclosure policies in overloaded and underloaded systems

Consider first case when we have system, where reward $R = 5$, cost per unit of time $C = 1$, arrival rate $\lambda = 5$, service rate $\mu = 2$ and provider has two servers, it means $k = 2$ and then $L = 20$, so compare policies

Policy	Idle Probability
X^0	0.3
X^∞	0.0243
X^{20}	**0.0032**
X^{10}	0.0345
X^{odd}	0.0391
X^{even}	0.0267
X^5	0.11

As you see that we have that X^L in our case X^{20} give the least value of idle stationary probability

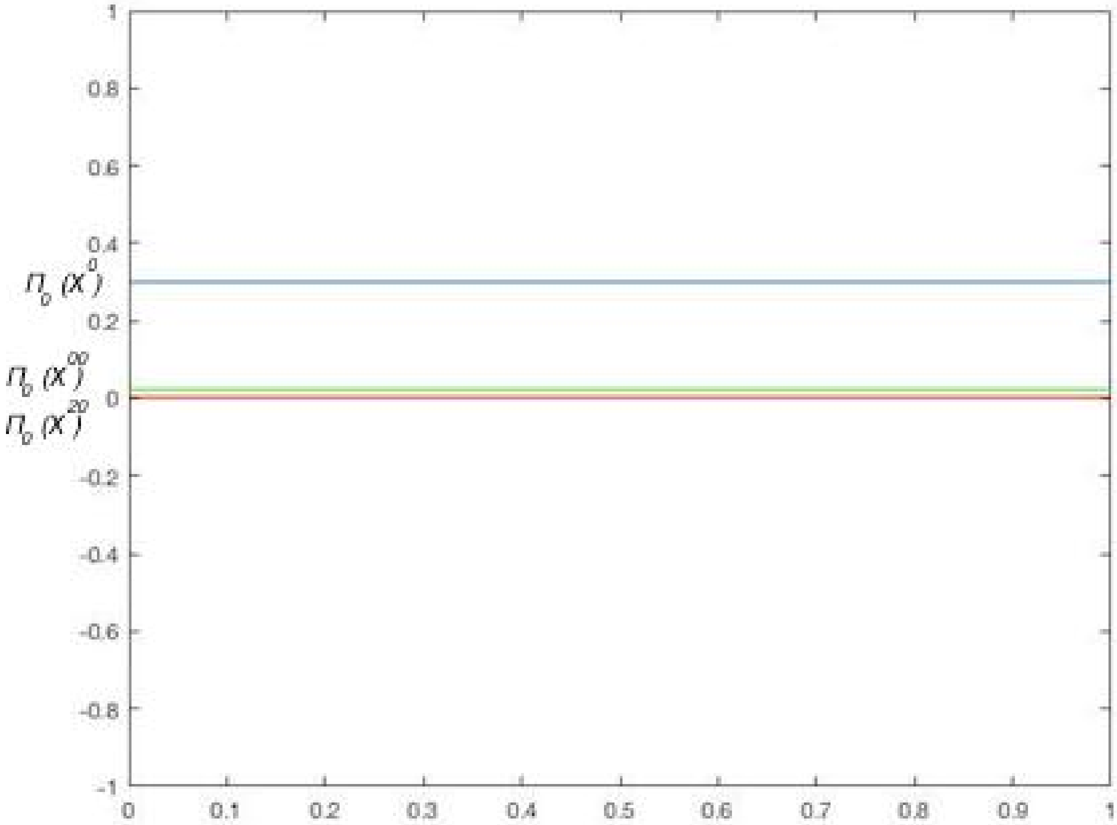

Figure 3.1: $\lambda = 5$, $\mu = 2$, $k = 2$ and$R = 5$, $C = 1$.

On Fig.3.1. presented following information disclosure policies X^0, X^∞ and X^{20}, so X^{20} provided by minimum of idle probability.

Now consider case when we have system, where reward $R = 6$, cost per unit of time $C = 1$, arrival rate $\lambda = 3$, service rate $\mu = 2$ and provider has two servers, it means $k = 2$ and then $L = 24$, so compare policies

Policy	Idle Probability
X^0	0.266
X^∞	0.321
X^{24}	**0.213**
X^{20}	0.345
X^{odd}	0.287
X^{even}	0.315
X^{10}	0.305

As you see that we have that X^L and in this case(when we have underloaded system) X^{24} give the least value of idle stationary probability

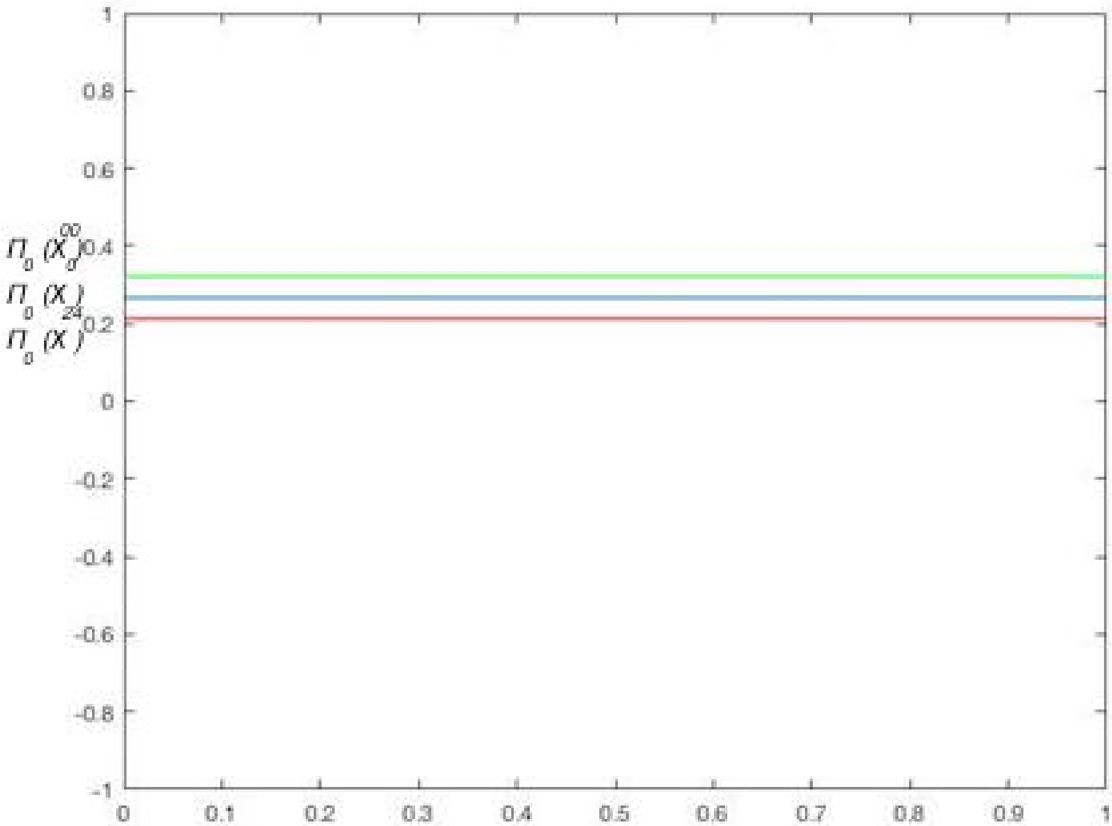

Figure 3.2: $\lambda = 3$, $\mu = 2$, $k = 2$ and$R = 6$, $C = 1$.

On Fig.3.2. we can see comparing of following information disclosure policies X^0, X^∞ and X^{24}, so X^{24} provided by minimum of idle probability.

Also we can noticed, that when we have overloaded system then all idle stationary probabilities more less then idle probabilities in underloaded system, this fact is intuitively understandable.

4 Conclusions

We consider a multiple-server queue (M/M/k) and found optimal information disclosure policies in two cases, when the system with partial observability queue, recall that it means that clients are aware about value of threshold, and when the system with unonservability queue.

We proved that in partial observability system, another words this system has experienced clients, we have a unique equilibrium(see Lemma 2.1.), it means that in game between clients exists a unique equilibrium. And proved that X^{∞} is optimal information disclosure policy(see Theorem 2.2) in overloaded system when the provider two servers, for one servers these results were found by Simhon [12], and X^{0} is optimal information disclosure policy in underloaded system, when provider has two servers(see Theorem 2.3). After that we proved that X^{L} is optimal information disclosure policy in the system with inexpirenced clients(see Theorem 3.1).

This results shows, that if clients visit very often the provider, then X^{L} is not optimal information disclosure policy, which is intuitively clear, and if we have system where clients do not often visit the provider, or each time we have new clients, then X^{L} is optimal information disclosure policy, and it is also reasonable.

For finding optimal information policy we use minimize of idle stationary probability, but here exists another case, when we should maximize utilization of the provider, so in the next researches we consider optimal information disclosure policy via maximizing utilization of the provider.

References

[1] Naor, Pinhas. 1969. The regulation of queue size by levying tolls. Econometrica: Journal of the Econometric Society 15–24.

[2] Edelson, Noel M, David K Hilderbrand. 1975. Congestion tolls for poisson queuing processes. Econometrica: Journal of the Econometric Society 81–92.

[3] Larsen, Christian. 1998. Investigating sensitivity and the impact of information on pricing decisions in an M/M/1/∞ queueing model. International journal of production economics 56 365–377.

[4] Afe'che, Philipp, Haim Mendelson. 2004. Pricing and priority auctions in queueing systems with a generalized delay cost structure. Management Science 50(7) 869–882.

[5] Hassin, Refael, Moshe Haviv. 2003. To queue or not to queue: Equilibrium behavior in queueing systems, vol. 59. Springer Science Business Media.

[6] Hassin, Refael. 2016. Rational Queueing. CRC Press.

[7] Hassin, Refael. 1986. Consumer information in markets with random product quality: The case of queues and balking. Econometrica: Journal of the Econometric Society 1185–1195.

[8] Chen, Hong, Murray Frank. 2004. Monopoly pricing when customers queue. IIE Transactions 36(6) 569–581.

[9] Shone, Rob, Vincent A Knight, Janet E Williams. 2013. Comparisons between observable and unobservable M/M/1 queues with respect to optimal customer behavior. European Journal of Operational Research 227(1) 133–141.

[10] Hassin, Refael, Ricky Roet-Green. 2017a. the impact of inspection cost on equilibrium, revenue, and social welfare in a single-server queue. Operations Research, forthcoming .

[11] Hu, Ming, Yang Li, Jianfu Wang. 2017. Efficient ignorance: Information heterogeneity in a queue. Manage- ment Science, forthcoming .

[12] Simhon, Eran, Yezekael Hayel, David Starobinski, Quanyan Zhu. 2016. Optimal information disclosure policies in strategic queueing games. Operations Research Letters 44(1) 109–113.

[13] Shiliang, Cui, Kaili Li, Jinting, Wang. 2016. Optimal disclosure of queue length information. Operations Research Letters 46(2) 100–117.

[14] Guo, Pengfei, Paul Zipkin. 2007. Analysis and comparison of queues with different levels of delay informa- tion. Management Science 53(6) 962–970.

[15] Economou, Antonis, Spyridoula Kanta. 2008. Optimal balking strategies and pricing for the single server markovian queue with compartmented waiting space. Queueing Systems 59(3-4) 237–269.

[16] Guo, Pengfei, Paul Zipkin. 2009. The effects of the availability of waiting-time information on a balking queue. European Journal of Operational Research 198(1) 199–209.

[17] Armony, Mor, Nahum Shimkin, Ward Whitt. 2009. The impact of delay announcements in many-server queues with abandonment. Operations Research 57(1) 66–81.

[18] Allon, G., A. Bassamboo, I. Gurvich. 2011. We will be right with you: Managing customer expectations with vague promises and cheap talk. Operations Research 59(6) 1382–1394.

[19] Cui, Shiliang, Senthil Veeraraghavan. 2016. Blind queues: The impact of consumer beliefs on revenues and congestion. Management Science 62(12) 3656–3672.

[20] Kremer, Mirko, Laurens Debo. 2015. Inferring quality from wait time. Management Science 62(10) 3023– 3038.

[21] Yu, Qiuping, Gad Allon, Achal Bassamboo, Seyed Iravani. 2017. Managing customer expectations and priorities in service systems. Management Science, forthcoming .

Appendix 1

```
l=11;
m=5;
k=2;
t=k-1;
R=2;
syms n
y=[];

z=pi0limit(l,m,k,R);
for D=1:10
y=[y pi0(D,l,m,k,R)];
end
y
z=y(1)+0.001
D=1:10
figure
plot(D, y);
hold on
plot(xlim, [z z], 'r')
function [outputArg1] = pi0(inputArg1, inputArg2, inputArg3,
inputArg4, inputArg5)
D=inputArg1;
l=inputArg2;
m=inputArg3;
k=inputArg4;
t=k-1;
R=inputArg5;
syms n
outputArg1 =(symsum(((l/m)^n)/(factorial(n)), n, 0, t)+symsum((((l/m)
/(factorial(k)*k^(n-k))), n, (D+2), Inf))^(-1);
```

```
end

function [outputArg1] = pi0limit(inputArg2, inputArg3,
inputArg4, inputArg5)
l=inputArg2;
m=inputArg3;
k=inputArg4;
t=k-1;
R=inputArg5;
syms n;
L=floor((1+m*(R-(1/m)))*k);
outputArg1 =(symsum(((l/m)^n)/(factorial(n)), n, 0, t)
+ symsum((((l/m)^n)/(factorial(k)*k^(n-k))), n, k, L))^(-1);
end

l=9.8;
m=5;
k=2;
R=2;
z=pi0unobservable(l, m, k, R)
y=[];
for D=1:10
y=[y pi0(D,l,m,k,R)];
end
D=1:10
figure
plot(D, y);
axis([1 10 0.009 0.04])
hold on
plot(xlim, [z z], 'r')

function [outputArg1] = pi0unobservable(inputArg2,
inputArg3, inputArg4, inputArg5)
l=inputArg2;
```

```
m=inputArg3;
k=inputArg4;
R=inputArg5;
t=k-1
syms q n
eq=(((((((l*q)/m)^k)*k)/(factorial(k)*(k-(((l*q)/m)^k))))
/(((((l*q)/m)^k)*k)/
 (factorial(k)*(k-(((l*q)/m)^k))) +
 symsum(((l/m)^n)
 /(factorial(n)), n, 0, k-1)))/((m*k)-(l*q)))+(1/m)==R;

Q=solve(eq, q)
Q=Q(1)

outputArg1=(symsum((((l*Q)/m)^n)/(factorial(n)), n, 0, t)
+ symsum(((((l*Q)/m)^n)/(factorial(k)*k^(n-k))), n, k, Inf))^(-1);

end

syms l m R q
eq=((((l*q)^2)/(m*(m*2-l*q)))/((1+((l*q)/(m)))*(2*m-l*q)+
(((l*q)^2)/m)))+(1/m)-R==0;
Q=solve(eq, q)
Q=Q(2)
```