An Invitation Control Policy for Proactive Service Systems: Balancing Efficiency, Value and Service Level

Yueming Xie

An Invitation Control Policy for Proactive Service Systems: Balancing Efficiency, Value and Service Level

Research Thesis

In partial Fulfillment of The Requirements for the Degree of Master of Science in Industrial Engineering

Yueming Xie

Submitted to the Senate of
the Technion – Israel Institute of Technology

Adar 5777 Haifa March 2017

This research was carried out under the supervision of Dr. Galit Yom-Tov and Prof. Liron Yedidsion in the Faculty of Industrial Engineering and Management.

The generous financial help of The Technion Graduate School is gratefully acknowledged.

Thanks My Dear Family, Advisors and Homeland China.

Contents

List of Figures

List of Tables

Abstract

Proactive service systems permit a controllable arrival rate managed by the service provider, which is different from classic service systems. Conceptually, some (or all) of the customers are invited to the system, so as to allow for a better control over operational indicators and profitability. Such a proactive service system is used, for example, to model an online chat service system, or for planning preventive care strategies for health care service providers.

Through an empirical study of a proactive chat service system, the validity of customer ranking information is elaborated for optimizing invitation control. It is also shown that service level measures can be formulated in terms of penalty for abandonment and cost of waiting. Hence, an infinite-time-horizon multiclass multiserver queueing system has been developed with impatient customers. We find an asymptotically optimal policy using a fluid approximation, by solving a linear programming problem that maximizes revenues. The asymptotic optimal invitation policy we developed invites customers by their $r\mu$ ranking in decreasing order until there are no idle servers. Then, an equivalent threshold policy is proposed that is easy to implement in practice. Numerical simulations were performed to demonstrate the performance of the policy and identify its limitations. We show that the fluid policy has a good performance but is also crude.

In order to refine the fluid policy, we analyzed a fluid approximation of the system under a more flexible threshold policy. The equilibrium is found to strongly depend on system parameters. In particular, it depends on the threshold value. It is also shown that the equilibrium is globally asymptotically stable via trajectory and Lyapunov analysis. Furthermore, in order to propose an invitation policy for proactive service systems that balances revenues and service level, the probability of implementing admission control is approximated, and several approximations of performance metrics are calculated. Simulations are performed to examine the performance of these approximations. All of them perform well especially in large-size systems.

Abbreviations and Notations

$\wedge$	:	Minimal Computation Signal
$\vee$	:	Maximal Computation Signal
$()^+$	:	The Larger Value Between the Result Inside Brackets and Zero
MDP	:	Markov Decision Process
DCP	:	Diffusion Control Problem
QED	:	Quality and Efficiency-Driven
ICT	:	Information Collection Time
LP	:	Linear Program
CSC	:	Customer Service Chat
IP	:	In-Process
ODE	:	Ordinary Differential Equation
ROC	:	Receiver Operating Characteristic
NA	:	Non-Admissible

Chapter 1

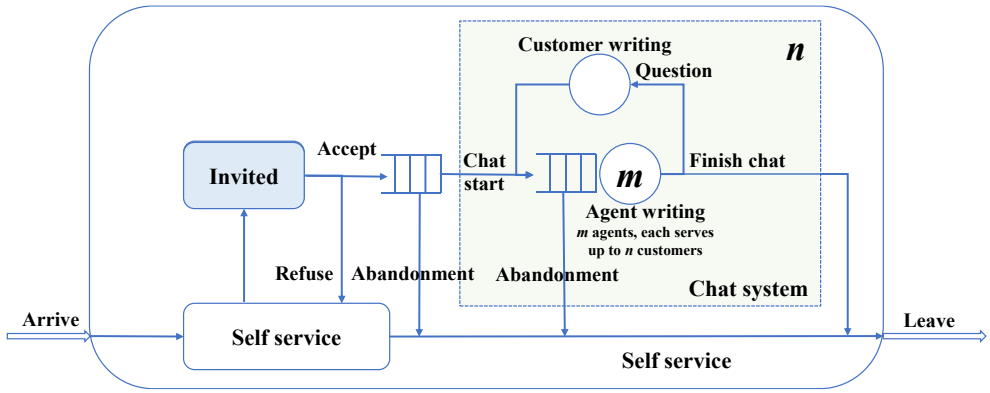
Introduction

Classic service models mostly consider cases in which the customers autonomously seek service from companies. Thus, the arrival rate is highly variable and exogenous to the system. Decision makers then make strategic decisions on how to cope with that stream of arrivals, regard service level and decide whether to serve all customers and what resource to assign. In contrast, new technology allows companies to control arrivals. We refer to service systems that can do so as proactive service systems. The new technology we refer to provides service system access to historical information regarding potential customers prior to their arrival to the system, for example, through their surf information on the internet. The companies use such information to classify potential customers, assess their current value, and *invite* them to the system for personalized assistance. In this type of system, the company has sound information indicating that the invited customer is likely to require or benefit from service. The agents are able to access the profile and data of the current customers they are serving, which helps them provide a meaningful interaction with their customers, so as to both promote the revenue and to improve the customer experiences. Proactive service systems are becoming more and more common. We give three examples from internet–based contact centers, law–enforcement systems, and healthcare systems.

Our first example is an internet–based proactive chat system. Many banks and retail companies encourage the use of internet or mobile platforms for providing self services. In addition, these companies usually provide other service channels by which a customer can reach them — either by phone, chat or mail. Such channels are required in order to solve problems or to complement the self service. Such a combination is beneficial both to companies and customers, as self services have the benefit of flexible timing, visualization and low cost, while a personal connection through other service channels is sometimes needed to solve more complicated problems or to enhance customer experience. The decision of how to combine correctly the platforms and when to move from one to another is an important strategic decision. These days, many companies adopt an online customer service chat (CSC) system as an attractive complement for the online self services, for its economy and immediacy. The chat can be initiated by the

customer, for example, by pressing a 'contact us' button or by the company that extends an invitation for a chat on the customer's screen. We concentrate on the latter. In order to decide to which customers to offer service, the company collects the customers' browsing behavior on their website and additional historical data. Then, the company evaluates that consumer's 'service value'; for example, if the end-user seems to have a problem we might infer that he can benefit from personal help greatly, while if he is 'doing fine', no chat is needed. The high-value visitors may then be invited to the chat based on the current service availability. From the customer's perspective, once an invitation was offered, during browsing the website, they are free to accept the invitation or alternatively decline it anytime before leaving the website. Note that after accepting the invitation customers enter a queue. They may abandon that queue at any time. The service itself is composed of interactions between the agent and customers. The number of interactions vary. An interaction can include for example a customer question and an agent answer. While the customer types the question, the agent is waiting. Hence it is customary that each agent, in such contact centers, manages multiple customers simultaneously. This might create a second in-service queue, in which customers wait for answers and may abandon if it is too long. The dynamics of such a system is described in Figure 1.1. The company needs to decide how many customers to invite. Too many customers can lower the service level (waiting and abandonment) and increase the staffing cost; on the other hand, low capacity is also unwelcome since the company may miss valuable customers. Hence, when discussing the invitation strategies, decision makers need to balance customer value, costs and service levels.

Figure 1.1: Proactive chat system description

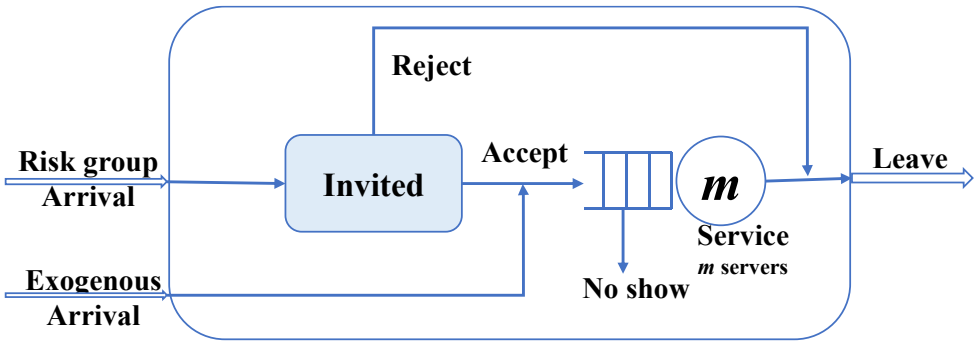


Our second example is from law-enforcement systems in Israel, specifically a traffic speed-control system. Many countries use a camera-based automated system to enforce speed limit laws. In such countries, the authorities decide on the minimal speed they want to enforce (which may be and usually is higher than the maximal speed allowed by law). One of the considerations when deciding on enforcing a speed limit is the load on the enforcing system, as some of the drivers who were caught passing the speeding limit may want to go to court and not just pay a fine. Hence, the lower the enforcing

limit, the higher the load on the court system. The speed limits can be considered as a proactive control policy to regulate the system's load. Each ticket is viewed as an invitation to engage a trail. The customer may decide to decline the invitation by not appealing for a trial and simply paying the fine ticket. Unlike the chat system, once appealing, one cannot abandon the system until the trial is over.

Our last example is a proactive healthcare system in which patients are invited for a periodic or a followup medical examination. Such preventive care policies aim to identify health problems before they become severe. Screening all people is wasteful. Therefore, the decision makers invite only an appropriate number of high–risk patients for the preventive checkup. This will both reduce the cost of further treatment, and improve health. However, even though the methodologies for evaluating which patient is more likely to need such services are improving, operational and economical considerations limit the implementation of such preventive care policies. One such limiting factor is the number of physicians who can do that checkup. A certain capacity should be kept available, at all times, for the regular patients who come with unexpected health issues. The dynamics of such a system is described in Figure 1.2. The challenge is to balance the two groups properly and plan the invited patients in a manner that will not overload the system but will take into account their medical risk properly.

Figure 1.2: Proactive health care system description



This research explores invitation policies for a proactive service system that balance revenue and service levels.

1.1 Literature Review

1.1.1 Admission Control

An invitation problem can also be considered as an admission control problem, because it can be interpreted as whether to accept or reject each potential arrival. There are many purposes that discuss admission control problems; we focus on those that consider also service levels (e.g. abandonment). Koole and Pot (2011), motivated by an inbound call center, discussed the admission control problem to maximize profit of an $M/M/s/n+M$ queueing system by controlling its trunk and agent number. They

assumed Poisson arrival and exponential servers. Later on, Ward and Kumar (2008) extended it into the general distribution arrival and service rates case in a conventional heavy-traffic regime. Koçağa and Ward (2010) tried to minimize the infinite horizon expected average cost associated with customer blocking, abandonments and server idleness of the Elrang-A queueing system. They used the Markov decision process (MDP) to show the optimal admission control policy in a threshold form. An efficient iterative algorithm was developed under certain constraints, which can guarantee the optimal solution to minimize the infinite horizon expected average cost. Then, by solving the diffusion control problem (DCP) in a Quality and Efficiency-Driven (QED) regime, an asymptotically optimal policy is obtained. Weerasinghe and Mandelbaum (2008) considered the $G/M/n/B + GI$ queueing system for the finite horizon cost minimization under QED regime. They developed a static control policy, in the form of a constraint on the system capacity. They showed by using DCP analysis that the solution asymptotically minimizes the cost that trades off blocking and abandonment over a finite time horizon. All the above papers assume blocked customers to have the same value while we assume otherwise.

There are several studies that discuss a control policy which not only controls admission but also some other operational parameters. Such hybrid control mechanisms (especially joint admission and service rate control) had been considered by Ata and Shneorson (2006) (adjustable arrival and service rates for the $M/M/1$ system), Ghosh and Weerasinghe (2007) (queue capacity and service rate control for the $M/M/n$ system), Ghosh and Weerasinghe (2010) (extend the system in Ghosh and Weerasinghe (2007) with impatient customers) and Lee and Kulkarni (2014) (controllable arrival and service rates for the $M/M/n$ system). Chan and Yom-Tov (2015) studied the admission control problem of a multi-server queueing system which allows speedup. They used dynamic programming to prove that a threshold policy is optimal. They first analyzed the system equilibrium under the fluid level and then used it to deduce the parameters in order to approximate the system into an existing stochastic model. Furthermore, a heuristic algorithm is explored to determine thresholds for admission control and speedup. Especially, they assumed a concave cost function, which means that each blocked customer may have a different value or impact. We will take a similar approach in our analysis but would like to maximize revenue instead of minimize costs. In addition,, they do not consider abandonment which is an important feature in our model.

Early work on the admission control in a nonidentical customer system can be found in Miller (1969), where an optimal threshold policy for a multi-server loss system was explored. In such a system, new arrivals will balk without entering the system if there are no free servers available. Hence, the system administrators would like to reject some arrivals in order to keep some strategic idleness for the higher value customers, so as to lift the total reward. Such a study develops by concerning different aspects of the system characters. For instance, some references tried to extend it into the non-stationary case

(Yoon and Lewis 2004), whereas some others mentioned the patience of each customer (Zayas-Cabán and Lewis 2016). Nevertheless, Zayas-Cabán and Lewis (2016) discussed the policy in a two-class loss system, in which abandonment happens both during queueing and service instead of while in queueing. This may happen in a health care system. Also, we are more interested in a service system with a queue, which has more general applications.

1.1.2 A Multi-class Customer Queueing System

Customer ranking is usually a typical feature of the proactive service system, thus yielding multi-class customer types. When we have the class information, it is meaningful to choose an appropriate queueing policy according to different aim, such as the cost/reward objective or service level requirement. The well-known $c\mu$ rule is a very important priority policy for a multi-class queueing system, in order to minimize system cost. This policy was proven optimal in both deterministic (Smith 1956) and stochastic (Pinedo 1983) environments to minimize linear cost criteria. Van Mieghem (1995) generalized this rule by using heavy traffic analysis to minimize more general nondecreasing convex cost structures. He proved that this policy is asymptotically optimal for minimizing cumulative delay cost.

Atar et al. (2004), Atar et al. (2010) and Atar et al. (2013) studied the multi-server system with several classes of impatient customers. Atar et al. (2010) investigated a linear program (LP) that leads to a lower bound on the long run average holding cost. Then, it was shown that a routing policy, which is referred to as the $c\mu/\theta$ rule, asymptotically attains the lower bound in both preemptive and non-preemptive cases. Both rules are independent of the arrival rates of the customers.

De Véricourt and Zhou (2005) extended the $c\mu$ rule to a multi-server system with return. Huang et al. (2015) studied this further in the context of emergency departments. In their research, the patients can either exogenously arrive or are in-process (IP). The performance measure, i.e., the cumulative costs, can be asymptotically minimized prioritizing new patients according to their triage score and IP patients by their progress.

Perry and Whitt (2011) also considered two-class arrivals, but in an overloaded X system, where each class of customers has its own queue and service pool, but service is on a first-come, first-serve basis. The private severs are only activated to help another class when an unexpected overload occurs. As a continuation of their work, Perry and Whitt (2009) proposed a threshold for the weighted queue-ratio to trigger the temporary routing for maintaining a certain queue ratio. One should notice that the authors focused on the fluid approximation of the system and developed the ordinary differential equation (ODE) of the system dynamic. Then, based on the fluid analysis, the steady-state queue lengths were approximated. Such a methodology is also used in Chan et al. (2014) and Chan and Yom-Tov (2015). In all cases, this approximation was proven effective.

In our research, we have to decide on the policy for invitation. We currently focus on the static ranking information for both decisions, namely, the queueing policy will follow the priority of the invitation policy. All the above papers discuss the routing policy for a multi-class system, whereas we are exploring its invitation policy. In other words, all customers within their service system are served through the $c\mu/\theta$-type policy. We concentrate on the decision of who we want to let into our system. Therefore, our study is more focused on maximizing the revenue of the system, which is a different objective than the references discussed.

1.1.3 Contact center

Proactive systems, in many cases, incorporate endogenous arrivals into the system so as to raise system efficiency. For example, in the contact center, the organization sometimes initiates service, termed 'outbound' calls. The balance between endogenous and exogenous customers was investigated previously. In order to achieve server efficiency, namely, reducing idleness, decision makers prefer to initiate new calls by an automatic dialer system even when all agents are occupied with other calls (Sarraf 1989). However, the consequent abandonment is not welcomed either customers or decision makers. Samuelson (1999) used queueing theory to maximize the number of dialing under an abandonment proportion constraint. Pang and Perry (2014) presented a logarithmic safe staffing policy for a large pool of agents who provide service to inbound and outbound calls, namely, blending call service. However, this system gives priority to the inbound calls; in addition, outbound calls are immediately lost if there is no agent available. That setting is significantly different from ours. As in all the examples discussed in the introduction there is no strict priority between in/out services. If the out-services are more valuable, then we might prioritize them instead.

A different type of call blending is balancing cross-selling opportunities. By cross-selling, we mean that during a customer initiated call, the agent proposes extra services the customer did not ask for. Hence, they increase the call length, but also take advantage of the customer's availability. Conceptually, the tradeoffs between whether or not to offer cross-selling opportunities is similar to the ones we consider here. The difference is that in such a system customers do not need to wait in queue. Armony and Gurvich (2010) found an asymptotically optimal threshold to balance the staffing requirement and the cross-selling opportunities, so as to maximize the profit while meeting a certain service level.

Even though proactive service systems can be applied in many service environments, our main motivation and data comes from contact centers. The main applications we address is the Customer Service Chat (CSC) system. The emerging CSC system has some unique features, which were discussed in several papers. For example, chat systems have a lower operational cost than telephone service support systems (Andrews and Haworth 2002); they show better performance, including average speed to answer

and user satisfaction, and allow for multitasking (namely one agent can serve multiple customers) (Shae et al. 2007)

Tezcan and Zhang (2014) addressed the implication of simultaneous service which results in service rates that depend on the number of customers each agent is serving. The customers may be impatient when waiting for the service to begin or for each answer while in service. In order to minimize the staffing level under a certain service level goal, they used a routing problem LP to minimize the abandonment probability. Then, a closely-related staffing LP was formulated, for which the corresponding number of agents was proven asymptotically optimal. Note that the same structure is also used when considering the staffing of an emergency department, where the service is given in a discontinuous manner (Yom-Tov and Mandelbaum 2014, KC 2013). However, in most literature, their systems have only exogenous customers and concentrate on optimizing either staffing or routing. In the next chapter, we introduce a case study of a CSC system, and show that an invitation policy is applicable and promotes the system reward. This shows that optimizing an invitation policy is a promising direction to study for CSC systems.

1.2 Research Objectives and Thesis Structure

We aim to find an inviting policy capable of optimizing certain performance measures. Several questions are of interest: Which performance measures should be taken into account when constructing the model? Which type of invitation policy should one use? When should the system invite customers? What is the tradeoff between system reward and service level?

Hence, the objective of this work is to develop an invitation policy for a multi-server system with impatient non-identical customers, so as to maximize revenue, taking into account customers' value, service level and system efficiency. By capturing the setting of a chat service system, the service level could be expressed in term of penalty for abandonment and cost of waiting. The system efficiency could be expressed by the operating cost of available agents in the system. We start with analyzing the fluid approximation of such a system, whose optimal policy leads us to a threshold policy. This policy is very simple, basically stating that customers should only be invited if an available agent appears. This is not a very realistic policy since customers do not enter the system immediately, and many customers reject invitations. Also, we show through simulation that this policy is not optimal. Hence, we refine the fluid model to allow for a larger variety of the thresholds to be considered. We continue by analyzing the fluid model equilibrium and approximate performance measures of the system operating with different thresholds. Under a certain service level requirement, the decision maker can then evaluate the revenue of different invitation thresholds to propose a better one. This study has the following contributions:

- We construct our model and revenue function based on an empirical study (Chapter

2). By using real data, we elaborate the importance of classifying the customers according to their values, the fact that indeed, providing service to the right customers enhance revenue, and the impact of different metrics of service levels. In particular, the case study provides justification for the use of small data information in optimizing operations. Such small data are collected by automated systems on potential customers, which sometimes is more effective than the big data (Lam et al. 2017).

• By solving a linear programming problem of a fluid model for our multi-server system with impatient non-identical arrivals, we determine an optimal invitation policy that ranks customers by the product of revenues multiplied by the service rate. We discuss its limitations in Chapter 3.

• Based on the fluid analysis, we propose a threshold policy for invitation control. Under this control, we leverage the fluid equilibrium result and develop a stochastic approximation of performance levels using the Filipov method (Filipov 1988). Such an approximated result provides an evaluation of various threshold controls. Those approximations are presented in Chapter 4.

• In Chapter 5, we develop more approximations based on the fluid equilibrium for both revenue evaluation and service level indication. All the acquired approximations perform well in the simulation.

Chapter 2

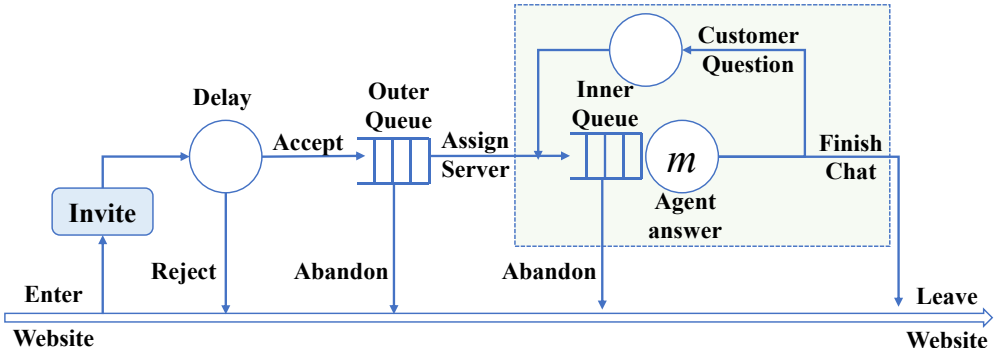
An Empirical Study of a Proactive Chat Service System

Before constructing the model for theoretic analysis, we first investigate an existing proactive service system, in order to sketch the most important features of such systems. To that end, we explore empirically the sensitivity of the revenue to some selected operational parameters.

2.1 System Overview

The customer dataset comprises of more than a half million chats of an airline company over one month. This company website provides both service and sales, through its contact center. Anyone who is interested in the business can visit this website at anytime. Customer flow in the system is described by Figure 2.1. The system traces all online

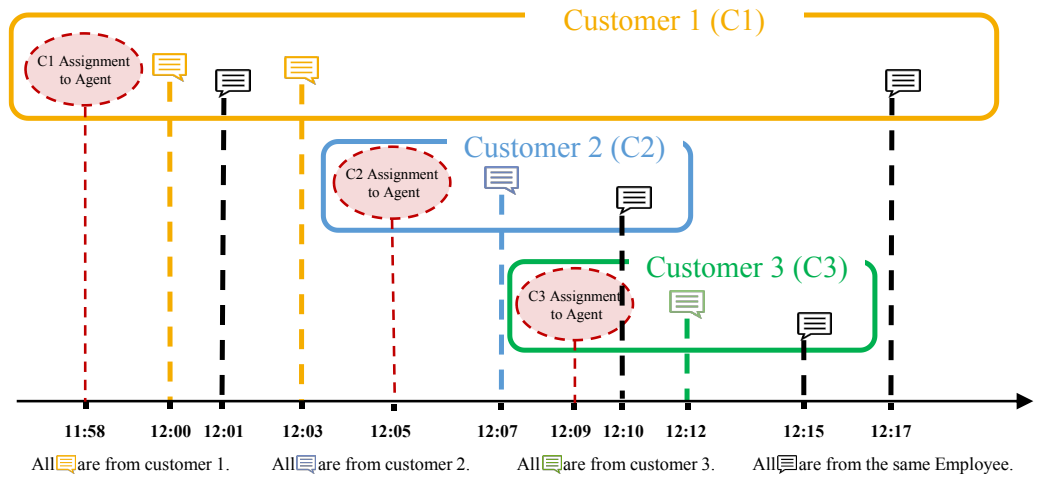
Figure 2.1: Customer perspective process description



customers and computes scores by an analysis of their browsing behavior and personal information (browser, location, etc.). According to customer score and chat service capacity, the system sends invitations to high-score customers. The invitations are sent by either a button displayed on the webpage or as a pop-up window. The customers, who receive the invitation, make a decision on whether to accept it. Once accepting

the invitation, a customer starts to wait in the outer queue until the system detects an available server and assigns this customer to that server. Note that every server can manage up to 3 customers simultaneously; therefore, the number of "in service" chats can be larger than the number of online servers. Customers have finite patience, hence, they may abandon the outer queue before they are assigned to a server. We may not know that a customer abandoned till the service began. Then, the chat service starts in the form of an alternate server and customer line. Since agents manage multiple customers simultaneously, customers may need to wait in the inner queue for the server during their dialog. In Figure 2.2, an example is demonstrated on the perspective of an agent who serves three customers simultaneously. Upon customer assignment, the customer and server talk one after the other continuously. The agent may wait for the customer entry (e.g. from 12:01 to 12:03, this server is idle). Because the maximal multi-task level is 3, from 12:09, this agent cannot get any extra customer. A customer may also wait for the busy server who is replying to some other parallel customer (e.g. from 12:12 to 12:15, customer 1 is waiting in the inner queue, because the server is busy serving customer 3).

Figure 2.2: Server perspective process description

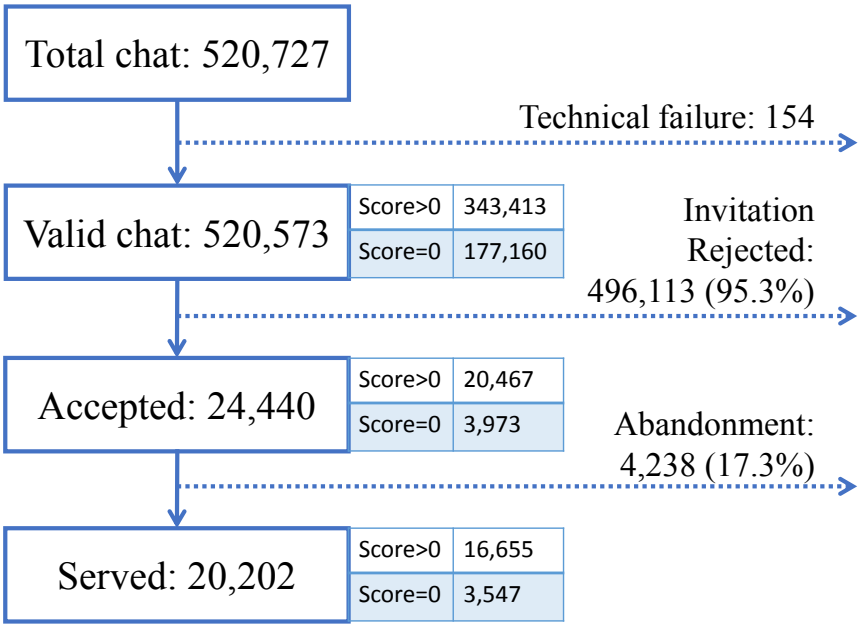


After the service is finished, the system receives the information that the corresponding service load is released and it starts to assign a new customer. After the chat, the customer may stay or leave the website. Before leaving, some customers may purchase commodities in this website, which we will refer to *conversion* in this context. In this analysis, we consider the conversion rate, which is the proportion of customers purchasing commodities, as the main output of the service. The aim is to investigate how the customer properties as well as operational decisions can impact this rate.

2.2 Data Description

The above chat system records information of each invited customer, The data include: personal information (browser, location, etc.), score, time stamp of all important events described in customer flow, conversion, and some other interested indicators. We collected 520,727 chats that were served during January 2016 (Figure 2.3). Each of them stands for an invited visit customer on this website. By screening the data, 154 error chats are excluded because of a technical failure.

Figure 2.3: Selection of chat sample



It can be noticed that there are three crucial events in the process: sending the invitation, accepting/rejecting the invitation and assigning a server. One can classify all invited customers into more specified groups: the accept/reject customer and the served/abandoned customers. Figure 2.3 shows that 95.3% customers ignore the invitation. Out of those who accepted an invitation, 17.3% abandon the queue. Finally, there are only 20,202 customers entering service, which is 3.88% out of all the invited customers.

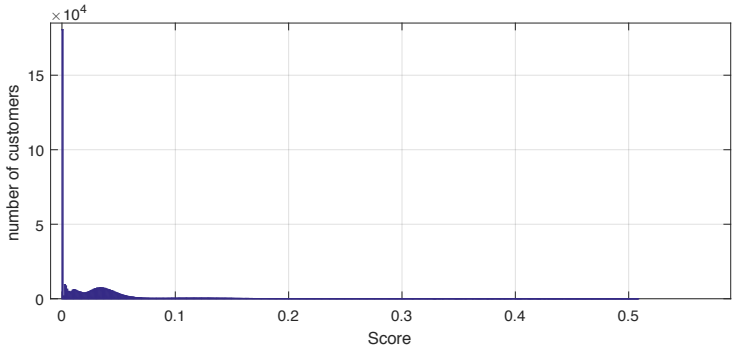
2.2.1 Descriptive Analysis of Customer Score

Customer score is the most unique characteristic of this system that maps all the information on customer activity on the website into a single value. Before going deeper into the service procedure, we want to initially verify the validity of the score as value representative. In other words, does the high score result in a higher income? Meanwhile, we also consider the following questions: How to characterize the score? How can we use the score in the analysis? How does the score relate to other indicators/features of the customer?

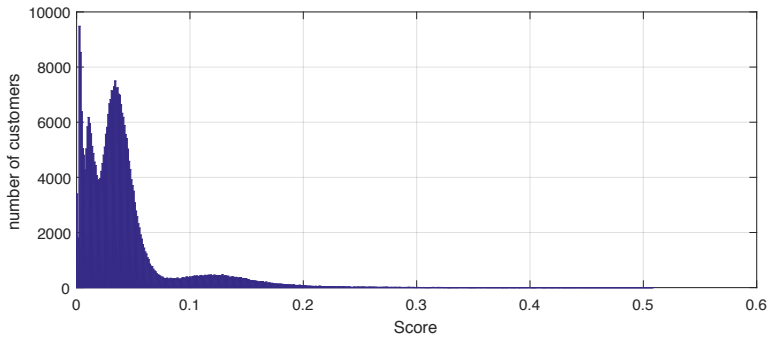
Customers scores range from 0 to 0.6. By checking the distribution of all customer scores (Figure 2.4a), one can notice that there is a large proportion (around 1/3) of customers with a score equal TO 0. A Zero score might be due to low value or lack of information. In order to identify if indeed the zero score customers are low value customers, we examine the information collect time (ICT), defined as the time interval between customer entering the website and receiving the invitation. The average ICT of all customers is 197.8 seconds. However, among all zero score customers, the average ICT is only 0.59 seconds, which is much lower than the average ICT among non-zero score customers – 299.6 seconds. Furthermore, there are around 71.6% zero score customers with 0 ICT, which means that they are invited on entering the website. This can happen when the customer arrives at a non-peak hour. During ICT, the system is collecting the customer information especially their online behavior. If the ICT is very short, one has reason to believe that the zero score is such due to the lack of information but not the lack of value. Therefore, we will consider the scored zero and non-zero score customers, separately, in our analysis.

Figure 2.4: Customer score distribution

(a) All customer score distribution



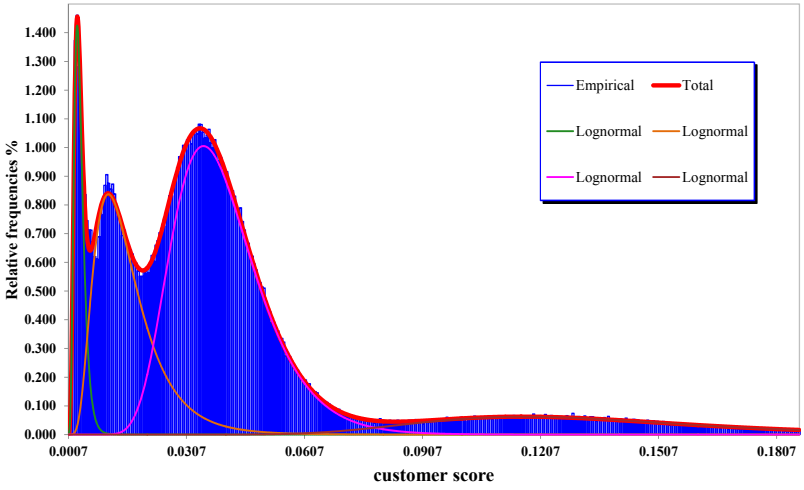
(b) Non-zero score customer distribution



After excluding the zero score customers, the score distribution is scaled (shown in Figure 2.4b). It can be observed that scores are concentrated in four regions. By following that observation, it is natural to divide customers into four groups based on

their score distribution pattern. How to define such division is not clear. One way, is to fit a mixture distribution for that distribution. Such a fitting is presented in Figure 2.5. We observe that a good fit results from mixing 4 log normal distributions. The relative proportion of each group is given by the weight 8.33%, 26.33%, 54.52% and 10.83%. The problem with this approach is that it does not provide a clear classification for specific customers, i.e., to which group that customer belongs. Hence, we take a simpler approach in which we use the local minima of the score distribution to determine three breakpoints (0.0082, 0.0208 and 0.082) that separate the non-zero scores into 4 groups (see the first two columns of data in Table 2.1). As a robustness check we repeat all the analysis by considering both score and group level.

Figure 2.5: Mixture distribution fitting of customer score



Before tracking customers' characteristics among the groups, we use the group information to overview the relationship between score and system output, so as to take the first step towards validating score as a representative value. The output information we have is an indicator of the conversion behavior, namely, whether the customer spent money or not during that visit to the website. Hence, we check the conversion rate, i.e., the proportion of the conversion customers out of the total group population, in each group. The results are listed in Table 2.1.

Table 2.1: Conversion rates among groups

		All				Served			Non-Served	
		Customer Number	Percent of all Population	Conversion Number	Conversion Rate	Served Number	Conversion Number	Conversion Rate	Conversion Number	Conversion Rate
Score = 0		177160	34.03%	4557	2.57%	3547	260	7.33%	4297	2.48%
Score >0	Group1	43955	8.44%	144	0.33%	3542	12	0.34%	132	0.33%
	Group2	60792	11.68%	545	0.90%	4454	46	1.03%	499	0.89%
	Group3	203243	39.04%	7526	3.70%	6688	535	8.00%	6991	3.56%
	Group4	35423	6.80%	10662	30.10%	1971	611	31.00%	10051	30.05%

For all non-zero score customers, the conversion rate has consistent growth as shown by their higher score. Especially in the highest-scored group, customers express 100

times more willingness to consume on this website compared to the lowest-scored group. Meanwhile, after considering the customers who get service and who do not get service separately, the effect of the service can also be identified. The conversion rate of the customers who receive service is generally higher than those who do not. Notably, for group 3 customers, which holds around 40% of the total population, their conversion rate soars twice after service. To sum up, the initial analysis suggests that it is worthy to provide proactive service to customers, in particular to high-scored customers.

Meanwhile, some operational parameters also show differences among groups (Table 2.2). Generally speaking, as scores increase, customers are more likely to reject the invitation, and their probability of abandonment decreases. The highest-scored customers (around 7%) show opposite behavior, which is worthy for further study. The average length of stay has some fluctuations (around 5%) among groups.

Table 2.2: Operational parameters among groups

		Number	Accept Number	Accept Rate	Serve Number	Abandon Rate	Average Abandon Time	Average Length of Stay
Score = 0		177160	3973	2.24%	3547	10.72%	110.2	680.6
Score >0	**Group1**	43955	4569	10.39%	3542	22.48%	195.5	747.3
	Group2	60792	5801	9.54%	4454	23.22%	181.2	749
	Group3	203243	7787	3.83%	6688	14.11%	150.9	726.2
	Group4	35423	2310	6.52%	1971	14.68%	153.5	765.6

2.3 Three Level Logistic Regression Mode

In order to draw a conclusion on customer behavior, a more robust statistic analysis is needed. We build a logistic model to explain how customer value and operational decisions impact conversion. Since the output, conversion is a binary indicator representing a purchasing / non-purchasing event, we use logistic regression for the analysis. Note that the majority of customers ignore the invitation; hence, most of their information do not exist. This also happens to the customers that abandon in the outer queue. Therefore, we built models in three levels of analysis: (1) all invited customers, (2) invitation accepted customers, and (3) served customers, separately.

2.3.1 Parameter and Data Selection

The predicted variable, as explained, is the conversion indicator. The explanatory variables we choose are listed in Table 2.3.

These variables are chosen from three categories: general control variables, customer indicators and system operational indicators. The first group includes the hour of day (HOUR) and day of week (DAY). The second group covers customer preference and characteristics. This includes the parameters mentioned before (SCORE, GROUP), The ACC_RECO and SEV_RECO are the indicators of invitation acceptance and receiving

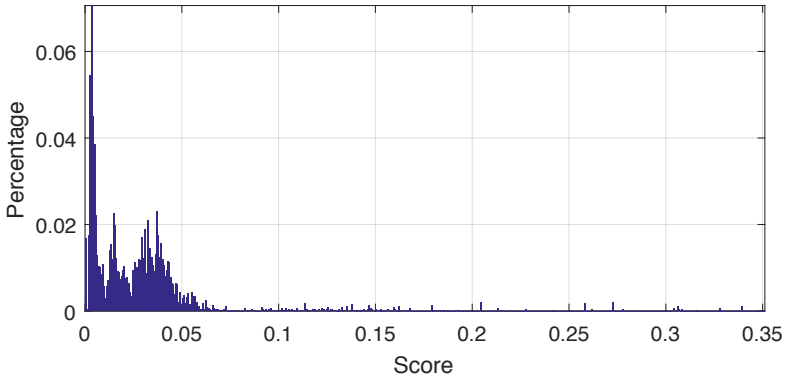
Table 2.3: List of parameters

Type	Name	Explanation	Mean	STD	Levels
Predicted variable	**CONVER**	Conversion 1: 4913/No conversion 0: 139692	/		**All**
GENERAL	**HOUR**	Hour of day (3∼17)	/		**All**
	DAY	Day of week (7 days)	/		**All**
Customer indicators	**SCORE**	Customer score	0.0125	0.0243	**All**
	GROUP	Divided by score: GROUP1 (0,0.0082]: 43955 / GROUP2 (0.0082,0.0208]: 27357 / GROUP3 (0.0208,0.082]: 67346 / GROUP4 (0.082, 0.6): 5947	/		**All**
	ACC_RECO	Accept the invitation 1: 14141 / Reject 0: 307624	/		**All**
	SEV_RECO	Served 1: 11658 / Abandon 0: 2483	/		**All**
	SKILL	Different service types: Service 1: 5139/ Sales 2: 6519	/		**Sev**
	AVG_SENT	Average sentiment score for all customer line	0.1104	0.3836	**Sev**
	END_SENT	Average sentiment for the last 10% cutomer line	0.2307	0.7356	**Sev**
	SENT_TREND	Increase 1: 3413/Nonchange 0: 5492/Decrease -1: 2753 sentiment between the last/first 10% customer line	/		**Sev**
	LOS	Chat duration between exit queue and chat end in seconds	726.8876	558.8086	**Sev**
	NO_WORDS	Number of words given by the agent during chat	162.7651	136.1779	**Sev**
System operational indicators	**INV_TYPE**	Button invite 1: 60626 or Window invite 2: 261139	/		**All**
	QUEUE_SEC	Waiting time for the outside queue	71.0063	163.5771	**Acc**
	PROP_INNERQ	Waiting time for the inner queue / LOS	0.1870	0.2101	**Sev**
	MULTI	Average multi-task level	2.4258	0.5797	**Sev**

service, respectively. We define SKILL as the purpose of the visit (seeking service or sales). The length of stay (LOS) is defined as the chat duration and NO_WORDS counts the number of words the agent wrote during one chat, which reflects the service workload of that specific chat. Apart from the score, the system also traces customer sentiment on sentence level while they are in service. Several emotion indicators were included in the model. We sum up the sentiment score (range from -10 to 7) to the chat level by average sentiment (AVG_SENT), the sentiment at the end of the conversation (END_SENT), and the sentiment change during the whole chat (SENT_TREND). In the last group, some operational parameters are selected, including the invitation type (INV_TYPE) (button displayed on the webpage or a pop-up window), the queueing time for outside (QUEUE_SEC), the proportion of inner queueing time in the total length of stay (PROP_INNERQ) and the average multi-task level of the server during that chat (MULTI). The last column in Table 2.3 describes to which level of analysis this variable is relevant.

The current invitation policy prioritize customers according to their score; hence the data is biased and includes a higher proportion of Group 3 and 4 than the general population. Meanwhile, the system invites customers also according to the system load, which is independent of the score distribution and, therefore, can be considered as a nature experiment. In order to eliminate the data bias, we use an importance sampling approach (Kroese and Rubinstein 2008) by which we sample our data according to the score distribution of all customers on the website (including both invited and non-invited customers, see Figure 2.6). From the distribution we can see that the customers still can be divided into the same 4 groups according to their score. After importance sampling, our sample includes 144605 chats.

Figure 2.6: Customer score distribution of all customers on website



2.3.2 Level 1: All Customers

In this level of analysis, we build the first model to check the validity of scoring, as predicting conversion.

$$Logit\left(P\left(CONVER_i\right)\right) = \beta_0 + \beta_1 \cdot SCORE_i + \varepsilon_i. \tag{2.1}$$

Model 2.1 predicts the probability of conversion using a logistic regression. The results are shown in Table 2.4. In the logistic model, the e to the power of the coefficient is the amplifier of the odds ratio of this variable. Hence, the result shows that the customer score has a significant positive effect on conversion.

Table 2.4: All customers fit logit Model 2.1

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
SCORE	28.2951	0.3015	93.83	<2e-16 ***

In Section 2.2.1, we classified the customers according to their score-group. Such approach is the one we use in our theoretical study. Therefore, we repeat the analysis with score-based class information to check its robustness:

$$Logit\left(P\left(CONVER_i\right)\right) = \beta_0 + \beta_1 \cdot GROUP_i + \varepsilon_i. \tag{2.2}$$

By fitting all customer data to Model 2.2, the result shows consistency with the result of Model 2.1 (see Table 2.5).

Table 2.5: All customers fit logit Model 2.2

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
GROUP2	0.91192	0.10728	8.501	<2e-16 ***
GROUP3	2.45334	0.08594	28.547	<2e-16 ***
GROUP4	5.08883	0.0878	57.961	<2e-16 ***

Next, we want to add more control variables to improve the prediction and the inter-operational decision impact on conversion. According to Table 2.3, only several variables are available for all customers. Thus, the next model is built to confirm the validity of scoring and the utility of service.

$$Logit\left(\Pr\left(CONVER_i\right)\right) = \beta_0 + \beta_1 \cdot SCORE_i + \beta_2 \cdot ACC\_RECO_i + \beta_3 \cdot SEV\_RECO_i + \beta_4 \cdot INV\_TYPE_i + \beta_5 \cdot GENERAL_i + \varepsilon_i. \tag{2.3}$$

The result (see Table 2.6) shows that as before the customer score has a significant positive effect on conversion. More important is the fact that providing service also has a positive effect. However, the acceptance does not impact conversion in a positive way, which means receiving an invitation is not enough to increase conversion, whereas reaching service is crucial. From the fact that acceptance and service show opposite effects, we conclude that the abandonments have a negative influence on conversion. Therefore, an abandonment penalty should be added when discussing the system revenue. Such a penalty can be considered as opportunity-loss costs. In addition, conversion is also significantly different between the two invitation types. Button invitation seems to perform better — it could be because a pop-up window may interrupt browsing. This phenomena is interesting and should be investigated in future research.

Table 2.6: All customers fit logit Model 2.3

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
SCORE	24.8488	0.3115	79.772	<2e-16 ***
INV_TYPE2	-2.11885	0.03697	-57.308	<2e-16 ***
ACC_RECO1	-0.58604	0.16918	-3.464	0.000532 ***
SEV_RECO1	0.44623	0.17991	2.48	0.013129 *
GENERAL	included			

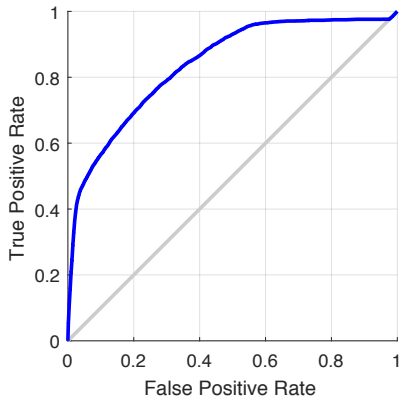
Moreover, we plot the Receiver Operating Characteristic (ROC) curve for the predictors of both Model 2.1 and 2.3 on conversion, in Figure 2.7. The value of the area under the ROC curve is the statistical measure of how much better that model can rank a randomly chosen positive instance higher than a randomly chosen negative one (Fawcett 2006). The larger area under the curve shows that the model is more accurate. The model with the operational variables predicts better, which means that pre-service information, although very good, does not include all impacts on conversion. Indeed having service is essential for explaining conversion accurately.

2.3.3 Level 2: Customers Who Accept the Invitation

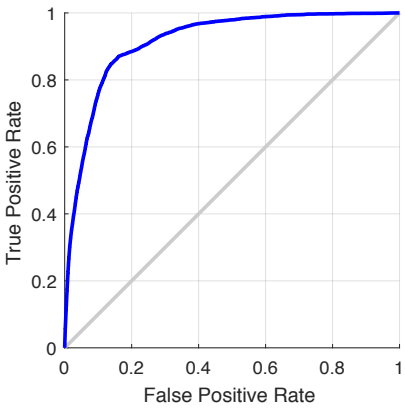
After invitation acceptance, customers are waiting in the outer queue until they are assigned to an available server. In this step, we investigate whether waiting in the outer

Figure 2.7: ROC curve for predicting conversion

(a) Predictor of Model 2.1



(b) Predictor of Model 2.3



queue affects conversion rate. For that purpose we propose the following model:

$$\begin{aligned} Logit\left(P\left(CONVER_i\right)\right) = \beta_0 + \beta_1 \cdot SCORE_i + \beta_2 \cdot QUEUE\_SEC_i \\ +\beta_3 \cdot SEV\_RECO_i + \beta_4 \cdot INV\_TYPE_i + \beta_5 \cdot GENERAL_i + \varepsilon_i \end{aligned} . \quad (2.4)$$

According to the statistical result (in Table 2.7), the longer a customer waits in the outer queue, the less chance a purchase will be made during the visit. In numbers, waiting 1 more minute (60 seconds) will decrease the odd ratio of the probability of conversion by 9.15%. Hence, it is reasonable to include holding cost, when discussing system optimization.

Table 2.7: Customers who accept the invitation fit logit Model 2.4

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
SCORE	22.9123	1.0441	21.9450	<2e-16 ***
INV_TYPE2	-1.7984	0.1585	-11.3440	<2e-16 ***
QUEUE_SEC	-0.0016	0.0006	-2.7260	0.006412 **
SEV_RECO1	0.2903	0.1835	1.5820	0.1137
GENERAL	included			

2.3.4 Level 3: Served Customers

On the service level, both customer and server characteristics are examined. From the point of view of the customer, we check: Is the inner queue waiting also negatively correlated with conversion? What is the impact of service time and how is customer sentiment during the chat associated with conversion rates? From the aspect of the server, we check the impact of workload and multi-task level on conversion. We thus

check the following model:

$$\begin{aligned} Logit\left(P\left(CONVER_i\right)\right) = \beta_0 + \beta_1 \cdot SCORE_i + \beta_2 \cdot INV\_TYPE_i + \beta_3 \cdot SKILL_i \\ +\beta_4 \cdot GENERAL_i + \beta_5 \cdot NO\_WORDS + \beta_6 \cdot MULTI_i \\ +\beta_7 \cdot AVG\_SENT_i + \beta_7 \cdot END\_SENT_i + \beta_9 \cdot SENT\_TREND_i \\ +\beta_{10} \cdot \log\left(LOS_i\right) + \beta_{11} \cdot PROP\_INNERQ_i + \varepsilon_i \end{aligned} . \tag{2.5}$$

Table 2.8 presents the model results. It seems that the sentiment factors have no significant effect. A surprising effect is observed in the inner waiting queue. While waiting in the outer queue had a negative impact on conversion, waiting in the inner queue is positively associated with conversion. This means that the customers with a larger proportion of wait during their total length of stay, have a higher probability of conversion. Note that waiting in an inner queue is practically waiting while being served, and the customers are less aware of such waiting. Hence, it may be that such a wait is reflected to customers as being served longer and not necessarily as waiting longer. This finding fits similar observations made in restaurants (Tan and Netessine 2014). It implies that the longer the perceived service, the higher probability of conversion. Last, as expected, the sales skill is associated with higher conversion.

Table 2.8: Customers who get service fit logit Model 2.5

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
SCORE	13.329468	1.203909	11.072	<2e-16 ***
INV_TYPE	-1.865105	0.169444	-11.007	<2e-16 ***
NO_WORDS	-1.69e-03	6.20e-04	-2.73	0.00634 **
MULTI	-0.102772	0.123616	-0.831	0.40576
AVG_SENT	0.278269	0.267378	1.041	0.298
END_SENT	0.065398	0.176364	0.371	0.71078
SENT_TREND20	0.110879	0.201034	0.552	0.58126
SENT_TREND21	0.143322	0.288031	0.498	0.61877
LOS	0.187859	0.118506	1.585	0.11292
PROP_INNERQ	0.956183	0.295775	3.233	0.00123 **
SKILL	3.740796	0.367691	10.174	<2e-16 ***
GENERAL	included			

To sum up, through the above empirical study:

• We show that the customer ranking information is acquirable and valid for optimizing invitation policy. Not all customers should be invited. The distribution of such information allows us to classify all customers into a limited number of groups which may simplify the theoretical analysis in the following chapters.

• Operational factors such as load results in waiting and abandonment. Both factors have negative correlation with conversion. Hence, the cost of waiting, as well as the penalty of abandonment should be considered when maximizing revenue of the system. Another option is to maximize revenue under some performance measure constraints

that will limit the negative effect of overload.

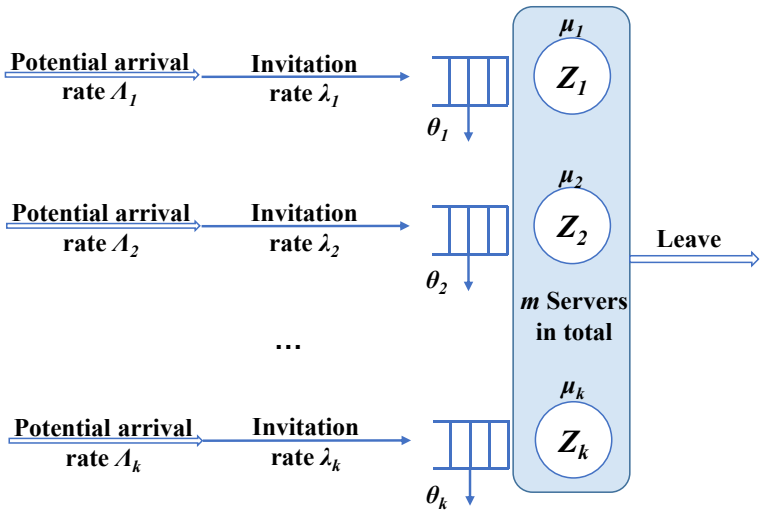
• The operation factors that are unique to the chats, i.e. parallel service and inner wait, should not be consider as costs.

Chapter 3

Model of a Proactive service System

Following the idea of classification, we group customers with similar characteristics together and start the analysis by finding the optimal invitation policy at the group level. Since all group information is available to the system, it is natural for the system to use customer value in the routing policy. Hence, from now on, we discuss a multi-server, multi-class queueing service system (Figure 3.1) that operates over a infinite-time horizon.

Figure 3.1: Equivalent system description



The system has several customer classes that differ in their customers' value. All potential class i customers arrive to the system according to a Poison process with rate Λ_i, where $i \in \mathcal{K} = \{1, ..., k\}$. Following a predetermined invitation policy, system invites class i customers according to a Poison process with rate $\lambda_i \leq \Lambda_i$. Customers wait in a class-dedicated queue with infinite capacity until they are assigned to a server. During waiting, customers may abandon the queue due to their exhausted patience. Class i

customer patience is exponentially distributed with rate θ_i. Each arriving customer requires a random amount of service. The customers' service times are exponentially distributed with mean $1/\mu_i$, for customer class i. Every service provided to class i customers brings a reward with a value of r_i. However, waiting time and abandonment execute a penalty with positive cost c_i^h per unit of waiting time and c_i^{ab} per renege customer, respectively. Aiming to maximize revenue—the difference between system reward and cost—a dynamic control policy was developed, to determine the effective invitation rates.

The above stochastic model can be described as a continuous time Markov process denoted by $\{\mathbb{X}(t), \mathbb{Q}(t), \mathbb{Z}(t)\} = (\{X_i(t), Q_i(t), Z_i(t)\}, t \geq 0)$: $X_i(t)$ and $Q_i(t)$ are the total headcount of class i customers in the system and in the queue at time t, respectively, and $Z_i(t)$ is the number of servers that serve class i customers at time t. All servers share a server pool with a total of m statistically identical servers who cater to all types of customers. Apparently, all stochastic variables are defined on the non-negative quadrant and for all $i \in \mathcal{K}$ satisfy

$$\begin{aligned} &X_i(t) = Q_i(t) + Z_i(t); \\ &\sum_{i \in K} Z_i(t) \leq m. \end{aligned} \tag{3.1}$$

Denote A_i as the arrival Poisson process with rate λ_i, and D_i and R_i, as the departure processes from service and abandonment, respectively. We denote the initial condition of the system by $X_i(0)$. The dynamics of the process of the number of each class customer can be characterized by Equation (3.2). Any proposed invitation policy has to satisfy the dynamics provided by Equations (3.1) and (3.2). Note that the system does not permit work conservation.

$$X_i(t) = X_i(0) + A_i(t) - D_i(t) - R_i(t), i \in \mathcal{K}. \tag{3.2}$$

By using the system state variables, the instantaneous cost of class i customers at time t can be computed by

$$C_i(t)\,dt = c_i^h \cdot Q_i(t)\,dt + c_i^{ab} \cdot dR_i(t). \tag{3.3}$$

Because the patience of any class of customer is exponentially distributed with rate θ_i, at time t, the expected abandonment rate of the class i customer can be written as $\theta_i \cdot Q_i(t)$ (Atar et al. 2010). Hence, the above cost function can be modified to Equation (3.4). For computational simplicity, $c_i = c_i^h + \theta_i \cdot c_i^{ab}$ is used as a unified cost parameter of class i from now on. According to the definition of cost parameters, it is clear that c_i is positive.

$$\begin{aligned} C_i(t) &= c_i^h \cdot Q_i(t) + c_i^{ab} \cdot (\theta_i \cdot Q_i(t)) \\ &= (c_i^h + \theta_i \cdot c_i^{ab}) \cdot Q_i(t) \\ &= c_i \cdot Q_i(t) \end{aligned} \quad . \tag{3.4}$$

Meanwhile, the system is rewarded by each customer who finishes service. As the service process is exponentially distributed with rate μ_i, respectively among classes, the customer service completion rate is

$$dD_i(t) = \mu_i \cdot Z_i(t). \tag{3.5}$$

Thus, the total instantaneous system revenue is the summation of the revenue of all classes, that can be expressed by

$$R_{total}(t) = \sum_{i \in \mathcal{K}} (r_u \cdot \mu_i \cdot Z_i(t) - c_i \cdot Q_i(t)). \tag{3.6}$$

Furthermore, by considering the problem over an infinite time horizon, our objective is to find an invitation policy satisfying system constraints (defined by Equation (3.1) and (3.2)) that achieves the maximum average revenue defined by Equation (3.7). The second equation is a result of the independence between customer class and time.

$$\begin{aligned} \bar{R}_{total} &= \lim_{T\to\infty} \tfrac{1}{T} \int_0^T \left(\sum_{i \in \mathcal{K}} (r_i \cdot \mu_i \cdot Z_i(t) - c_i \cdot Q_i(t)) \right) dt \\ &= \sum_{i \in \mathcal{K}} \lim_{T\to\infty} \tfrac{1}{T} \int_0^T (r_i \cdot \mu_i \cdot Z_i(t) - c_i \cdot Q_i(t))\, dt. \end{aligned} \tag{3.7}$$

Next, we optimize the fluid scale of this stochastic model to acquire some understanding of the invitation policy.

3.1 The Optimal Fluid Policy

Denote x_i, q_i and z_i as the long run fluid averages of the process $X_i(t)$, $Q_i(t)$ and $Z_i(t)$ for class i customers, $i \in \mathcal{K}$. These fluid functions satisfy, in steady state, the following set of equations

$$\begin{cases} x_i = q_i + z_i; \\ \lambda_i = \mu_i z_i + \theta_i q_i; \\ \lambda_i \le \Lambda_i; \\ \sum\limits_{i \in \mathcal{K}} z_i \le m; \\ x_i, z_i, q_i \ge 0. \end{cases} \tag{3.8}$$

Under the above constraints, our objective is to maximize the total revenue over all sets $(\lambda_i, x_i, q_i, z_i)$. After simplification, the corresponding linear program (LP) is

$$\begin{aligned} \max_{\forall z_i, q_i} \quad & \sum_{i \in \mathcal{K}} (r_i \mu_i z_i - c_i q_i) \\ s.t. \quad & \sum_{i \in \mathcal{K}} z_i \le m \\ & \mu_i z_i + \theta_i q_i \le \Lambda_i, \forall i \in \mathcal{K} \\ & z_i, q_i \ge 0, \forall i \in \mathcal{K}. \end{aligned} \tag{3.9}$$

Since the above LP includes several inequality constraints, we use Karush–Kuhn–Tucker (KKT) conditions (Karush 1939) to determine the necessary optimality conditions of this convex problem. The Lagrangian is

$$\begin{aligned}\mathcal{L}\left(z_i, q_i, \alpha, \beta_i, \gamma_i, \sigma_i\right) = &- \sum_{i\in\mathcal{K}} \left(r_i\mu_i z_i - c_i q_i\right) + \alpha\left(\sum_{i\in\mathcal{K}} z_i - m\right) \\ &+ \sum_{i\in\mathcal{K}} \beta_i\left(\mu_i z_i + \theta_i q_i - \Lambda_i\right) - \sum_{i\in\mathcal{K}} \gamma_i z_i - \sum_{i\in\mathcal{K}} \sigma_i q_i.\end{aligned} \tag{3.10}$$

The KKT conditions are

$$\begin{cases} -r_i\mu_i + \alpha + \beta_i\mu_i - \gamma_i = 0 \\ c_i + \beta_i\theta_i - \sigma_i = 0 \\ \alpha\left(\sum_{i\in\mathcal{K}} z_i - m\right) = 0 \\ \beta_i\left(\mu_i z_i + \theta_i q_i - \Lambda_i\right) = 0 \\ \gamma_i z_i = 0 \\ \sigma_i q_i = 0 \\ \alpha, \beta_i, \gamma_i, \sigma_i \le 0 \end{cases}, \forall i \in \mathcal{K}. \tag{3.11}$$

Because c_i is positive, and β_i, θ_i are non-negative, all σ_i must be positive to keep the second condition of Equations (3.11) holding. Therefore, according to the sixth condition, any q_i must equal 0. Because the KKT conditions are necessary conditions, the above result means that the optimal fluid invitation policy does not permit a queue for any class of customers.

Let $q_i = 0$, then the original LP Equations (3.9) becomes

$$\begin{aligned} \max_{\forall z_i} \quad & \sum_{i\in\mathcal{K}} \left(r_i \cdot \mu_i\right) \cdot z_i \\ s.t. \quad & \sum_{i\in\mathcal{K}} z_i \le m \\ & z_i \le \Lambda_i/\mu_i, \forall i \in \mathcal{K} \\ & z_i \ge 0, \forall i \in \mathcal{K}. \end{aligned} \tag{3.12}$$

Let's relabel the classes according to a decreasing order of their rank, defined by the product $r_i\mu_i$, i.e.:

$$r_1 \cdot \mu_1 \ge r_2 \cdot \mu_2 \ge \ldots \ge r_k \cdot \mu_k. \tag{3.13}$$

Obviously, the optimal solution is to assign all the available servers, of which the number equals Λ_i/μ_i, to serve class i customers in decreasing order until all servers are occupied. Assuming k_0 is the last class that invite all customers and * denotes the optimal result of LP (3.9):

$$\begin{cases} z^* = \left(\frac{\Lambda_1}{\mu_1}, \frac{\Lambda_2}{\mu_2}, \ldots, \frac{\Lambda_{k_0}}{\mu_{k_0}}, m - \sum_{i=1}^{k_0-1} \frac{\Lambda_i}{\mu_i}, 0, \ldots, 0\right) \\ q^* = (0, 0, \ldots, 0) \end{cases}. \tag{3.14}$$

In other words, in the fluid scale, the optimal invitation policy is: rank customer by $r_i\mu_i$,

then use all system service capacity to invite customers with as high ranking customers as possible, so that the system runs in the critical load regime. We call this policy $r\mu$.

3.2 Effectiveness of the Fluid Policy

Next, we examine the obtained fluid policy via simulation. Note that the optimal invitation policy (3.14) categorizes the customer into three types: all invited, partially invited and non-invited. We are not interested in the third type. Therefore, in the following experiments, we simplify our model into a two-class system in which the higher-ranked class customers are all invited and the lower-ranked customers are partially invited.

We simulate both a large system with 200 servers ($m = 200$) and a medium size system with 40 servers ($m = 40$), with non-preemptive prioritize queues. We assume that for both class customers, the average patience is longer than the average service time. The high-ranked customers have $\mu_1 = 1$ and $\theta_1 = 0.5$. For the low-ranked customers we simulate two conditions that vary in their *relative* demand: a) $\mu_{22} = 0.8, \theta_{22} = 0.4$ – in this case, the service demand of class 2 customers is lower and their patience is shorter. b) $\mu_{21} = 1.25, \theta_{21} = 0.625$, in which class 2's service demand is higher and they are more patient. For the large system, the potential arrival rate for high/low value customers is 150 and 100, respectively. In the smaller system, the rate is 30 and 20, respectively. Meanwhile, in order to capture the pattern of optimal policy, we test several sets of reward/cost parameters. All simulated parameter sets are listed in Table 3.1. Generally speaking, we use in total 6 types of reward/cost parameter sets. The pair of reward parameters has 3 combinations: both high, class 1 high and class 2 low, and both low. In each pair of reward parameters we test both high class 2 cost and low class 2 cost.

Table 3.1: Parameter sets for simulations of fluid policy

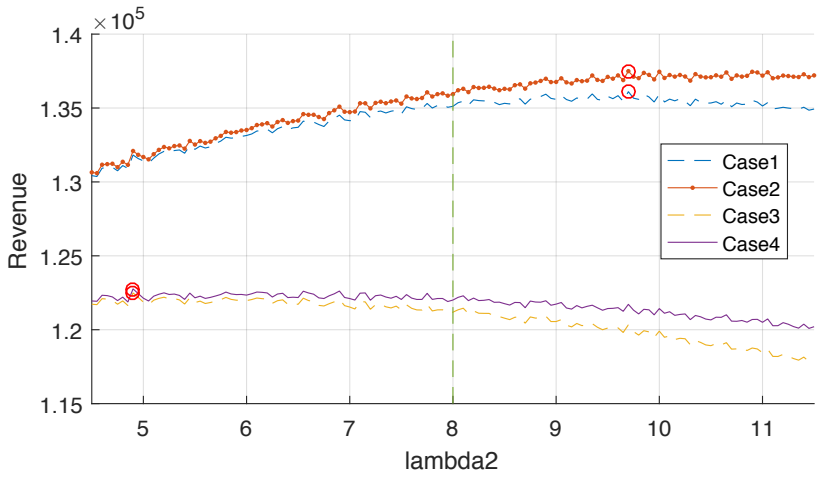
Set Number	**Number of servers** m	**High-value customer**								**Low-value customer**								**Fluid Expected Reward Total**
		Λ_1	μ_1	**Offered load$_1$**	θ_1	r_1	c_1^h	c_1^{ab}	**Fluid Expected Reward$_1$**	Λ_2	μ_2	**Offered Load$_2$**	θ_2	r_2	c_2^h	c_2^{ab}	**Fluid Expected Reward$_2$**	
1	40	30	1	30	0.5	8	0.5	1.2	120000	20	0.8	16	0.4	6	0.6	1.5	24000	144000
2															0.2	1		
3														2	0.6	1.5	8000	128000
4															0.2	1		
5						3			45000						0.6	1.5		53000
6															0.2	1		
7						8			120000		1.25	25	0.65	6	0.6	1.5	37500	157500
8															0.2	1		
9														2	0.6	1.5	12500	132500
10															0.2	1		
11	200	150		150		8			240000	100	0.8	80	0.4	6	0.6	1.5	48000	288000
12															0.2	1		
13														2	0.6	1.5	16000	256000
14															0.2	1		
15						3			90000						0.6	1.5		106000
16															0.2	1		

3.2.1 The Original Fluid Policy

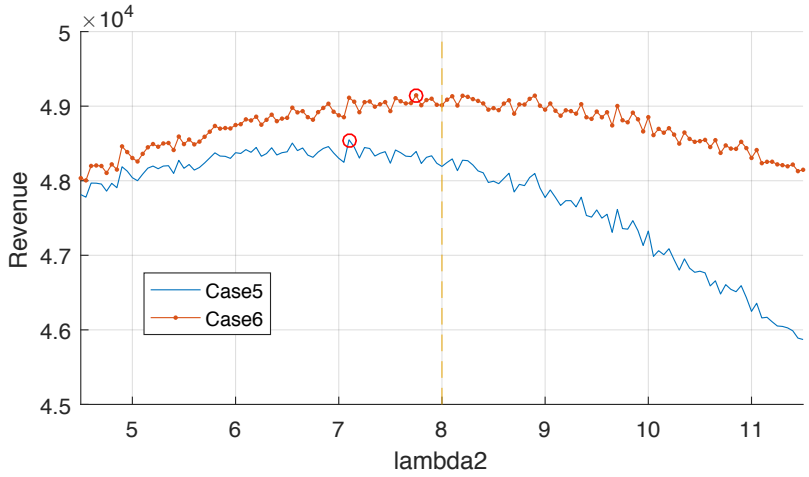
According to the first six parameter sets in Table 3.1, the fluid policy can be interpreted as: inviting high-ranked customers by rate 30 and inviting low-value customers by rate $\lambda_2 = 8$ (this is the result of equation: $40 - \Lambda_1/\mu_1 = \lambda_2/\mu_2$). We also check the revenue of the system for a range of λ_2 values around 8 – between 4.5 and 11.5, so as to examine the policy performance. The system revenue for different sets of cost parameters are illustrated in Figure 3.2. For each set of parameters, the optimal arrival rate (under this policy) is marked by *.

Figure 3.2: System revenue with arrival rate control ($m = 40, \mu_2 = 0.8$)

(a) Parameter sets Case 1 – 4



(b) Parameter sets Case 5 – 6



We can observe that for all cases, the fluid policy is not optimal. In particular, for cases 3 and 4, the fluid policy deviates from optimality. Obviously, such inaccuracy of the fluid policy is caused by the stochasticity of the system. On the fluid level, the system

is always supposed to be critically loaded. However, in reality, all customers arrive stochastically. Thus, the queues for both class customers are accumulated occasionally, whereas the servers are sometimes idle as well.

Meanwhile, due to different parameter combinations, the optimal invitation policy can be overestimated or underestimated. By comparing cases 3 and 4 (Figure 3.2a) to cases 5 and 6 (Figure 3.2b), we observe that when decreasing the reward of high-ranked customers, it is more beneficial to invite a higher rate of the low-value customers, because the relative value of low-ranked customers. However, all optimal rates are lower than the fluid optimal solution. It means that a queue is not welcome for those cases. By comparing cases 5 and 6 to cases 1 and 2 in Figure 3.2a, in which all optimal invitation rates shift higher, we learn that when the difference between cost and reward of both classes is getting larger, inviting more customers to the system becomes more and more profitable. Sometimes, we would rather keep an overloaded system to reduce the probability that the server is idle. In addition, we find that case 6 has more welcome low-ranked customers than case 5, which indicates that the invitation rate for low-value customers negatively depends on its waiting penalty. This is because when the invitation rate is getting higher, the expected queue length is also increasing.

Figure 3.3 demonstrates cases 7–10, where low-ranked customers have higher service and impatience rates than the high-ranked customers. According to previous analysis, customers are evaluated by $r\mu$. Therefore, an increase of μ_2 can be considered as the increasing of the reward rate per server for low-value customers. Hence, the increase in optimal invitation rates is not surprising.

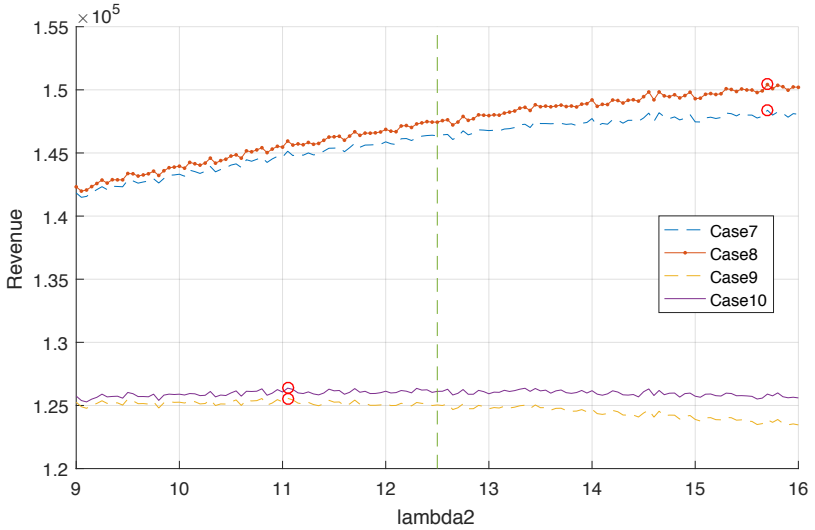


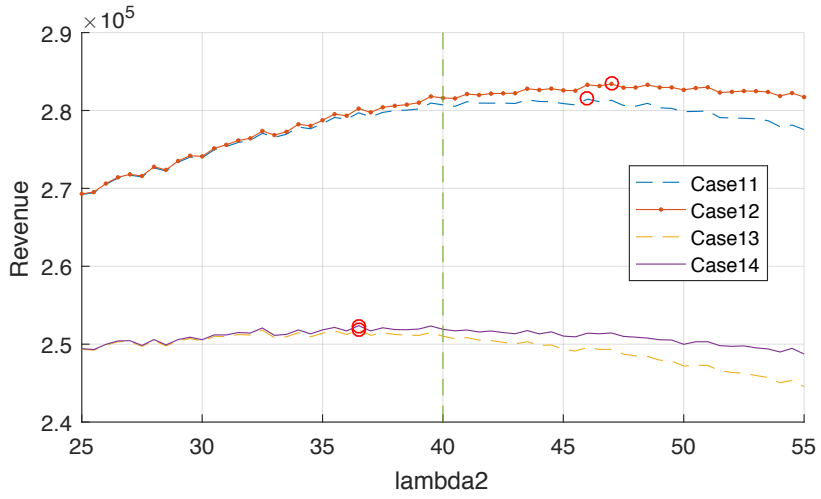
Figure 3.3: System revenue with arrival rate control ($m = 40, \mu_2 = 1.25$)

In addition, we check the same parameter sets as in Figure 3.2 for a large size system with $m = 200$. From the result shown in Figure 3.4, we find that the fluid policy becomes better for all cases, but is still not optimal. Apparently, in large systems

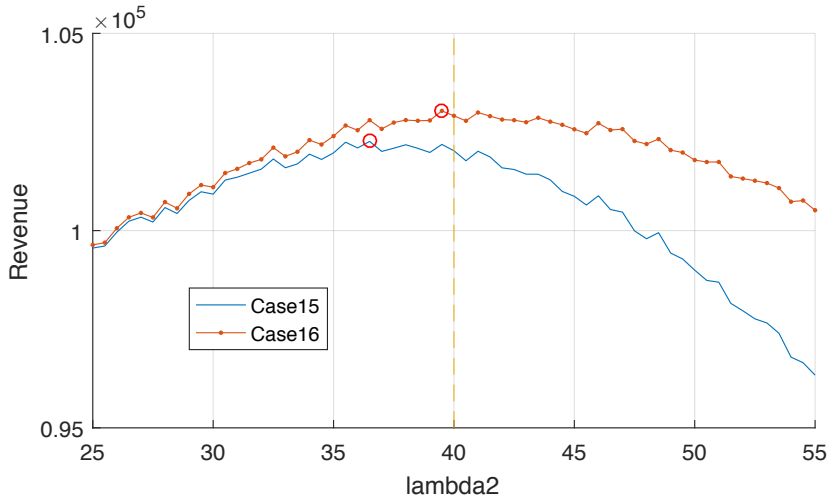
the fluid approximation is better. In large systems, when the reward/cost parameters change, the optimal invitation rate changes in the same way as in the medium size systems. However, we can find that for the same set of reward/cost parameters, the optimal rate moves close to the fluid policy invitation rate. Because in large systems, the risks of both queue accumulation and server idleness are lower than the risks in small systems.

Figure 3.4: System revenue with arrival rate control ($m = 200, \mu_2 = 0.8$)

(a) Parameter sets Case 11 – 14



(b) Parameter sets Case 15 – 16



3.2.2 The Applicable Threshold Policy

In reality, the potential customers arrive to the system stochastically. Therefore, we need to find realistic ways to implement this policy. The analysis suggests that a threshold

policy is approximately optimal. In this case, the fluid policy can be interpreted as: invite all potential high-ranked customers and stop inviting the low-ranked customers when all agents are busy, namely, the admission thresholds for high-ranked customers is infinity and $x_1 + x_2 \geq m$ for the low-ranked customer. As in the last subsection, we compare revenue under the fluid threshold to other policies with the same threshold structure but in which the value of that threshold varies below/above the theoretical one. By simulating both small and large systems with the same parameter sets listed in Table 3.1, the revenue of a threshold policy with different values are obtained and an optimal threshold is found. By comparing the revenue between different policies (see Table 3.2), we can find that in general, the threshold policy (columns 6–9) performs better than the original arrival rate control policy (columns 2–5). Moreover, for all parameter sets, the revenue under the fluid policy's equivalent threshold (column 6 and 7), though is not an optimal threshold (columns 8 and 9), is higher than the revenue under the optimal arrival rate control policy (columns 4 and 5). This good performance of the threshold policy is because the threshold control is a dynamic control. The admission of low-value customers is adjusted by system state, and is able to achieve a lower variance around the targeted load value, as we shall see in Section 4.

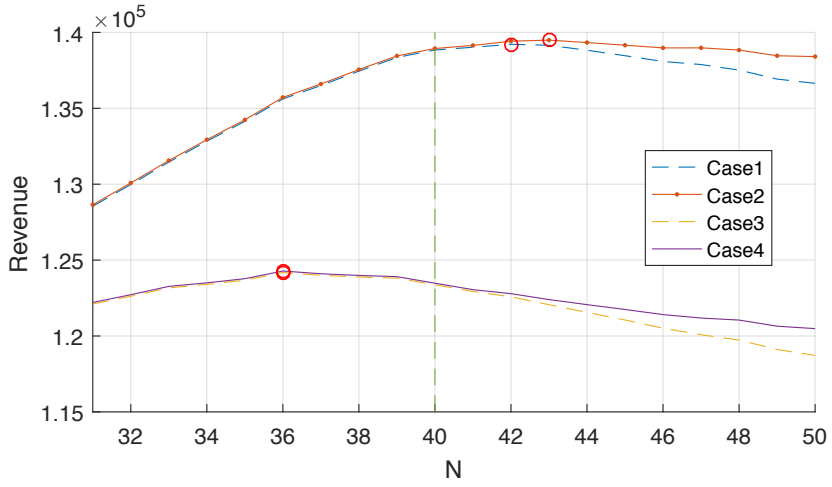
Table 3.2: Comparison of the revenue between arrival rate and threshold control policy

Set Number	**Arrival Control Fluid Policy**		**Arrival Control Optimal**		**Threshold Control Fluid Policy**		**Threshold Control Optimal**	
	Rate	**Revenue**	**Rate**	**Revenue**	**Threshold**	**Revenue**	**Threshold**	**Revenue**
1	8	135116.011	9.7	136107.7	40	138839.7	42	139214.5
2	8	135937.43	9.7	137494.4	40	138939.5	43	139490.8
3	8	121186.945	4.9	122488.9	40	123378.5	36	124180.3
4	8	122008.363	4.9	122762.4	40	123478.4	36	124280.1
5	8	48191.4448	7.1	48548.75	40	50404.75	40	50404.75
6	8	49012.8635	7.75	49145.54	40	50504.6	40	50504.6
7	12.5	146362.181	15.7	148387	40	149179	47	151554.2
8	12.5	147437.965	15.7	150417.5	40	149279	47	152503.9
9	12.5	125042.314	11.05	125566.4	40	127902.9	41	127916.8
10	12.5	126118.098	11.05	126372	40	128002.9	42	128052.2
11	40	280721.641	46	281454.5	200	285117.8	203	285699.9
12	40	281605.898	47	281605.9	200	285157.7	205	285935.4
13	40	251022.974	36.5	251834.6	200	253633.2	197	254176.7
14	40	251907.232	36.5	252372.4	200	253673.2	197	254216.7
15	40	102024.807	36.5	102261.2	200	104618.4	199	104742.5
16	40	102909.065	39.5	103035.7	200	104658.4	199	104782.4

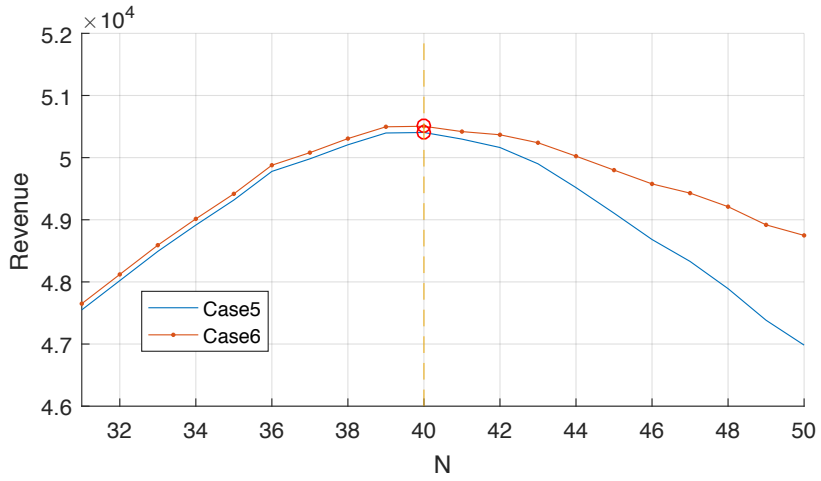
Figure 3.5 presents the revenue when applying a threshold policy with identical parameter sets as in Figure 3.2. The optimal thresholds are much closer to the corresponding fluid one, while the changing pattern is still similar. Such accuracy of the fluid policy is more obvious in large size systems (see the last six rows in Table 3.2). However, though its accuracy improved greatly, unfortunately, in all parameter sets we tested, none of the fluid policy is optimal.

In addition, in all cases, the difference between the simulation revenue of fluid arrival rate or threshold control policy (columns 3 and 7 in Table 3.2) and fluid expected total

Figure 3.5: System revenue with threshold control ($m = 40, \mu_2 = 0.8$)



(a) Parameter sets Case 1 – 4



(b) Parameter sets Case 5 – 6

reward (last column in Table 3.1) is no more than 10% and 5%, respectively. This means that though the fluid policy is not optimal, it performs well. However, in a practical sense, 10% is usually a considerable loss, which promotes us to seek refinement.

To sum up, through the fluid level analysis:

• We determine an asymptotic optimal invitation policy: inviting customers by their $r\mu$ ranking in decreasing order until there is no idle server. Notice that the abandonment rate does not seem to be a factor in the fluid optimal policy.

• We proposed an equivalent threshold policy, namely, setting a threshold for only one partially invited customer class. Such a policy is not only easy to implement in practice but also performs better than the original fluid arrival rate control policy.

• Using simulation, we show that the fluid optimal policy by the controlling threshold

performs well also for stochastic environments. However, it is not optimal in such situations. Especially for small size systems, it usually has some loss. Therefore, we propose and analyze, in the next section, a refinement to the fluid policy.

Chapter 4

Analysis of System Dynamics

Referring to the fluid result, the threshold policy can be considered as a promising type of policy for invitations. Such a policy is easy to implement by setting a threshold to some of the customer classes. We consider a possible refinement of the fluid policy in which we optimize the threshold value. The fluid suggested that the threshold shall only affect the partially-invited customer class. All customers of higher-ranked classes should be invited and can be merged into one type. Hence, we need two classes: class 1 of high-ranking customers which we always invite and class 2 of low-ranking customers which we partially invite. We classify all candidate customers based on the $r\mu$ policy that we proposed in Chapter 3. We consider all the fully-invited customer classes together as high-ranking customers, denoted as Class 1, and the partially-invited customer class as low-ranking customers, denoted as Class 2. Thus we can reduce the multi-class model into the following two-class system (Figure 4.1):

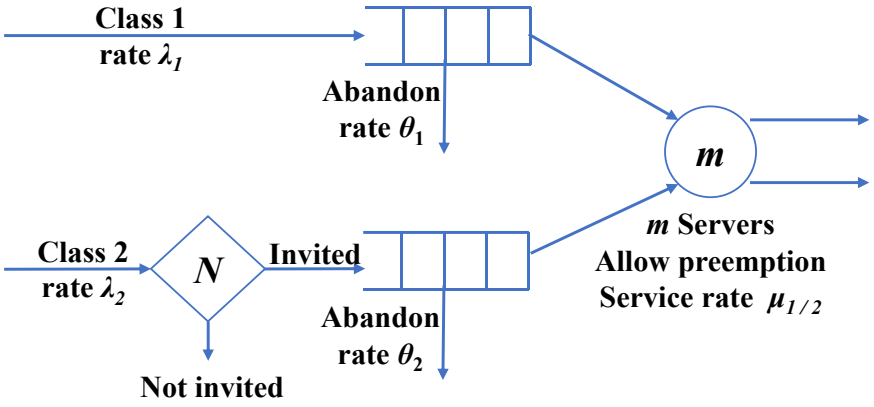


Figure 4.1: The simplified 2 class model of threshold policy

Two classes of customers arrive according to 2 independent Poisson processes at rates λ_1 and λ_2. Class i ($i = 1, 2$) customers are served with rates μ_i and lose waiting patience with rate θ_i. There are m identical servers in the system that serve both classes. Class 1 customers have a higher ranking, i.e., $r_1\mu_1$ is greater than $r_2\mu_2$. A predetermined threshold N controls the admission of Class 2 customers. Namely, the

system admits lower ranking customers only if *the total number of customers* in the system is lower than the threshold. Note that the original fluid policy optimization suggested that $N = m$ (no queues). After entering the system, if all servers are busy, the customers wait in a priority queue in accordance with their ranking. We assume that preemption is permitted, which means that the higher-ranked customer can interrupt a lower-ranked customer in service and get service first when the system is overloaded, and the interrupted service of the lower-ranked customer resumes at a later time when the service load is released. A Class i ($i = 1, 2$) customer's service completion brings rewards r_i to the system. The penalty on waiting time is c_i per unit of time waiting of customer of type i. We assume all the reward/cost parameters are positive.

Let $x_i(t)$, $z_i(t)$ and $q_i(t)$ denote the fluid contents of customer i in the system, service, and queue at time t, respectively. We aim to determine the optimal admission threshold, N, to maximize system revenue over an infinite horizon. Such revenue is influenced both by reward and penalty. Hence, we need to examine the effect of the threshold on different performance metrics such as the expected number of customers of class i in the system, $E(x_i)$, and the proportion of uninvited customers, $P(x_1 + x_2 \geq N)$. The dynamics of this model is captured by the following differential equations:

$$\begin{cases} \dot{x}_1(t) = \lambda_1 - \mu_1 z_1(t) - \theta_1 q_1(t) \\ \dot{x}_2(t) = I_{\{(x_1(t)+x_2(t))<N\}}\lambda_2 - \mu_2 z_2(t) - \theta_2 q_2(t) \\ q_1(t) = x_1(t) - z_1(t) \\ q_2(t) = x_2(t) - z_2(t) \\ z_1(t) = x_1(t) \wedge m \\ z_2(t) = x_2(t) \wedge (m - x_1(t))^+ \end{cases}, \tag{4.1}$$

where $I_{\mathbf{x}}$ is a 0-1 indicator that represents whether condition $\mathbf{x}$ is true or false, the symbol $\wedge$ is the minimal operator and $()^+$ means the larger value between the result inside brackets and zero. Those equations can be simplified into:

$$\begin{cases} \dot{x}_1(t) = \lambda_1 - \mu_1 (x_1(t) \wedge m) - \theta_1 (x_1(t) - m)^+ \\ \dot{x}_2(t) = I_{\{(x_1(t)+x_2(t))<N\}}\lambda_2 - \mu_2 \left(x_2(t) \wedge (m - x_1(t))^+\right) - \theta_2 \left(x_2(t) - (m - x_1(t))^+\right)^+ \end{cases}. \tag{4.2}$$

The above dynamics (4.2) is discontinuous on the right-hand side of $\dot{x}_2$ when $x_1 + x_2 = N$. We want to examine the long-term behavior of the system in the fluid level and determine the steady state of $\mathbf{x}(t) \triangleq [x_1(t), x_2(t)]^T$, denoted as $\bar{\mathbf{x}} = \lim_{t\to\infty} \mathbf{x}(t) = (\bar{x}_1, \bar{x}_2)$.

In order to analyze this long-term behavior, several definitions are needed. Consider a dynamic system that is represented by $\dot{\mathbf{x}} = f(\mathbf{x})$. Denote $\mathbf{x}(t)$ as the flow at time t. Slotine and Li (1991) defined equilibrium state (or point), stability (and instability), asymptotically stable and globally asymptotically stable in Definition 3.2 - 4, 6, as:

Definition 4.0.1. A state $\mathbf{x}^*$ is an *equilibrium state* (or *equilibrium point*) of the system if $f(\mathbf{x}^*) = \mathbf{0}$.

Denote in state-space ball $\mathbf{B}_R = \{\mathbf{x} | \|\mathbf{x}\| < R\}$, and sphere $\mathbf{S}_R = \{\mathbf{x} | \|\mathbf{x}\| = R\}$.

Definition 4.0.2. The equilibrium state is said to be *stable* if for any $R > 0$, there exists $r > 0$, such that if $\|\mathbf{x}(0)\| < r$, then $\|\mathbf{x}(t)\| < R$ for all $t \geq 0$. Otherwise, the equilibrium point is *unstable*.

Definition 4.0.3. The equilibrium point $\bar{\mathbf{x}}$ is *asymptotically stable* if it is stable, and if in addition there exists $r > 0$, such that $\|\mathbf{x}\| < r$ implies that $\mathbf{x}(t) \to \bar{\mathbf{x}}$ as $t \to \infty$.

Definition 4.0.4. If asymptotic stability holds for any initial states, the equilibrium point is said to be *globally* asymptotically stable.

Before focusing on this discontinuous system, we start by analyzing two extreme cases: a system that invites all low-ranked customers, i.e., no admission control, $N = \infty$, and a system that always applies admission control, i.e., $N = 1$.

4.1 Without the Threshold Policy

4.1.1 System Without Admission Control

When the system never implements admission control ($N = \infty$), its dynamics can be simplified into

$$\dot{\mathbf{x}} = \begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{cases} \lambda_1 - \mu_1 (x_1 \wedge m) - \theta_1 (x_1 - m)^+ \\ \lambda_2 - \mu_2 \left(x_2 \wedge (m - x_1)^+\right) - \theta_2 \left(x_2 - (m - x_1)^+\right)^+ \end{cases} . \tag{4.3}$$

The dynamics (4.3) can have three possible forms in different regions of the state space (see Figure 4.2):

Region A, $\Omega_A = \{(x_1, x_2) | x_1 \geq m\}$

$$\dot{\mathbf{x}} = \begin{cases} \lambda_1 - \mu_1 m - \theta_1 (x_1 - m) \\ \lambda_2 - \theta_2 x_2 \end{cases} , \tag{4.4}$$

Region B, $\Omega_B = \{(x_1, x_2) | x_1 < m \leq x_1 + x_2\}$

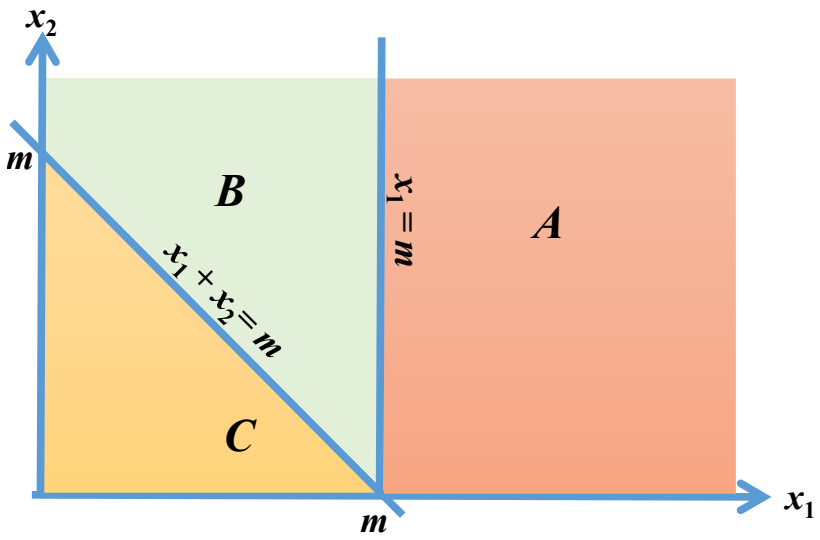
$$\dot{\mathbf{x}} = \begin{cases} \lambda_1 - \mu_1 x_1 \\ \lambda_2 - \mu_2 (m - x_1) - \theta_2 (x_1 + x_2 - m) \end{cases} , \tag{4.5}$$

Region C, $\Omega_C = \{(x_1, x_2) | x_1 + x_2 < m\}$

$$\dot{\mathbf{x}} = \begin{cases} \lambda_1 - \mu_1 x_1 \\ \lambda_2 - \mu_2 x_2 \end{cases} , \tag{4.6}$$

The equilibrium for each of the above three dynamics, denoted by $\bar{\mathbf{x}}_A, \bar{\mathbf{x}}_B$ and $\bar{\mathbf{x}}_C$,

Figure 4.2: Regions of system state



respectively, can be computed as

$$\bar{\mathbf{x}}_A = \begin{pmatrix} \frac{\lambda_1 - \mu_1 m}{\theta_1} + m \\ \frac{\lambda_2}{\theta_2} \end{pmatrix}, \bar{\mathbf{x}}_B = \begin{pmatrix} \frac{\lambda_1}{\mu_1} \\ \frac{\lambda_2 - (\mu_2 - \theta_2)(m - \lambda_1/\mu_1)}{\theta_2} \end{pmatrix}, \bar{\mathbf{x}}_C = \begin{pmatrix} \frac{\lambda_1}{\mu_1} \\ \frac{\lambda_2}{\mu_2} \end{pmatrix}. \tag{4.7}$$

These equilibria may not be admisible if the equilibria are out of the defined region. We denoted by $\bar{\mathbf{x}}^L$ the equilibrium of system (4.3).

Theorem 4.1. *The fluid (4.3) converges to the following globally asymptotically stable equilibrium:*

$$\bar{\mathbf{x}}^L = \begin{cases} \bar{\mathbf{x}}_A, & m \leq \lambda_1/\mu_1 \\ \bar{\mathbf{x}}_B, & \lambda_1/\mu_1 + \lambda_2/\mu_2 > m > \lambda_1/\mu_1 \\ \bar{\mathbf{x}}_C, & m \geq \lambda_1/\mu_1 + \lambda_2/\mu_2 \end{cases}. \tag{4.8}$$

In the first condition, the system is loaded even with just class 1; under the second condition, the system is underloaded if only high-ranked customers are ordered, but overloaded in general; when the third condition applies the system is underloaded. The most interesting case is the second one.

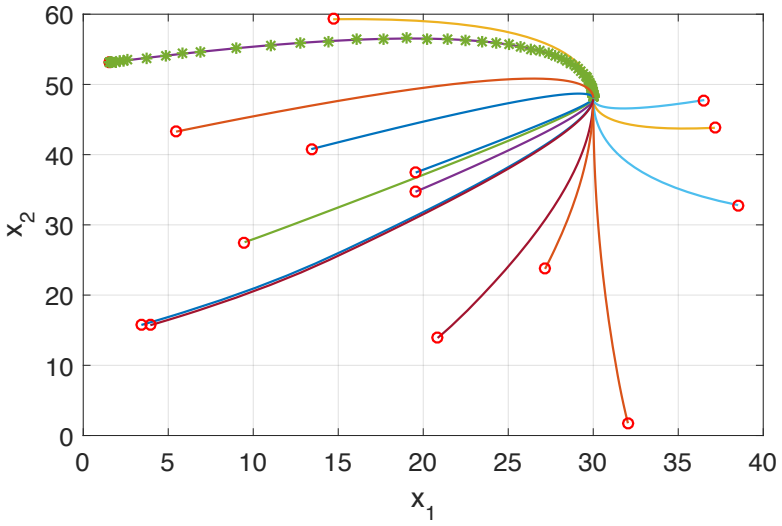
Before proving Theorem 4.1, we first want to qualitatively understand the flow in the phase space, in particular on the borders between the different regions. Because each of the dynamics (4.4, 4.5, 4.6) are linear and all their eigenvalues are negative, according to Lyapunov's stability theorem, the equilibria (4.7) are all asymptotically stable for each of the dynamics (Khalil 1996). Note also that the equilibria (4.7) are all asymptotically stable if there are no restrictions on the defined region, i.e., the dynamics are valid in the full phase space. In other words, the trajectory starting inside each region leaves that region after a finite time if the equilibrium doesn't reside in that region.

Because $\dot{x}_1$ is independent of x_2 and it is a linear equation, we have:

Lemma 4.1.1. *In the system (4.3), $x_1(t)$ monotonically decreases (increases).*

Therefore, in any situation, the flow of $\mathbf{x}$ can only cross the border between region A and region B in one direction. However, $x_2(t)$ may not be monotonic. In Figure 4.3, we simulate several trajectories of the system when $\lambda_1/\mu_1 + \lambda_2/\mu_2 > m > \lambda_1/\mu_1$ starting with random initial points. We find that all trajectories converge to the equilibrium (their intersection point). Note that the trajectory marked with '*' is of interest as it shows a case where x_2 is non-monotonic. It first increases for a while and then decreases to the equilibrium.

Figure 4.3: Trajectories of system without admission control



Thus, we examine the behavior round the border $x_1 + x_2 = m$ between Region B and C.

Lemma 4.1.2. *The trajectory $\mathbf{x}(t)$ can cross the border between region B and C at most twice.*

Proof. On $x_1 + x_2 = m$, the vector field $\dot{\mathbf{x}}$ degenerates into:

$$\dot{\mathbf{x}} = \begin{cases} \lambda_1 - \mu_1 x_1 \\ \lambda_2 - \mu_2 (m - x_1) \end{cases} . \tag{4.9}$$

Denote δ_m as the projection of vector field $\dot{\mathbf{x}}$ on the gradient of $x_1 + x_2 = m$. Then on the boundary $x_1 + x_2 = m$, one has

$$\delta_m = \dot{\mathbf{x}}^T \begin{pmatrix} 1 \\ 1 \end{pmatrix} = (\mu_2 - \mu_1) x_1 + (\lambda_1 + \lambda_2 - \mu_2 m) . \tag{4.10}$$

If $\delta_m > 0$, the flow crosses the border from region C to region B; and if $\delta_m < 0$, the flow crosses the border from region B to region C. In other words, the trajectory of $\mathbf{x}(t)$ crosses the borders multiple times if the sign of δ_m changes. However, from equation (4.10), δ_m is a linear function of x_1. According to Lemma 4.1.1, δ_m is also monotonic, namely, the sign of δ_m can change at most once. Hence, any trajectory of $\mathbf{x}(t)$ can cross between region B and C at most twice. □

Now we finally prove Theorem 4.1:

Proof. In the system with dynamics (4.3), from the above analysis, one knows that an equilibrium exists in only one of the regions A, B or C, denoted here by S. For each region that the equilibrium does not reside in, a trajectory starting in that region leaves that region after finite time (however with a possibility of returning). The reason is that the equilibria (4.7) are all asymptotically stable and do not lie in that region. It implies that the trajectory crosses one of the two borders $x_1 = m$ or $x_1 + x_2 = m$ in finite time. From Lemmas 4.1.1 and 4.1.2, the number of crossings is limited. Thus, after a sufficiently long time, the trajectory will reside in only one region and never leaves. We claim that region must be S, otherwise there may be a contradiction regarding the number of crossings.

Then the trajectory converges to the equilibrium in S because it is asymptotically stable. □

4.1.2 the System Always Applies Admission Control

When the admission control is always implemented ($N = 0$), system dynamics can be simplified to the following continuous form:

$$\dot{\mathbf{x}} = \begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{cases} \lambda_1 - \mu_1 \left(x_1(t) \wedge m\right) - \theta_1 (x_1(t) - m)^+ \\ -\mu_2 \left(x_2(t) \wedge (m - x_1(t))^+\right) - \theta_2 \big(x_2(t) - (m - x_1(t))^+\big)^+ \end{cases} . \quad (4.11)$$

Because of the abandonment, this system is always stable and converges to its equilibrium, denoted as $\bar{\mathbf{x}}^H = \left(\bar{x}_1^H, \bar{x}_2^H\right)$. Note that since there is no admission for class 2 customers, $\dot{x}_2(t) < 0, \forall t$, and after some finite time ε, x_2 will become 0 and from that time on, the system behaves like an Erlang-A queue for which the equilibrium is $(\lambda_1 - \mu_1 m)/\theta_1 + m$ if the system is overloaded and λ_1/μ_1 if the system is underloaded.

Theorem 4.2. *In system (4.11), the following equilibrium is globally asymptotically stable, i.e. the system fluid converges to:*

$$\bar{\mathbf{x}}^H = \begin{cases} \left((\lambda_1 - \mu_1 m)/\theta_1 + m, 0\right) & m < \lambda_1/\mu_1 \\ \left(\lambda_1/\mu_1, 0\right) & m \geq \lambda_1/\mu_1 \end{cases} . \quad (4.12)$$

Proof. $\bar{\mathbf{x}}^H$ defined by (4.12) is the equilibrium of system (4.11) since its a solution to $\dot{\mathbf{x}} = \mathbf{0}$. In order to show its globally asymptotical stability, the following Lyapunov

function candidate is used (Lyapunov 1992):

$$V(\mathbf{x}) = \left|x_1 - \bar{x}_1^H\right| + \left|x_2 - \bar{x}_2^H\right|. \tag{4.13}$$

We want to show that, $\forall \mathbf{x} \neq \bar{\mathbf{x}}^H$, $\dot{V}(\mathbf{x}) < 0$. The state space $\{x_1 \geqslant 0, x_2 \geqslant 0\}$ can be divided into two domains according to the value of m. In both cases:

Case A: $m < \lambda_1/\mu_1$. It has the following subcases:

1. $x_1 > \bar{x}_1^H, x_2 \geq \bar{x}_2^H$
   $\dot{V}(\mathbf{x}) = \dot{x}_1 + \dot{x}_2 = \lambda_1 - \mu_1 m - \theta_1 (x_1 - m) - \theta_2 x_2$
   $< \lambda_1 - \mu_1 m - \theta_1 \left(\bar{x}_1^H - m\right) - \theta_2 \bar{x}_2^H = 0$

2. $x_1 = \bar{x}_1^H, x_2 > \bar{x}_2^H$
   $\dot{V}(\mathbf{x}) = \dot{x}_2 = -\mu_2 \left(x_2 \wedge (m - x_1)^+\right) - \theta_2 \big(x_2 - (m - x_1)^+\big)^+ < 0$

3. $x_1 \geq \bar{x}_1^H, x_2 < \bar{x}_2^H$
   $\dot{V}(\mathbf{x}) = \dot{x}_1 - \dot{x}_2 = \lambda_1 - \mu_1 m - \theta_1 (x_1 - m) + \theta_2 x_2$
   $< \lambda_1 - \mu_1 m - \theta_1 \left(\bar{x}_1^H - m\right) + \theta_2 \bar{x}_2^H = 0$

4. $x_1 < \bar{x}_1^H, x_2 \geq \bar{x}_2^H$
   $\dot{V}(\mathbf{x}) = -\dot{x}_1 + \dot{x}_2$
   $= -\lambda_1 + \mu_1 (x_1 \wedge m) + \theta_1 (x_1 - m)^+ - \mu_2 \left(x_2 \wedge (m - x_1)^+\right) - \theta_2 \big(x_2 - (m - x_1)^+\big)^+$
   $< -\lambda_1 + \mu_1 \left(\bar{x}_1^H \wedge m\right) + \theta_1 \big(\bar{x}_1^H - m\big)^+ = 0$

5. $x_1 < \bar{x}_1^H, x_2 < \bar{x}_2^H$
   $\dot{V}(\mathbf{x}) = -\dot{x}_1 - \dot{x}_2$
   $= -\lambda_1 + \mu_1 (x_1 \wedge m) + \theta_1 (x_1 - m)^+ + \mu_2 \left(x_2 \wedge (m - x_1)^+\right) + \theta_2 \big(x_2 - (m - x_1)^+\big)^+$
   $< -\lambda_1 + \mu_1 \left(\bar{x}_1^H \wedge m\right) + \theta_1 \big(\bar{x}_1^H - m\big)^+ + \mu_2 \left(\bar{x}_2^H \wedge (m - x_1)^+\right) + \theta_2 \big(\bar{x}_2^H - (m - x_1)^+\big)^+$
   $= 0$

Case B $m \geq \lambda_1/\mu_1$. It has the following subcases:

1. $x_1 > \bar{x}_1^H, x_2 \geq \bar{x}_2^H$
   $\dot{V}(\mathbf{x}) = \dot{x}_1 + \dot{x}_2$
   $= \lambda_1 - \mu_1 (x_1 \wedge m) - \theta_1 (x_1 - m)^+ - \mu_2 \left(x_2 \wedge (m - x_1)^+\right) - \theta_2 \big(x_2 - (m - x_1)^+\big)^+$
   $\leq \lambda_1 - \mu_1 (x_1 \wedge m) - \theta_1 (x_1 - m)^+$
   $< \lambda_1 - \mu_1 \left(\bar{x}_1^H \wedge m\right) - \theta_1 \big(\bar{x}_1^H - m\big)^+ = 0$

2. $x_1 = \bar{x}_1^H, x_2 > \bar{x}_2^H$
   $\dot{V}(\mathbf{x}) = \dot{x}_2 = -\mu_2 \left(x_2 \wedge (m - x_1)^+\right) - \theta_2 \big(x_2 - (m - x_1)^+\big)^+ < 0$

3. $x_1 \geq \bar{x}_1^H, x_2 < \bar{x}_2^H$
   $\dot{V}(\mathbf{x}) = \dot{x}_1 - \dot{x}_2$
   $= \lambda_1 - \mu_1 (x_1 \wedge m) - \theta_1 (x_1 - m)^+ + \mu_2 \left(x_2 \wedge (m - x_1)^+\right) + \theta_2 \big(x_2 - (m - x_1)^+\big)^+$
   $< \lambda_1 - \mu_1 \cdot \bar{x}_1^H + \mu_2 \left(\bar{x}_2^H \wedge \left(m - \bar{x}_1^H\right)^+\right) + \theta_2 \big(\bar{x}_2^H\big)^+ = 0$

4. $x_1 < \bar{x}_1^H, x_2 \geq \bar{x}_2^H$
$\dot{V}(\mathbf{x}) = -\dot{x}_1 + \dot{x}_2$
$= -\lambda_1 + \mu_1 (x_1 \wedge m) + \theta_1 (x_1 - m)^+ - \mu_2 \left(x_2 \wedge (m - x_1)^+\right) - \theta_2 \left(x_2 - (m - x_1)^+\right)^+$
$< -\lambda_1 + \mu_1 \left(\bar{x}_1^H \wedge m\right) + \theta_1 \left(\bar{x}_1^H - m\right)^+ = 0$

5. $x_1 < \bar{x}_1^H, x_2 < \bar{x}_2^H$
$\dot{V}(\mathbf{x}) = -\dot{x}_1 - \dot{x}_2$
$= -\lambda_1 + \mu_1 (x_1 \wedge m) + \theta_1 (x_1 - m)^+ + \mu_2 \left(x_2 \wedge (m - x_1)^+\right) + \theta_2 \left(x_2 - (m - x_1)^+\right)^+$
$< -\lambda_1 + \mu_1 \left(\bar{x}_1^H \wedge m\right) + \theta_1 \left(\bar{x}_1^H - m\right)^+ + \mu_2 \left(\bar{x}_2^H \wedge (m - m)^+\right) + \theta_2 \left(\bar{x}_2^H\right)^+ = 0$

Thus in all cases, $\forall \mathbf{x} \neq \bar{\mathbf{x}}^H$, $\dot{V}(\mathbf{x}) < 0$. Hence, the system is globally asymptotically stable. □

4.2 Applying Threshold Policy

After plugging in the admission control, the system (4.2) becomes discontinuous on its right-hand side. Therefore, we fit our model into Filippov's framework (Filipov 1988) for analysis.

The system state space, $\left\{\mathbb{R}_+^2 : x_1 \geq 0, x_2 \geq 0\right\}$, is divided by the switching boundary s:

$$s \triangleq \{x : x_1 + x_2 - N = 0\} \tag{4.14}$$

into two regions: $\mathcal{R}^L$ and $\mathcal{R}^H$, where $\mathcal{R}^L \triangleq \{(x_1, x_2) \,|x_1 + x_2 - N < 0\}$ and $\mathcal{R}^H \triangleq \{(x_1, x_2) \,|x_1 + x_2 - N > 0\}$. We denote the fluid function in regions $\mathcal{R}^L$ and $\mathcal{R}^H$ by $\mathbf{f}^L(\mathbf{x})$ and $\mathbf{f}^H(\mathbf{x})$, respectively. They are continuous and piecewise smooth ODE:

$$\begin{aligned} \mathbf{f}^H(\mathbf{x}) &= \begin{pmatrix} \lambda_1 - \mu_1 (x_1 \wedge m) - \theta_1 (x_1 - m)^+, \\ -\mu_2 \left(x_2 \wedge (m - x_1)^+\right) - \theta_2 \left(x_2 - (m - x_1)^+\right)^+ \end{pmatrix}; \\ \mathbf{f}^L(\mathbf{x}) &= \begin{pmatrix} \lambda_1 - \mu_1 (x_1 \wedge m) - \theta_1 (x_1 - m)^+, \\ \lambda_2 - \mu_2 \left(x_2 \wedge (m - x_1)^+\right) - \theta_2 \left(x_2 - (m - x_1)^+\right)^+ \end{pmatrix}. \end{aligned} \tag{4.15}$$

By applying the Filippov theory, the dynamics on the switching boundary s can be defined as a convex inclusion:

$$\dot{\mathbf{x}} = \begin{cases} \mathbf{f}^H(\mathbf{x}) & \mathbf{x} \in \mathcal{R}^H \\ \varphi \mathbf{f}^H(\mathbf{x}) + (1 - \varphi) \mathbf{f}^L(\mathbf{x}), \varphi \in [0, 1] & \mathbf{x} \in s \\ \mathbf{f}^L(\mathbf{x}) & \mathbf{x} \in \mathcal{R}^L \end{cases}, \tag{4.16}$$

which we write explicitly when $\mathbf{x} \in s$,

$$\dot{\mathbf{x}} = \begin{cases} \lambda_1 - \mu_1 (x_1 \wedge m) - \theta_1 (x_1 - m)^+ \\ (1 - \varphi) \lambda_2 - \mu_2 \left(x_2 \wedge (m - x_1)^+\right) - \theta_2 \left(x_2 - (m - x_1)^+\right)^+ \end{cases}, \varphi \in [0, 1]. \tag{4.17}$$

By using $\bar{\mathbf{x}}^L = (\bar{x}_1^L, \bar{x}_2^L)$ and $\bar{\mathbf{x}}^H = (\bar{x}_1^H, \bar{x}_2^H)$ that are defined in Theorem 4.2 and

4.1, respectively, the equilibrium of system (4.16) can be stated as follows. Note that $\bar{x}_2^H = 0$ and $\bar{x}_1^L = \bar{x}_1^H$. Thus, $\bar{x}_1^H + \bar{x}_2^H \leq \bar{x}_1^L + \bar{x}_2^L$.

Theorem 4.3. *In system (4.16), the following equilibrium is globally asymptotically stable, i.e. the system fluid converges to:*

$$\bar{\mathbf{x}} = \begin{cases} \bar{\mathbf{x}}^L & \bar{x}_1^L + \bar{x}_2^L \leq N \\ \alpha \bar{\mathbf{x}}^H + (1-\alpha)\, \bar{\mathbf{x}}^L & \bar{x}_1^H + \bar{x}_2^H < N < \bar{x}_1^L + \bar{x}_2^L \\ \bar{\mathbf{x}}^H & \bar{x}_1^H + \bar{x}_2^H \geq N \end{cases}, \tag{4.18}$$

where

$$\alpha = \begin{cases} \frac{\lambda_1\theta_2 + \lambda_2\theta_1 + m\theta_1\theta_2 - m\mu_1\theta_2 - N\theta_1\theta_2}{\lambda_2\theta_1}, & m \leq \lambda_1/\mu_1 \\ \frac{\lambda_1\mu_2 + \lambda_2\mu_1 + m\mu_1\theta_2 - m\mu_1\mu_2 - N\mu_1\theta_2}{\lambda_1\mu_2 + \lambda_2\mu_1 + m\mu_1\theta_2 - \lambda_1\theta_2 - m\mu_1\mu_2}, & \lambda_1/\mu_1 + \lambda_2/\mu_2 > m > \lambda_1/\mu_1 \\ \frac{\lambda_1\mu_1\mu_2 + \lambda_2\mu_1{}_2 - N\mu_1\mu_2}{\lambda_2\mu_1}, & m \geq \lambda_1/\mu_1 + \lambda_2/\mu_2 \end{cases}.$$

Because the dynamics (4.16) is discontinuous on border s, we use the following variables to investigate the behavior of system flow around s. Denote $\delta_N(\mathbf{x}) \equiv \dot{\mathbf{x}}^T \nabla s$ as the projection of vector field $\dot{\mathbf{x}}$ on the gradient of s. It is the angle between gradient and system flow. The angle is less than 90 degrees if $\delta_N(\mathbf{x}) > 0$ and larger than 90 degrees if $\delta_N(\mathbf{x}) < 0$. When this value equals 0, the system flow is perpendicular to the gradient of s. Thus, $\delta_N(\mathbf{x})$ indicates whether the flow moves toward or away from s when approached from $\mathcal{R}^H, s$ or $\mathcal{R}^L$. This measure is often referred to as the Lie derivative of s along the field defined by (4.16). In order to evaluate $\delta_N(\mathbf{x})$ on s, we denote the following two values:

$$\begin{cases} \delta^H(x_1) \equiv \nabla s^T \mathbf{f}^H(x_1, N - x_1) \\ \delta^L(x_1) \equiv \nabla s^T \mathbf{f}^L(x_1, N - x_1) \end{cases},$$

and let $\delta(x_1, \varphi)$ be the convex combination of $\delta^H(x_1)$ and $\delta^L(x_1)$

$$\delta(x_1, \varphi) \equiv \varphi \delta^H(x_1) + (1-\varphi)\, \delta^L(x_1), \varphi \in [0,1].$$

Note that $\delta(x_1, 0) = \delta^L(x_1), \delta(x_1, 1) = \delta^H(x_1)$. We evaluate the value of $\delta(x_1, \varphi)$ at a special point on s, namely, $x_1 = \bar{x}_1$.

$$\begin{aligned} \delta(\bar{x}_1, \varphi) &= \lambda_1 - \mu_1(\bar{x}_1 \wedge m) - \theta_1(\bar{x}_1 - m)^+ + (1-\varphi)\lambda_2 \\ &\quad -\mu_2\left((N-\bar{x}_1)^+ \wedge (m-\bar{x}_1)^+\right) - \theta_2\left((N-\bar{x}_1)^+ - (m-\bar{x}_1)^+\right)^+ \\ &= (1-\varphi)\lambda_2 - \mu_2\left((N-\bar{x}_1)^+ \wedge (m-\bar{x}_1)^+\right) - \theta_2\left((N-\bar{x}_1)^+ - (m-\bar{x}_1)^+\right)^+ \end{aligned}.$$

We analyze the value of $\delta(\bar{x}_1, \varphi)$ in the following cases:

1. $m \leq \lambda_1/\mu_1, x_1 = \bar{x}_1 = (\lambda_1 - \mu_1 m)/\theta_1 + m$
   $\delta(\bar{x}_1, \varphi) = (1-\varphi)\lambda_2 - \theta_2(N - \bar{x}_1)^+$
   - $N \leq \bar{x}_1$

$\delta(\bar{x}_1, \varphi) = (1-\varphi)\,\lambda_2 \geq 0$.
Note that the equality holds only when $\varphi = 1$.

- $\bar{x}_1 < N < \bar{x}_1 + \lambda_2/\theta_2$
  $0 < \theta_2\,(N - \bar{x}_1)\,/\lambda_2 < 1$.
  When $\varphi = 1 - \theta_2\,(N - \bar{x}_1)\,/\lambda_2$, $\delta(\bar{x}_1, \varphi) = 0$ and $\varphi \in (0, 1)$.
- $N \geq \bar{x}_1 + \lambda_2/\theta_2$
  $\delta(\bar{x}_1, \varphi) \leq (1-\varphi)\,\lambda_2 - \theta_2\,(\lambda_2/\theta_2) = -\varphi\lambda_2 \leq 0$.
  Note that the equality holds only when $\varphi = 0$.

2. $\lambda_1/\mu_1 + \lambda_2/\mu_2 > m > \lambda_1/\mu_1, x_1 = \bar{x}_1 = \lambda_1/\mu_1$

   - $N \leq \bar{x}_1$
     $\delta(\bar{x}_1, \varphi) = (1-\varphi)\,\lambda_2 \geq 0$.
     Note that the equality holds only when $\varphi = 1$.
   - $\bar{x}_1 < N \leq m$
     $0 < \mu_2\,(N - \bar{x}_1)\,/\lambda_2 < 1$.
     When $\varphi = 1 - \mu_2\,(N - \bar{x}_1)\,/\lambda_2$, $\delta(\bar{x}_1, \varphi) = 0$ and $\varphi \in (0, 1)$.
   - $m < N < \bar{x}_1 + (\lambda_2 - (\mu_2 - \theta_2)\,(m - \lambda_1/\mu_1))\,/\theta_2$
     $0 < (\mu_2\,(m - \bar{x}_1) + \theta_2\,(N - m))/\lambda_2 < 1$.
     When $\varphi = 1 - (\mu_2\,(m - \bar{x}_1) + \theta_2\,(N - m))\,/\lambda_2$, $\delta(\bar{x}_1, \varphi) = 0$ and $\varphi \in (0, 1)$.
   - $N \geq \bar{x}_1 + (\lambda_2 - (\mu_2 - \theta_2)\,(m - \lambda_1/\mu_1))\,/\theta_2 > m$
     $\delta(\bar{x}_1, \varphi) \leq (1-\varphi)\,\lambda_2 - \mu_2\,(m - \bar{x}_1) - \theta_2\,((\lambda_2 - \mu_2\,(m - \lambda_1/\mu_1))\,/\theta_2) \leq 0$.
     Note that the equality holds only when $\varphi = 0$.

3. $m \geq \lambda_1/\mu_1 + \lambda_2/\mu_2, x_1 = \bar{x}_1 = \lambda_1/\mu_1$

   - $N \leq \bar{x}_1$
     $\delta(\bar{x}_1, \varphi) = (1-\varphi)\,\lambda_2 \geq 0$.
     Note that the equality holds only when $\varphi = 1$.
   - $\bar{x}_1 < N < \bar{x}_1 + \lambda_2/\mu_2$
     $0 < \mu_2\,(N - \bar{x}_1)\,/\lambda_2 < 1$.
     When $\varphi = 1 - \mu_2\,(N - \bar{x}_1)\,/\lambda_2$, $\delta(\bar{x}_1, \varphi) = 0$ and $\varphi \in (0, 1)$.
   - $\bar{x}_1 + \lambda_2/\mu_2 \leq N \leq m$
     $0 < (\mu_2\,(m - \bar{x}_1) + \theta_2\,(N - m))/\lambda_2 < 1$.
     $\delta(\bar{x}_1, \varphi) = (1-\varphi)\,\lambda_2 - \mu_2\,(N - \bar{x}_1) \leq (1-\varphi)\,\lambda_2 - \mu_2\,(\lambda_2/\mu_2) = -\varphi\lambda_2 \leq 0$.
     Note that the equality holds only when $\varphi = 0$.
   - $N > m$
     $\delta(\bar{x}_1, \varphi) = (1-\varphi)\,\lambda_2 - \mu_2\,(m - \bar{x}_1) - \theta_2\,(N - m) \leq 0$.
     Note that the equality holds only when $\varphi = 0$.

The above results depend on system parameters, which are summarized by the following cases:

1. $m \leq \lambda_1/\mu_1$

   (a) $\dfrac{\lambda_2}{\theta_2} + \dfrac{\lambda_1 - (\mu_1 - \theta_1)\, m}{\theta_1} \leq N$

   (b) $\dfrac{\lambda_1 - (\mu_1 - \theta_1)\, m}{\theta_1} < N < \dfrac{\lambda_2}{\theta_2} + \dfrac{\lambda_1 - (\mu_1 - \theta_1)\, m}{\theta_1}$

   (c) $\dfrac{\lambda_1 - (\mu_1 - \theta_1)\, m}{\theta_1} \geq N$

2. $\lambda_1/\mu_1 + \lambda_2/\mu_2 > m > \lambda_1/\mu_1$

   (a) $\dfrac{\lambda_2 - \mu_2\,(m - \lambda_1/\mu_1)}{\theta_2} + m \leq N$

   (b) $\dfrac{\lambda_1}{\mu_1} < N < \dfrac{\lambda_2 - \mu_2\,(m - \lambda_1/\mu_1)}{\theta_2} + m$

   (c) $\frac{\lambda_1}{\mu_1} \geq N$

3. $m \geq \lambda_1/\mu_1 + \lambda_2/\mu_2$

   (a) $\dfrac{\lambda_1}{\mu_1} + \dfrac{\lambda_2}{\mu_2} \leq N$

   (b) $\dfrac{\lambda_1}{\mu_1} < N < \dfrac{\lambda_1}{\mu_1} + \dfrac{\lambda_2}{\mu_2}$

   (c) $\dfrac{\lambda_1}{\mu_1} \geq N$

It can be evaluated that:

In all Case (a), $\forall \varphi \in [0,1]$, one has $\delta\,(\bar{x}_1, \varphi) < 0$.

In all Case (b), $\exists \varphi \in (0,1)$, such that $\delta\,(\bar{x}_1, \varphi) = 0$.

In all Case (c), $\forall \varphi \in [0,1]$, one has $\delta\,(\bar{x}_1, \varphi) > 0$.

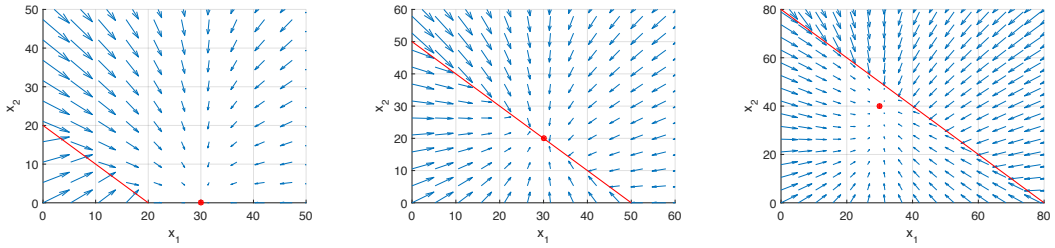
Note that we consider the breakpoint into Case (a) or Case (c). Because on all breakpoints, $\delta\,(\bar{x}_1, \varphi) = 0$ only if $\varphi = 0$ or 1. In this case, system dynamics on the border s has the same function as $\mathbf{f}^H(\mathbf{x})$ or $\mathbf{f}^L(\mathbf{x})$. Also, s can be considered as it belongs to $\mathcal{R}^H$ or $\mathcal{R}^L$. Hence the flow in the breakpoint has the same properties as either Case (a) or Case (c).

Case 2 is the most interesting case. We use the phase portrait (Figure 4.4) to sketch the system flow. The boundary s is plotted by a red line. The equilibrium is marked by a dot. The arrows represent the derivatives of the system states. In Figure 4.4a, the equilibrium doesn't lie on the switching line. The arrows appear to penetrate the switch line. In Figure 4.4b, the equilibrium is on s and all the arrows in $\mathcal{R}^L$ and $\mathcal{R}^H$ point to s in a small region around s. Figure 4.4c is similar to Figure 4.4a, i.e., the direction of arrows are similar to Case 2(a), penetrating the switch line, from the region without equilibrium ($\mathcal{R}^H$) to the region with equilibrium ($\mathcal{R}^L$).

Therefore, in order to prove the asymptotical stability, we are more interested in the system dynamics in a vertical stripe around the equilibrium. Meanwhile, in all

Figure 4.4: Phase portraits of Case 2

(a) Case 2(a) (b) Case 2(b) (c) Case 2(c)



the above cases, x_1 shows monotonicity. Because $\dot{x}_1$ is always independent of x_2 and its dynamics are piecewise-linear in x_1, similar to the analysis in the previous section, $\lim_{t\to+\infty} x_1(t)$ exists. In other words, we have the following:

Lemma 4.2.1. *In the system defined by Equation (4.2),* $\forall\varepsilon > 0$, $\exists T > 0$, *such that* $\mathbf{x}(t > T) \in \mathcal{R}_\varepsilon = \{(x_1, x_2) \,|x_1 \in (\bar{x}_1 - \varepsilon, \bar{x}_1 + \varepsilon), \varepsilon > 0\}$.

As before, x_1 has an equilibrium that depends on the level of load of class 1 customers only.

$$\bar{x}_1 = \begin{cases} (\lambda_1 - \mu_1 m)/\theta_1 + m & m \le \lambda_1/\mu_1 \\ \lambda_1/\mu_1 & m > \lambda_1/\mu_1 \end{cases} . \tag{4.19}$$

Thus, it is enough to consider the states inside region $\mathcal{R}_\varepsilon$. To that end, we first examine the states close to the switching boundary s for Cases (a), (b), and (c).

Lemma 4.2.2. *In system (4.2), if* $\forall\varphi \in [0,1]$, $\delta(\bar{x}_1, \varphi) > 0$ *(or* < 0 *), then* $\exists\varepsilon > 0$, *such that* $\forall x_1 \in (\bar{x}_1 - \varepsilon, \bar{x}_1 + \varepsilon)$, $\delta(x_1, \varphi) > 0$ *(or,* $\delta(x_1, \varphi) < 0$ *) holds for* $\forall\varphi \in [0,1]$.

Proof. We prove the case $\delta(x_1, \varphi) > 0$. According to the definition, $\delta(x_1, \varphi) = \varphi\delta^H(x_1) + (1-\varphi)\delta^L(x_1)$. Since $\forall\varphi \in [0,1]$: $\delta(\bar{x}_1, \varphi) > 0$, $\delta^H(\bar{x}_1) > 0$ and $\delta^L(\bar{x}_1) > 0$. By continuity, there always exists $\varepsilon^H > 0$ and $\varepsilon^L > 0$, such that $\delta^H(x_1) > 0$ for $x_1 \in (\bar{x}_1 - \varepsilon^H, \bar{x}_1 + \varepsilon^H)$, and $\delta_2^L(x_1) > 0$ for $x_1 \in (\bar{x}_1 - \varepsilon^L, \bar{x}_1 + \varepsilon^L)$. Taking $\varepsilon = \min(\varepsilon^L, \varepsilon^H)$, one has $\delta_2^H(x_1) > 0$ and $\delta^L(x_1) > 0$ for $x_1 \in (\bar{x}_1 - \varepsilon, \bar{x}_1 + \varepsilon)$. Furthermore, the $\delta(x_1, \varphi) > 0$ since it is a convex combination of $\delta^H(x_1)$ and $\delta^L(x_1)$. □

Under the conclusion of Lemma 4.2.1, Lemma 4.2.2 implies that for all Cases (a) and (c), there always exists a small neighborhood of $(\bar{x}_1, N - \bar{x}_1)^T$, where the system trajectory can only cross s in one determined direction. However, in Case (b), s becomes uncrossable:

Lemma 4.2.3. *In system (4.2), if* $\exists\varphi_0 \in (0,1)$ *such that* $\delta(\bar{x}_1, \varphi_0) = 0$, *then* $\exists\varepsilon > 0$ *such that for all* $x_1 \in (\bar{x}_1 - \varepsilon, \bar{x}_1 + \varepsilon)$, $\delta(x_1, \varphi) = 0$ *has a solution with* $\varphi \in (0,1)$.

Proof. From system dynamics (4.2), one has $\dfrac{\partial \delta(x_1,\varphi)}{\partial \varphi} = -\lambda_2 \neq 0$. Then the existence of φ as an explicit function of x_1 is guaranteed by the implicit function theorem (Munkres 1997), i.e., there exist $\varepsilon > 0$ for $x_1 \in (\bar{x}_1 - \varepsilon, \bar{x}_1 + \varepsilon)$ such that $\varphi(x_1)$ satisfies $\delta(x,\varphi(x_1)) = 0$. Note also that $\varphi(\bar{x}_1) = \varphi_0$ and $\varphi(x)$ is continuous at $\bar{x}_1$; thus $\varphi(x_1) \in (0,1)$ for all x_1 in a sufficiently small neighborhood of $\bar{x}_1$. □

Reference Bernardo et al. (2008) defines such a region on s where $\delta(x_1,\varphi) = 0, \varphi \in (0,1)$, as a sliding region. Therefore, Lemma 4.2.3 shows that in Case (b), $\exists \varepsilon > 0$, for which $x_1 \in (\bar{x}_1 - \varepsilon, \bar{x}_1 + \varepsilon)$ is a sliding region. If the system has a sliding region around $\bar{x}_1$, the system flow converges to this region in finite time.

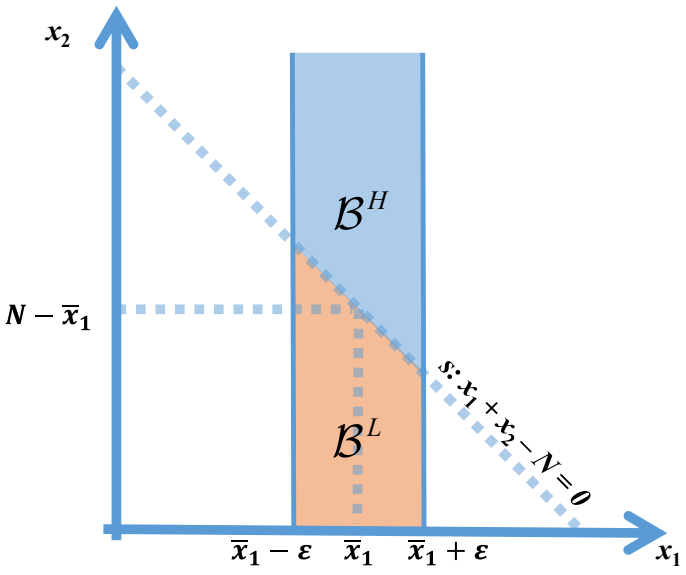
Lemma 4.2.4. *In system (4.2), if a sliding region exists, then* $\exists \varepsilon > 0$, *such that* $\forall \mathbf{x} \in \mathcal{B}^H$ *(see Figure 4.5),* $\delta_N(\mathbf{x}) < 0$ *where*

$$\mathcal{B}^H = \{(x_1,x_2)\,|x_1 \in (\bar{x}_1 - \varepsilon, \bar{x}_1 + \varepsilon)\} \cap \mathcal{R}^H,$$

and $\forall \mathbf{x} \in \mathcal{B}^L$, $\delta_N(\mathbf{x}) > 0$, *where*

$$\mathcal{B}^L = \{(x_1,x_2)\,|x_1 \in (\bar{x}_1 - \varepsilon, \bar{x}_1 + \varepsilon)\} \cap \mathcal{R}^L.$$

Figure 4.5: Regions of system state when sliding region exists



Proof. According to Lemma 4.2.3, a sliding region exists in Case (b). One can easily evaluate that

$$\begin{cases} \delta^H(\bar{x}_1, N - \bar{x}_1) = \gamma^H < 0 \\ \delta^L(\bar{x}_1, N - \bar{x}_1) = \gamma^L > 0 \end{cases}.$$

Here γ^H and γ^L are functions of parameters, where

$$\gamma^H = \begin{cases} -\theta_2 \left(N - \bar{x}_1\right) & m \leq \lambda_1/\mu_1 \\ -\mu_2((N \wedge m) - \bar{x}_1)^+ - \theta_2(N-m)^+ & \lambda_1/\mu_1 + \lambda_2/\mu_2 > m > \lambda_1/\mu_1 \\ -\mu_2 \left(N - \bar{x}_1\right) & m \geq \lambda_1/\mu_1 + \lambda_2/\mu_2 \end{cases},$$

$$\gamma^L = \begin{cases} \lambda_2 - \theta_2 \left(N - \bar{x}_1\right) & m \leq \lambda_1/\mu_1 \\ \lambda_2 - \mu_2((N \wedge m) - \bar{x}_1)^+ - \theta_2(N-m)^+ & \lambda_1/\mu_1 + \lambda_2/\mu_2 > m > \lambda_1/\mu_1 \\ \lambda_2 - \mu_2 \left(N - \bar{x}_1\right) & m \geq \lambda_1/\mu_1 + \lambda_2/\mu_2 \end{cases}.$$

Hence, any $x \in \mathcal{B}^H$ can be written as:

$$x = \begin{pmatrix} \bar{x}_1 + \Delta x_1^H \\ N - \bar{x}_1 + \Delta x_2^H \end{pmatrix},$$

and $\left|\Delta x_1^H\right| < \varepsilon$ and $\Delta x_2^H > -\Delta x_1^H$. Similarly, $\forall x \in \mathcal{B}^L$ can be written as

$$x = \begin{pmatrix} \bar{x}_1 + \Delta x_1^L \\ N - \bar{x}_1 + \Delta x_2^L \end{pmatrix},$$

where $\left|\Delta x_1^L\right| < \varepsilon$ and $\Delta x_2^L < -\Delta x_1^L$. Note also that $\delta_N(\mathbf{x})$ is a linear function of x_1 and x_2 of the form

$$\delta_N(\mathbf{x}) = \begin{cases} a^H x_1 + b^H x_2 + c^H \triangleq \delta_N^H(\mathbf{x}), x \in \mathcal{B}^H \\ a^L x_1 + b^L x_2 + c^L \triangleq \delta_N^L(\mathbf{x}), x \in \mathcal{B}^L \end{cases},$$

where a^H, a^L, b^H, b^L, c^H, and c^L represent constants and $b^H, b^L < 0$. Thus, $\forall \mathbf{x} \in \mathcal{B}^H$,

$$\begin{array}{l} \delta_N^H(\mathbf{x}) = a^H x_1 + b^H x_2 + c^H = a^H \Delta x_1^H + b^H \Delta x_2^H + \gamma^H \\ < \left(\left|a^H\right| + \left|b^H\right|\right)\varepsilon + \gamma^H \end{array},$$

and $\forall x \in \mathcal{B}^L$,

$$\begin{array}{l} \delta_N^L(\mathbf{x}) = a^L x_1 + b^L x_2 + c = a^L \Delta x_1^L + b^L \Delta x_2^L + \gamma^L \\ > -\left(\left|a^L\right| + \left|b^L\right|\right)\varepsilon + \gamma^L \end{array}.$$

Thus, one can choose

$$\varepsilon = \min\left(\frac{-\gamma^H}{2\left(|a^H| + |b^H|\right)}, \frac{\gamma^L}{2\left(|a^L| + |b^L|\right)}\right),$$

such that

$$\begin{cases} \delta_N(\mathbf{x}) = \delta_N^H(\mathbf{x}) < \dfrac{\gamma^H}{2} < 0, \mathbf{x} \in \mathcal{B}^H \\ \delta_N(\mathbf{x}) = \delta_N^L(\mathbf{x}) > \dfrac{\gamma^L}{2} > 0, \mathbf{x} \in \mathcal{B}^L \end{cases}.$$

. □

Lemma 4.2.4 shows that, in all Case (b), there exists an $\varepsilon > 0$, such that $\forall \mathbf{x} \in \{(x_1, x_2) \,|x_1 \in (\bar{x}_1 - \varepsilon, \bar{x}_1 + \varepsilon)\}$, the system state converges to s in finite time.

Now we move to prove Theorem 4.3:

Proof. The system can be categorized into two types: there exists /does not exist a sliding region on s which contains $(\bar{x}_1, N - \bar{x}_1)$.

A. The sliding region does not exist

From Lemma 4.2.2, we know that there exists $\varepsilon_{NSR} > 0$; all the trajectories inside the region $\mathcal{B}_{NSR} = \{(x_1, x_2) \,|x_1 \in (\bar{x}_1 - \varepsilon_{NSR}, \bar{x}_1 + \varepsilon_{NSR})\}$, can only pass through s in one direction. According to Lemma 4.2.1, any trajectory arrives to $\mathcal{B}_{NSR}$ in finite time. Similarly to the proof of Theorems 4.2 and 4.1, all the trajectories inside $\mathcal{B}_{NSR}$ converge to $\bar{\mathbf{x}}^H$ ($\bar{\mathbf{x}}^L$). Therefore, in all Case (a), $\bar{\mathbf{x}}^H$ is a globally asymptotically stable equilibrium, and in all Case (c), $\bar{\mathbf{x}}_L$ is a globally asymptotically stable equilibrium.

B. The sliding region exists

From Lemma 4.2.3, we know that there exists $\varepsilon_{SR} > 0$, such that $\mathcal{B}_{SR} \cap S$ is a sliding region, where $\mathcal{B}_{SR} = \{(x_1, x_2) \,|x_1 \in (\bar{x}_1 - \varepsilon_{SR1}, \bar{x}_1 + \varepsilon_{SR1})\}$. Meanwhile, from Lemma 4.2.1, we know that all the trajectories reach $\mathcal{B}_{SR}$ in finite time. Lemma 4.2.4 maintains that all the trajectories inside $\mathcal{B}_{SR}$ reach s in finite time. In addition, on the sliding region, the evolution of system state is always along s, namely, $x_1 + x_2 = N$ always holds. Since x_1 converges to $\bar{x}_1$, all the trajectories will converge to $(\bar{x}_1, N - \bar{x}_1)$. The value of α can be evaluated to satisfy $\alpha\bar{\mathbf{x}}^H + (1-\alpha)\,\bar{\mathbf{x}}^L = (\bar{x}_1, N - \bar{x}_1)^T$. Therefore, in all Case (b), $\alpha\bar{\mathbf{x}}^H + (1-\alpha)\,\bar{\mathbf{x}}^L$ is the globally asymptotically stable equilibrium. □

To sum up, the equilibria defined by Theorem 4.3 depends on all system parameters. They can be illustrated in the space of the threshold N and capacity m, shown in Figure 4.6. There are in total three striped horizon regions separated by solid lines, which are referred to as upper zone (Case (a)), middle zone (Case (b)) and lower zone (Case (c)). Inside the middle region, there exists a sliding region and the equilibrium is $\alpha\bar{\mathbf{x}}^H + (1-\alpha)\,\bar{\mathbf{x}}^L$. In the upper region, the trajectory of the system state can only move towards $\mathcal{R}^L$ when its x_1 is close to $\bar{x}_1$, and converges to the equilibrium, $\bar{\mathbf{x}}^L$. In the lower region, the state trajectory has to pass from s to $\mathcal{R}^H$, after a finite time, and converge to the equilibrium $\bar{\mathbf{x}}^H$.

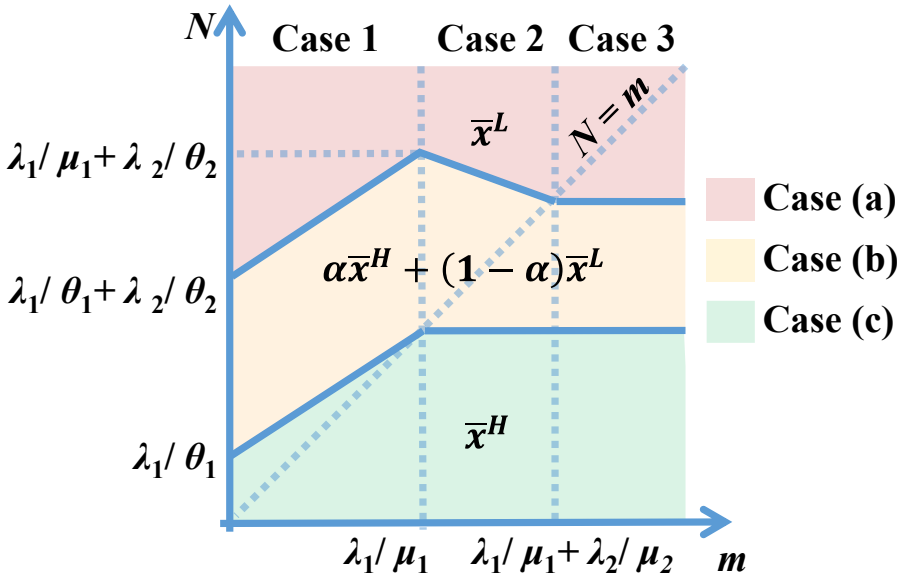
Furthermore, by using Theorem 4.3 and the results of Bernardo et al. (2008), on the fluid level, we can obtain:

Corollary 4.4. *The proportion of time using admission control in the system defined by equations 4.16, is given by:*

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T I_{\{(x_1(t)+x_2(t))<N\}} dt = \begin{cases} 0 & \bar{x}_1^L + \bar{x}_2^L \leq N \\ \alpha & \bar{x}_1^H + \bar{x}_2^H < N < \bar{x}_1^L + \bar{x}_2^L \\ 1 & \bar{x}_1^H + \bar{x}_2^H \geq N \end{cases}. \tag{4.20}$$

This proportion can help us approximate the probability of implementing admission

Figure 4.6: Equilibrium of various threshold values and server numbers



control for the original stochastic model (4.1), as:

$$P\left(Admission\right)=P\left(x_1+x_2\geq N\right)\approx\lim_{T\to\infty}\frac{1}{T}\int_0^T I_{\{(x_1(t)+x_2(t))<N\}}dt. \tag{4.21}$$

To sum up, through the system dynamics analysis:

• We propose a threshold policy that works only on the partially invited customer class according to a fluid policy, so as to try to find a refinement of a fluid policy. We simplify the original multi-class system into a two-class system with threshold policy that controls the admission of lower-ranking customers.

• We use the definition to find and prove the fluid globally stable equilibria of the number of customers (both classes) in the system, for the simplified two-class system. The globally stable equilibria highly depends on system parameters and the threshold; especially, in some cases, where the equilibria are found on sliding regions.

• Using the equilibria of customer numbers, we approximate the probability of implementing admission control by the stochastic two-class system.

Chapter 5

Discussion

5.1 Fluid Equilibrium

In this section, we aim to understand how the value of equilibrium is affected by system parameters. The equilibrium defined by Theorem 4.3 can be depicted using a bifurcation diagram (Figure 5.1). In our system, the values of $\bar{x}^L$ and $\bar{x}^H$ determine two breakpoints on N that separate the equilibrium into three cases – (a), (b) and (c). Both $\bar{x}^L$ and $\bar{x}^H$ depend only on system parameters as defined by Theorems 4.1 and 4.2. In Cases (a) and (c), the equilibrium is obtained as if the system that always uses / never uses admission control, respectively. In Case (b), the equilibrium depends on the value of the threshold, N, which is linearly increasing in N from $\bar{x}^H$ to $\bar{x}^L$.

Figure 5.1: Bifurcation diagram of Cases 1, 2 and 3 as a function of N

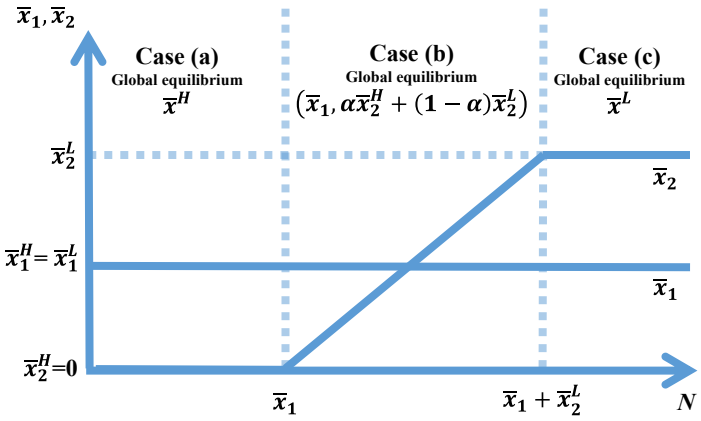
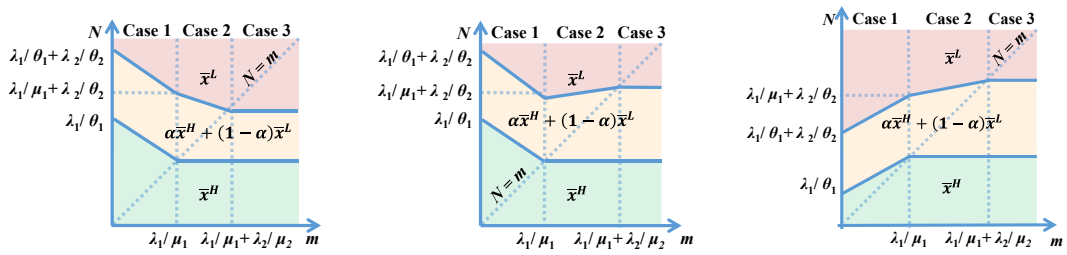


Figure 4.6 showed the distribution of equilibrium for different combinations of N and m; that diagram is actually for the cases of $\theta_1 > \mu_1$ and $\theta_2 < \mu_2$. Figure 5.2 (each color represents the same case as defined in Figure 4.6) presents the same analysis for other combinations. There are three other possibilities for relationships between θ_1, μ_1 and θ_2, μ_2. We find in different combinations, the monotonicity of the upper and lower bounds of Case(b) (the yellow region) changes as we increase the capacity m. In Case 1 ($m \leq \lambda_1/\mu_1$), both bounds decrease when $\mu_1 > \theta_1$ and increase when $\mu_1 < \theta_1$. In

Case 2 ($\lambda_1/\mu_1 < m < \lambda_1/\mu_1 + \lambda_2/\mu_2$), the upper bound decreases when $\mu_2 > \theta_2$ and increases when $\mu_2 < \theta_2$; the lower bound does not change and always equals λ_1/μ_1. In Case 3, the upper and lower bounds are constants. The system equilibrium is in Case(b) when $\lambda_1/\mu_1 < N < \lambda_1/\mu_1 + \lambda_2/\mu_2$, which is insensitive to m. Meanwhile, for all $N < \lambda_1/(\mu_1 \vee \theta_1)$, the system is always in Case(c) (the green region), whereas for all $N > \lambda_1/(\mu_1 \wedge \theta_1)/+\lambda_2/\mu_2$, the system is always in Case(a) (the red region). Therefore, when the threshold N is relatively large/small, it stops affecting the equilibrium.

Figure 5.2: The dependence of equilibrium distribution on different parameters



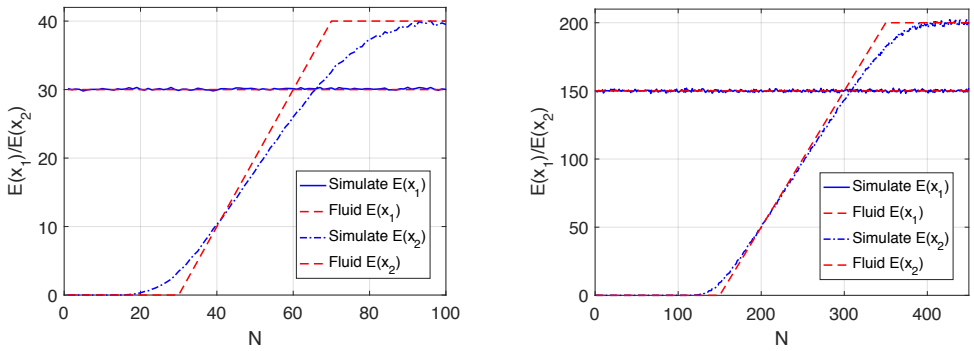
(a) $\theta_1 < \mu_1, \theta_2 < \mu_2$ (b) $\theta_1 < \mu_1, \theta_2 > \mu_2$ (c) $\theta_1 > \mu_1, \theta_2 > \mu_2$

5.2 Fluid-Based Performance Measures

We use simulation of the original two-class stochastic system to examine the accuracy of the fluid approximation. Concentrating on Case 2, in which class 1 customers are underloaded and in general the system is overloaded, we check the long-term behavior of both medium ($m = 40, \lambda_1 = 30$ and $\lambda_2 = 20$) and large ($N = 200, \lambda_1 = 150$ and $\lambda_2 = 100$) systems. For both systems, $\mu_1 = 1, \mu_2 = 0.8, \theta_1 = 0.5$ and $\theta_2 = 0.4$. Figure 5.3 presents the comparison of simulation and approximation of the expected number of people in the system for classes—$E(x_1)$ and $E(x_2)$.

Figure 5.3: Simulate vs. fluid: $E(x_1)$ and $E(x_2)$ as a function of N

(a) $m = 40$ (b) $m = 200$

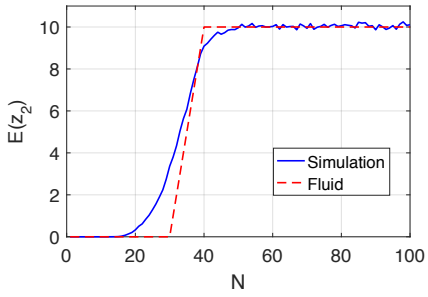


By comparing the simulation result, we can see that in both size systems, the approximation of the equilibrium of x_1 ($\bar{x}_1 = 30, 150$ in the medium and large size system, respectively) is very accurate. This is due to the fact that class 1 is underloaded at all times. Such accuracy is insensitive to the value of threshold N, which is evident for the independency of the equilibrium of x_1 in N. Thus, from now on, we focus on the performance metrics of class 2 customers. The approximated equilibrium of x_2 becomes more accurate as the system size increases. When N is close to $\bar{x}_1 + \bar{x}_2^L$ ($\bar{x}_2^L = 0$) and $\bar{x}_1 + \bar{x}_2^H$ ($\bar{x}_2^H = 40$ and 200, respectively), the accuracy of approximation decreases. This is because the dynamics of the fluid approximation is nonsmooth when $N = \bar{x}_1 + \bar{x}_2^H, \bar{x}_1 + \bar{x}_2^L$.

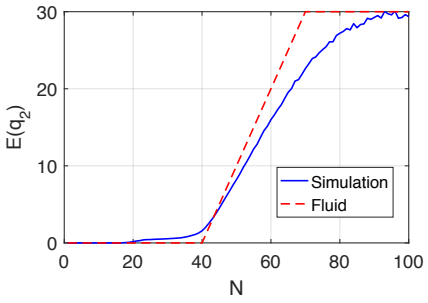
By substituting the equilibrium of x_1 and x_2 into the original system dynamics (4.1), we can determine the equilibria of the number of customer i in service ($\bar{z}_i$) and in queue ($\bar{q}_i$). In Case 2, $\bar{z}_1 = \bar{x}_1, \bar{q}_1 = 0$, which are constants. According to the simulation of the above two systems, the average queue length of class 1 customers is very close to 0 ($E(q_1) = 0.5674$ and 0.3460 in medium and large size systems, respectively) and the average number of servers who serve class 1 customers is very close to $\bar{z}_1$ ($E(z_1) = 29.8538, \bar{z}_1 = 30$ and $E(z_1) = 149.9278, \bar{z}_1 = 150$ in medium and large size systems, respectively). Similar to $\bar{x}_1$, the fluid approximation of $\bar{z}_1$ and $\bar{q}_1$ also perform very accurately. Figure 5.4 shows the comparison of simulation (solid line) and approximation (dotted line) of $\bar{z}_2$ and $\bar{q}_2$ for medium and large size systems.

Figure 5.4: Simulation vs. fluid: $E(z_2)$ and $E(q_2)$ as a function of N

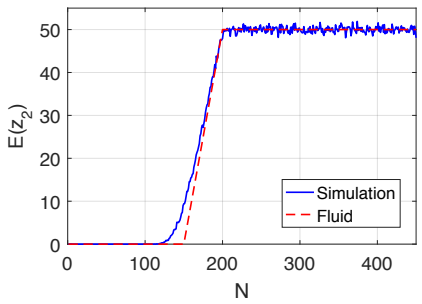
(a) $m = 40$: $E(z_2)$



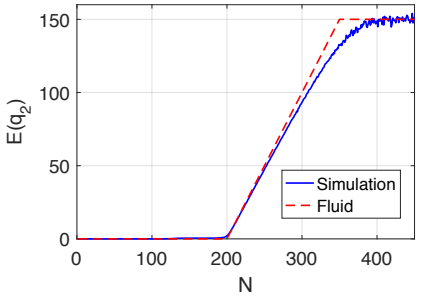
(b) $m = 40$: $E(q_2)$



(c) $m = 200$: $E(z_2)$



(d) $m = 200$: $E(q_2)$

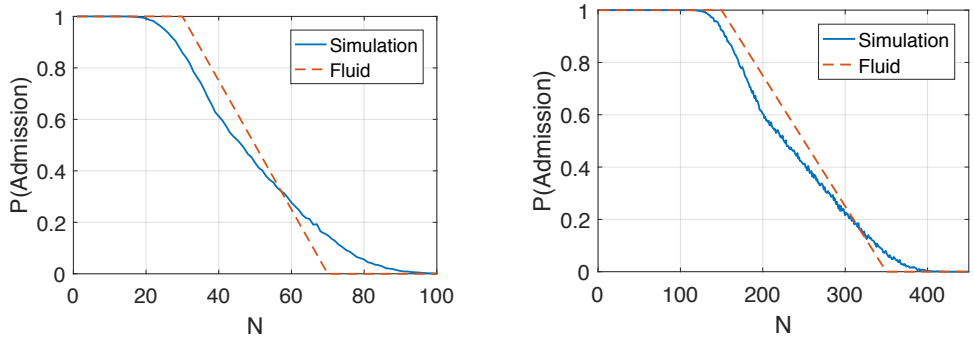


Similarly, the accuracies of $\bar{z}_2$ and $\bar{q}_2$ both improve when the system size increases. From Figure 5.4a and 5.4c, we observe that $\bar{z}_2$ loses more accuracy when N is close to $\bar{z}_1$. More precisely, when N is around $\bar{z}_1$, $\bar{z}_2$ is underestimated. This is typical of fluid approximations, when N is lower than $\bar{z}_1$ (which equals $\bar{x}_1$); the system does not give any service load to class 2 customers. However, because class 1 customers arrive and get service stochastically, the number of class 1 customers in the system sometimes is less than $\bar{x}_1$. As N increases, it is more and more likely that the admission control constraint is not satisfied during the arrival of class 2 customers. Thus, on the stochastic level, there are more class 2 customers admitted into the system and get served than we approximate on a fluid level, when N is around $\bar{z}_1$. In Figure 5.4b and 5.4d, there is more inaccuracy observed when $N = \bar{x}_1 + \bar{x}_2^H$. This is caused by a nonsmooth fluid approximation, just like its influence on the accuracy of $\bar{x}_2$ in the same area.

Note that the goal of this thesis is to determine an invitation policy for a proactive service system that balances revenue and service level. From the perspective of revenue, we notice that the revenue highly depends on the rate of arrival customers. In our two-class model, the revenue of class 1 customers is constant for each set of determined system parameters. The reason is that we neither control the arrival of class 1 customers, nor do we depend depends on class 2 customers. Therefore, we can only focus on how the revenue of class 2 customers is changed by different thresholds. According to the model, only certain class 2 customers are admitted to the system. In order to find the effective arrival rate of class 2 customers, we need the probability of the usage of admission control. In the end of Chapter 4, we approximate the probability that the admission of class 2 customers will be denied—P(Admission). Figure 5.4 shows the performance (solid line) of this approximation for both system sizes. Similar to the phenomena of $\bar{x}_2$, larger system approximations perform better, and inaccuracy happens when the value of N is round $\bar{x}_1 + \bar{x}_2^L$ or $\bar{x}_1 + \bar{x}_2^H$. These can be explained by similar arguments as before.

Figure 5.5: Simulation vs. fluid: P(Admission) function of N

(a) $m = 40$ (b) $m = 200$



In addition, based on the result of the fluid equilibrium, we can calculate approx-

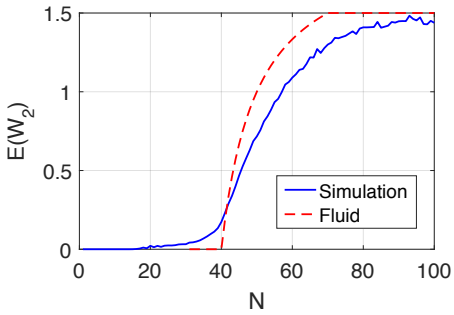
imations of serval service level indicators, then use them for policy balancing. The most common metrics are the expected waiting time, $E(W_2)$, and the probability of abandonment, $P(Ab_2)$. Because in Case 2, the fluid approximating queue length for class 1 customers is 0, we ignore class 1. In our model dynamics, those performance measures are approximated by

$$\begin{aligned} E(W_2) &= \bar{q}_2/(\lambda_2(1-\alpha)); \\ P(Ab_2) &= \theta_2\bar{q}_2/(\lambda_2(1-\alpha)). \end{aligned} \tag{5.1}$$

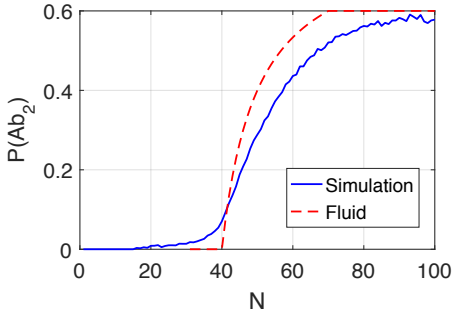
We simulate medium and large systems with the same parameters as before to examine the performance of these approximations. Results are shown in Figure 5.6.

Figure 5.6: Simulation vs. fluid: $E(W_2)$ and $P(Ab_2)$ function of N

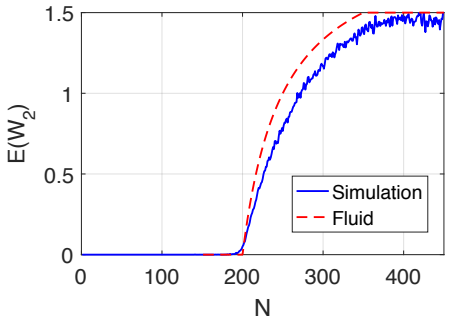
(a) $m = 40$: $E(W_2)$



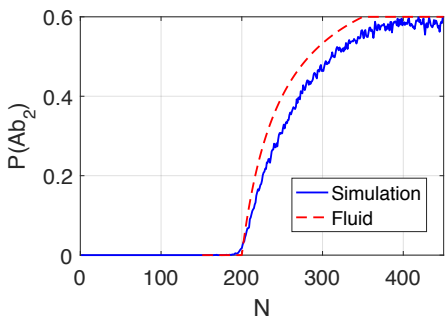
(b) $m = 40$: $P(Ab_2)$



(c) $m = 200$: $E(W_2)$



(d) $m = 200$: $P(Ab_2)$



Note that the fluid approximation (dotted line) of $E(W_2)$ is in fact $P(Ab_2)$ scaled with a constant value θ_2. The simulation results, i.e., the comparison of the curve between $E(W_2)$ and $P(Ab_2)$ with the same m (between Figure 5.6a and 5.6b, and between Figure 5.6c and 5.6b), also support such an observation. Thus, it is sufficient to discuss $E(W_2)$. From the approximation (5.1), we find that, $\bar{q}_2$ and α depend on N on the right-hand side of $E(W_2)$. Meanwhile, α is a denominator. According to the simulation results we acquired before, when N is near $\bar{x}_1 + \bar{x}_2^H$, $\bar{q}_2$ is underestimated whereas α is overestimated. Therefore, we expect an underestimation to occur here. The simulation results shown in Figure 5.6a and 5.6c verify, that regardless of system

size, the inaccuracy of $E(W_2)$ is more obvious when N is close to $\bar{x}_1 + \bar{x}_2^H$. Moreover, these service level approximations are getting better with the system size.

We examine all the performances of the approximation obtained in Chapter 4, and deduced for the equilibrium. In general, all of them perform well, especially in large-size systems. Meanwhile, we have observed some interesting phenomena of the performance as N is varying. For instance, the inaccuracy level of the approximation is not symmetric. We discuss this next.

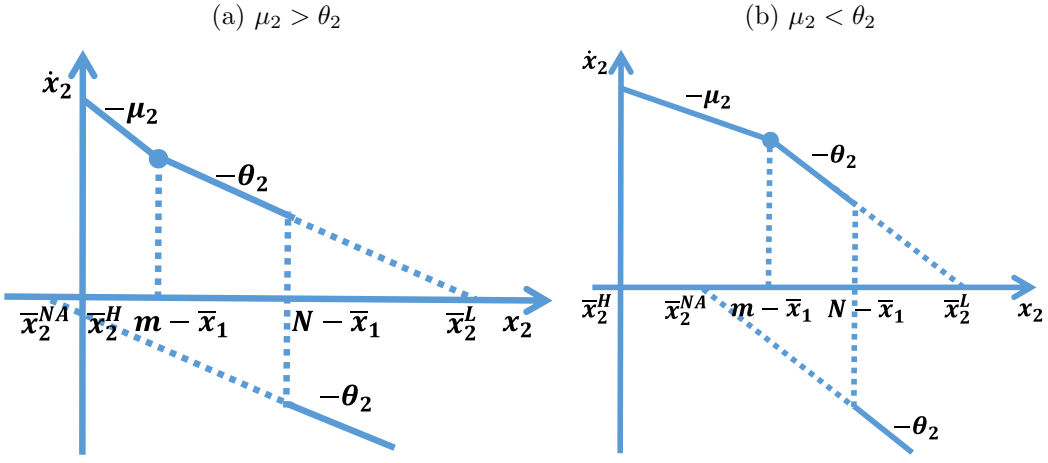
5.3 Further Discussion of the Performance

Still in Case 2, in Figure 5.3, we have observed an asymmetric mismatch around $\bar{x}_2^L$ and $\bar{x}_2^H$ for both medium and large size systems. Let's assume that the number of class 1 customers in the system constantly equals $\bar{x}_1$. All of them are in service. The remaining servers, $m - \bar{x}_1$ in total, serve class 2 customers. Thus, the system dynamics (4.2) can be simplified into a single-variable ODE

$$\dot{x}_2 = I_{\{x_2 < N - \bar{x}_1\}}\lambda_2 - \mu_2 \left(x_2 \wedge (m - \bar{x}_1)\right) - \theta_2 (x_2 - (m - \bar{x}_1))^+. \tag{5.2}$$

If the threshold, N, satisfies Case 2(b), its dynamics can be captured by Figure 5.7.

Figure 5.7: The dynamics analysis of x_2 with a given $\bar{x}_1$ in case 2(b)



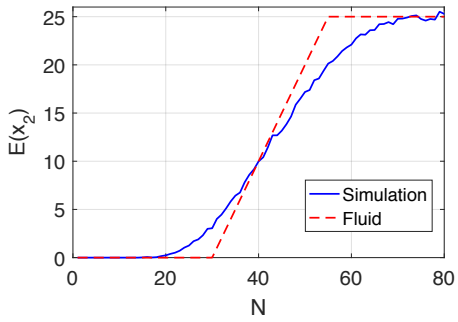
In the case that the admission control (according to the threshold $N - \bar{x}_1$) is applied (the lower lines in both diagrams), its equilibrium is not feasible. We can still obtain its value, denoted as a non-admissible (NA) equilibrium. We have $\bar{x}_2^{NA} = (\theta_2 - \mu_2)(m - \bar{x}_1)/\theta_2$. By comparing with $\bar{x}_2^H = 0$, $\bar{x}_2^{NA}$ can either be less or greater than $\bar{x}_2^H$. However, according to the original system constraints, when applying admission control, case 2(b) disappears and the system converges to $\bar{x}_2^H = 0$. Thus, when $\bar{x}_2^{NA}$ is positive, namely $\mu_2 < \theta_2$, one admission control is applied and the system converges to $\bar{x}_2^{NA}$ faster than it converges to $\bar{x}_2^H$. In the case of negative $\bar{x}_2^{NA}$, it converges slower

than it converges to $\bar{x}_2^H$.

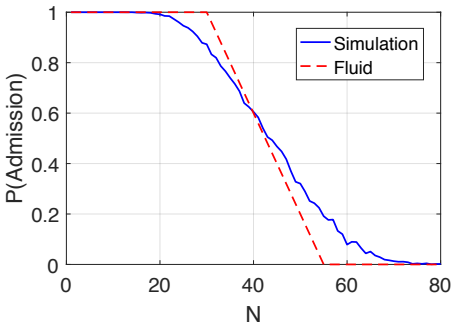
In Figure 5.3, $\mu_2 > \theta_2$. Thus, the simulation result of $E(x_2)$ is only slightly more than the fluid result when N is close to its lower bound of case 2(b), whereas the inaccuracy of P(Admission) is large. When θ_2 increases, the cross point of the simulation curve and fluid approximation is increasing in $E(x_2)$ but is decreasing in P(Admission) (see Figure 5.8). Moreover, we have observed that in the simulation of P(Admission), the slope of the trace has a small but significant change, as seen in Figure 5.5. This change disappears when $\bar{x}_2^{NA}$ overlaps $\bar{x}_2^H$ and emerges again when $\mu_2 < \theta_2$ (see Figure 5.8b and 5.8d). Though this change does not adversely affect the accuracy of fluid approximation significantly, the reason of this phenomenon is also worthy of future study.

Figure 5.8: Simulation and fluid of case 2 with different pairs of μ_2 and θ_2

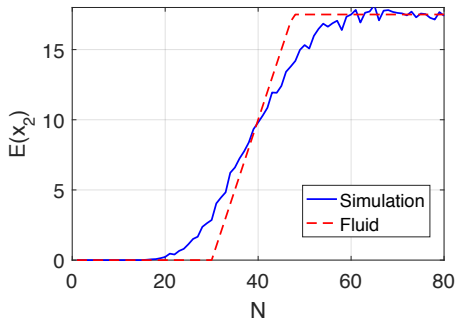
(a) $\mu_2 = \theta_2$: $E(x_2)$)



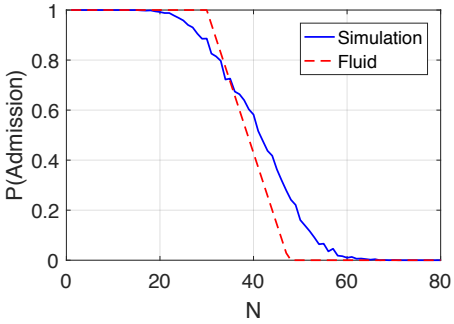
(b) $\mu_2 = \theta_2$: P(Admission)



(c) $\mu_2 < \theta_2$: $E(x_2)$



(d) $\mu_2 < \theta_2$: P(Admission)



Chapter 6

Conclusion

Motivated by various applications in chat services, law-enforcement and healthcare systems, we developed an invitation policy in the form of a threshold for a proactive service system to promote system revenues while considering the service level provided to customers.

Based on the analysis of a realistic proactive chat service system, we constructed a multiclass multiserver model with impatient customers and built an objective revenue function. According to the model, we first found an optimal fluid policy—$r\mu$ rule—by solving a linear programming problem of the fluid model. Through simulation of the fluid policy, we proposed an easily applicable threshold policy that applies to only one class of customers to control their admission. It is found that the system equilibrium under such a policy is globally asymptotically stable. This result is obtained in Theorem 4.3. Such an equilibrium helps us approximate the probability of the implementation of admission control with different thresholds. Furthermore, we discussed the performance of these approximations and deduced approximations for service level metrics as well. All approximations perform well, especially in large systems. Therefore, one can use such approximations to determine an invitation policy that maximizes revenue while the system service level satisfies specified constraints.

The above conclusion is obtained by analyzing a simplified version of the original system. The system we studied empirically, is more complex and allows, for example, the agents to serve multiple customers in parallel. When we built the model, we did not take that feature into account. Nevertheless, it is very common in a chat service system, which we suggest to be added in future research. In addition, in chat systems, customers need a random time to make their decision after they receive an invitation. We neglected such decision time. Therefore, further analysis is needed to understand the impact of this decision time delay.

Note that we discussed the equilibrium under a preemptive assumption. In most cases, the preemptive and non-preemptive cases converge to the same equilibrium when size goes to infinity. However, the difference appears when the load of the higher-ranked customers is very close to be critically loaded. If the system has many classes, such

exceptional cases can happen. Thus, the non-preemptive case is also worth exploration in the future.

So far, we verified validity of the approximation we obtained using simulation. In the next step, we suggest to use our case study to check the effectiveness of our results for the determination of threshold in practice. One can also investigate some other approaches that are based on the equilibrium result, to analyze our model stochastically and improve the approximation (see Chan and Yom-Tov (2015)).

The implementation of our policy in practice is not straightforward as our classes so far were only based on scores; classes by the model should be set by differences in μ as well. Also, the number of classes in practice is not well defined; we suggested to rely on the empirical observation that the score seems to be a mixture of several distributions, but other approaches might be considered too. For example, if one can forecast not only score but also service time for each individual, maybe revenues can be enhanced even further.

The application to a healthcare environment suggests several further extensions. In the healthcare system, the invitation policy needs to take into account also exogenous unplanned arrivals. Therefore, the service level for both invited and unexpected patients needs to be considered. A first solution for such environments could be to consider those types of customers as having the highest priority regardless of their $r\mu$ value.

Bibliography

Andrews, DC, KN Haworth. 2002. Online customer service chat: Usability and sociability issues. *Journal of Internet Marketing* **2**(1) 1–20.

Armony, Mor, Itai Gurvich. 2010. When promotions meet operations: Cross-selling and its effect on call center performance. *Manufacturing & Service Operations Management* **12**(3) 470–488.

Ata, Baris, Shiri Shneorson. 2006. Dynamic control of an m/m/1 service system with adjustable arrival and service rates. *Management Science* **52**(11) 1778–1791.

Atar, Rami, Chanit Giat, Nahum Shimkin. 2010. The cμ/θ rule for many-server queues with abandonment. *Operations Research* **58**(5) 1427–1439.

Atar, Rami, Haya Kaspi, Nahum Shimkin. 2013. Fluid limits for many-server systems with reneging under a priority policy. *Mathematics of Operations Research* **39**(3) 672–696.

Atar, Rami, Avi Mandelbaum, Martin I Reiman. 2004. Scheduling a multi-class queue with many exponential servers: Asymptotic optimality in heavy traffic. *The Annals of Applied Probability* **14**(3) 1084–1134.

Bernardo, Mario, Chris Budd, Alan Richard Champneys, Piotr Kowalczyk. 2008. *Piecewise-smooth Dynamical Systems: Theory and Applications*, vol. 163. Springer Science & Business Media.

Chan, Carri W., Galit Yom-Tov. 2015. How to balance admission control, speedup, and waiting Working paper, Technion.

Chan, Carri W, Galit Yom-Tov, Gabriel Escobar. 2014. When to use speedup: An examination of service systems with returns. *Operations Research* **62**(2) 462–482.

De Véricourt, Francis, Yong-Pin Zhou. 2005. Managing response time in a call-routing problem with service failure. *Operations Research* **53**(6) 968–981.

Fawcett, Tom. 2006. An introduction to roc analysis. *Pattern Recognition Letters* **27**(8) 861–874.

Filipov, Aleksei Fedorovich. 1988. Differential equations with discontinuous right-hand side. *Amer. Math. Soc.* 191–231.

Ghosh, Arka P, Ananda P Weerasinghe. 2007. Optimal buffer size for a stochastic processing network in heavy traffic. *Queueing Systems* **55**(3) 147–159.

Ghosh, Arka P, Ananda P Weerasinghe. 2010. Optimal buffer size and dynamic rate control for a queueing system with impatient customers in heavy traffic. *Stochastic Processes and Their Applications* **120**(11) 2103–2141.

Huang, Junfei, Boaz Carmeli, Avishai Mandelbaum. 2015. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research* **63**(4) 892–908.

Karush, William. 1939. Minima of functions of several variables with inequalities as side conditions. Ph.D. thesis.

KC, Diwas Singh. 2013. Does multitasking improve performance? evidence from the emergency department. *Manufacturing & Service Operations Management* **16**(2) 168–183.

Khalil, Hassan K. 1996. *Noninear Systems*. Prentice-Hall, New Jersey.

Koçağa, Yaşar Levent, Amy R Ward. 2010. Admission control for a multi-server queue with abandonment. *Queueing Systems* **65**(3) 275–323.

Koole, Ger, Auke Pot. 2011. Technical note—a note on profit maximization and monotonicity for inbound call centers. *Operations research* **59**(5) 1304–1308.

Kroese, Dirk P, RY Rubinstein. 2008. *Simulation and the Monte Carlo Method*. Wiley New York.

Lam, Son K, Stefan Sleep, Thorsten Hennig-Thurau, Shrihari Sridhar, Alok R Saboo. 2017. Leveraging frontline employees' small data and firm-level big data in frontline management an absorptive capacity perspective. *Journal of Service Research* Forthcoming.

Lee, Nelson, Vidyadhar G Kulkarni. 2014. Optimal arrival rate and service rate control of multi-server queues. *Queueing Systems* **76**(1) 37–50.

Lyapunov, Aleksandr Mikhailovich. 1992. The general problem of the stability of motion. *International Journal of Control* **55**(3) 531–534.

Miller, Bruce L. 1969. A queueing reward system with several customer classes. *Management Science* **16**(3) 234–245.

Munkres, James R. 1997. *Analysis on Manifolds*. Westview Press.

Pang, Guodong, Ohad Perry. 2014. A logarithmic safety staffing rule for contact centers with call blending. *Management Science* **61**(1) 73–91.

Perry, Ohad, Ward Whitt. 2009. Responding to unexpected overloads in large-scale service systems. *Management Science* **55**(8) 1353–1367.

Perry, Ohad, Ward Whitt. 2011. A fluid approximation for service systems responding to unexpected overloads. *Operations Research* **59**(5) 1159–1170.

Pinedo, Michael. 1983. Stochastic scheduling with release dates and due dates. *Operations Research* **31**(3) 559–572.

Samuelson, Douglas A. 1999. Predictive dialing for outbound telephone call centers. *Interfaces* **29**(5) 66–81.

Sarraf, Mohsen. 1989. Performance analysis of the outbound call management system. *INFOCOM'89. Proceedings of the Eighth Annual Joint Conference of the IEEE Computer and Communications Societies. Technology: Emerging or Converging, IEEE*. IEEE, 373–381.

Shae, Zon-Yin, Deepak Garg, Rajarshi Bhose, Rohan Mukherjee, Sinem Guven, Gopal Pingali. 2007. Efficient internet chat services for help desk agents. *Services Computing, 2007. SCC 2007. IEEE International Conference on*. IEEE, 589–596.

Slotine, Jean-Jacques E, Weiping Li. 1991. *Applied Nonlinear Control*, vol. 199. prentice-Hall Englewood Cliffs, NJ.

Smith, Wayne E. 1956. Various optimizers for single-stage production. *Naval Research Logistics Quarterly* **3**(1-2) 59–66.

Tan, Tom Fangyun, Serguei Netessine. 2014. When does the devil make work? an empirical study of the impact of workload on worker productivity. *Management Science* **60**(6) 1574–1593.

Tezcan, Tolga, Jiheng Zhang. 2014. Routing and staffing in customer service chat systems with impatient customers. *Operations Research* **62**(4) 943–956.

Van Mieghem, Jan A. 1995. Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *The Annals of Applied Probability* 809–833.

Ward, Amy R, Sunil Kumar. 2008. Asymptotically optimal admission control of a queue with impatient customers. *Mathematics of Operations Research* **33**(1) 167–202.

Weerasinghe, Ananda, Avishai Mandelbaum. 2008. Abandonment vs. blocking in many-server queues: asymptotic optimality in the qed regime. Tech. rep., Working paper.

Yom-Tov, Galit B, Avishai Mandelbaum. 2014. Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management* **16**(2) 283–299.

Yoon, Seunghwan, Mark E Lewis. 2004. Optimal pricing and admission control in a queueing system with periodically varying parameters. *Queueing Systems* **47**(3) 177–199.

Zayas-Cabán, Gabriel, Mark E Lewis. 2016. Admission control in a two class loss system with periodically varying parameters and abandonments Working paper.

תקציר

מערכות שירות פרו-אקטיביות, בשונה ממערכות שירות קלאסיות, מאפשרות שליטה ובקרה על קצב ההגעה של לקוחות למערכת. במערכות אלו, חלק מהלקוחות (ולעיתים כולם) מגיעים עקב הזמנות יזומות שמפעילה המערכת, הזמנות בהן המערכת מציעה ללקוחות לצרוך את שירותיה, באופן שיביא לשליטה טובה יותר במדדים תפעוליים שונים וברווחיות. מערכות שירות מסוג זה משמשות, לדוגמה, למידול שירותי צ׳אט באינטרנט, או לתכנון אסטרטגיות טיפול מונע במערכות בריאות.

במסגרת מחקר אמפירי של מערכת צ׳אט פרו-אקטיבית אנו מתקפים את החשיבות של סיווג הלקוחות למחלקות עדיפות שונות, ומשתמשים בסיווג זה לצורך אופטימזציה על מדיניות הזמנת הלקוחות למערכת. הראינו כי ניתן להגדיר מדדי שירות שונים באמצעות עלויות נטישה והמתנה, ולבצע אופטמיזציה על הרווח המקסימלי תחת אילוצים על מדדי השירות הללו, באופן שיגביל את ההשפעה השלילית של עומס יתר במערכת. מדדים תפעוליים אחרים, כולל מדדים ייחודיים לשירותי צ׳אט כגון, כמות השירות המקבילי (שירות שבו נציג אחד משרת מספר לקוחות במקביל) והמתנות פנימיות (המתנות לאחר קבלת השירות הראשוני, בין שירות אחד למשנהו), לא נחשבו כעלויות במודל שלנו .

במחקרנו, מידלנו את מערכת הצ׳אט כתור מרובה שרתים, עם לקוחות ממחלקות עדיפות שונות וסבלנות סופית. הנחנו כי סבלנות הלקוחות מתפלגת מעריכית וכי הפרמטר (קצב) של התפלגות זו תלוי במחלקת העדיפות אליה שייך הלקוח. לצורך הפשטת המודל הנחנו כי עלויות ההמתנה והנטישה זהות עבור כלל הלקוחות. השתמשנו במודל נוזלים ופתרנו בעיית תכנות לינארי למקסום הרווח (ההפרש בין הכנסות המערכת והעלויות), שהובילה למציאת המדיניות האופטימלית הגבולית. לפי מדיניות זו, יש להזמין לקוחות למערכת לפי סדר יורד של ערכי ה - $r\mu$ שלהם (כאשר r הינו הרווח משירות הלקוח ו - μ הינו קצב השירות של הלקוח), לפיכך נקראת מדיניות זו ״מדיניות $r\mu$ ״.

בנוסף, אנו מציעים לשלב מדיניות זו עם מדיניות סף, הקובעת שיש לעצור את הזמנת הלקוחות כאשר כל השרתים עסוקים. הראינו כי מדיניות סף זו הינה פשוטה ליישום וכן מביאה לביצועים טובים יותר מאשר מדיניות ה - $r\mu$ המשולבת עם סוגים אחרים של בקרה על קצב ההגעה. באמצעות סימולציה אנו מראים את היתרונות והמגבלות של מדיניות זו וכן מראים כי מדיניות ה - $r\mu$ האופטימלית, שנגזרה מקירוב הנוזלים ואינה כוללת מדיניות סף, מביאה אמנם לביצועים טובים אך אינה עדינה מספיק ברמה הסטוכסטית. בכדי לעדנה ניתחנו את מודל הנוזלים של המערכת בשילוב עם מדיניות סף, תחת ערכי סף שונים. איחדנו את מחלקות העדיפות המקוריות לשתי מחלקות בלבד, כאשר מדיניות הסף מגבילה את קצב ההגעה של הלקוחות בעלי העדיפות הנמוכה בלבד. לקוחות אשר נכנסו למערכת מקבלים שירות לפי העדיפות שלהם (כולל הפרעה לשירות של לקוח אחר ממדיניות נמוכה יותר).

הדינמיקה של מודל הנוזלים הינה אי רציפה מימין. מרחב המצב של המערכת מחולק על ידי הגבול לשני תחומים עם משוואות דיפרנציאליות רגילות רציפות וחלקות למקוטעין. מצאנו כי שיווי המשקל בין שני התחומים תלוי מאוד בפרמטרים של המערכת (קצבי ההגעה, קצבי השירות ועוד) ובעיקר בערך הסף. בנוסף מצאנו כי נקודת שיווי המשקל הינה יציבה אסימפטוטית, באופן גלובאלי, עם מסלולים יציבים במובן ליאפונוב. הראינו זאת על ידי התמקדות בדינמיקה של המערכת בפס אנכי מסביב לשיווי המשקל. תוצאה זו מוצגת במשפט 4.3. בפרט, במקרים מסויימים, שיווי המשקל נמצא באזורי הזזה.

על מנת להציע מדיניית הזמנה למערכות שירות פרואקטיביות אשר מאזנת בין רווחיות ורמת שירות חישבנו בקירוב את ההסתברות להפעלת בקרה על קצב ההגעה (כלומר, ההסתברות שיהיה צורך לעצור את ההזמנות למערכת) וכן מצאנו קירובים למספר מדדים המצביעים על רמת השירות, כגון זמן ההמתנה וההסתברות לנטישה. קירובים אלו נבחנו באמצעות סימולציה ונמצאו טובים, במיוחד עבור מערכות גדולות. בנוסף, אנו דנים בעבודתנו בהשפעה של פרמטרי המערכת על הדיוק של קירוב הנוזלים.

המחקר נעשה בהנחיית ד״ר גלית שם טוב ופרו״פ לירון ידידציון בפקולטה להנדסת תעשייה וניהול.
אני מודה לבית הספר ללימודי הסמכה בטכניון על התמיכה הכספית הנדיבה בהשתלמותי.

מדיניות הזמנת לקוחות למערכת שירות יוזמת: איזון יעילות, רווחיות ורמת שירות

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר מגיסטר למדעים
בהנדסת תעשיה (עם תזה)

יומינג שיא

הוגש לסנט הטכניון – מכון טכנולוגי לישראל

אדר תשע״ז, חיפה, פברואר 2017

מדיניות הזמנת לקוחות למערכת שירות יוזמת: איזון יעילות, ריווחיות ורמת שירות

יומינג שיא