Probabilistic Matching Systems: Stability, Fluid and Diffusion Approximations and Optimal Control

THE UNIVERSITY *of* EDINBURGH

Hanyi Chen

Supervisor:
Dr. Burak Büke

Doctor of Philosophy
The University of Edinburgh
2015

Probabilistic Matching Systems: Stability, Fluid and Diffusion Approximations and Optimal Control

Doctoral thesis

Hanyi Chen

SCHOOL OF MATHEMATICS
The University of Edinburgh

Hanyi Chen
James Clerk Maxwell Building, School of Mathematics, The University of Edinburgh
E-mail: *hanyii.chen@gmail.com*
Website: *http://www.maths.ed.ac.uk/~s1045742/*

Supervisor: Dr. Burak Büke
James Clerk Maxwell Building, School of Mathematics, The University of Edinburgh
E-mail: *B.Buke@ed.ac.uk*

Examiners:
I wish to thank
(External) Dr. Ayalvadi Ganesh
(Internal) Dr. Gonçalo dos Reis
for kindly agreeing to be my examiners and providing many valuable and insightful comments on the improvements of the thesis.

Doctoral thesis
Initial Submission for examination: February 11, 2015
Final Submission: June 10, 2015
Typeset in LaTeX

Declaration

This thesis, which was composed by the candidate herself, is submitted to The University of Edinburgh in partial fulfilment of the requirements for the degree of Doctor of Philosophy in the School of Mathematics.

The candidate hereby declares that the work presented in this thesis is, to the best of her knowledge and belief, original and her own, except where explicitly stated otherwise in the text. This thesis is based on Buke and Chen [12], [10] and [11]. The candidate contributed heavily to every result of each paper that this thesis is based on. She further asserts that none of the material contained in this thesis has been submitted, either in part or whole, for any other degree or professional qualification.

Hanyi Chen
Edinburgh, February 2015

To my beloved Mum and Dad

Abstract

In this work we introduce a novel queueing model with two classes of users in which, instead of accessing a resource, users wait in the system to match with a candidate from the other class. The users are selective and the matchings occur probabilistically. This new model is useful for analysing the traffic in web portals that match people who provide a service with people who demand the same service, e.g. employment portals, matrimonial and dating sites and rental portals.

We first provide a Markov chain model for these systems and derive the probability distribution of the number of matches up to some finite time given the number of arrivals. We then prove that if no control mechanism is employed these systems are unstable for any set of parameters. We suggest four different classes of control policies to assure stability and conduct analysis on performance measures under the control policies. Contrary to the intuition that the rejection rate should decrease as the users become more likely to be matched, we show that for certain control policies the rejection rate is insensitive to the matching probability. Even more surprisingly, we show that for reasonable policies the rejection rate may be an increasing function of the matching probability. We also prove insensitivity results related to the average queue lengths and waiting times.

Further, to gain more insight into the behaviour of probabilistic matching systems, we propose approximation methods based on fluid and diffusion limits using different scalings. We analyse the basic properties of these approximations and show that some performance measures are insensitive to the matching probability agreeing with the results found by the exact analysis.

Finally we study the optimal control and revenue management for the systems with the objective of profit maximization. We formulate mathematical models for both unobservable and observable systems. For an unobservable system we suggest a deterministic optimal control, while for an observable system we develop an optimal myopic state dependent pricing.

Keywords: Matching systems · Stability · Admission control policies · Rejection rates · Fluid approximation · Diffusion approximation · Optimal control · Unobservable queue · Observable queue · State dependent pricing · Myopic strategy

Lay Summary

In recent years there has been a great increase in the popularity of web portals as a meeting place for individuals who provide services and those who require services. Some commonly used examples are employment portals (efinancialcareers), classified advertisement portals (Gumtree), rental portals (Zoopla) and dating websites. A common feature of these systems is that each pair of users has a probability to match with each other and the system operator has no control on who matches with whom. We introduce a new queueing model, which we define as a probabilistic matching system, to model the traffic in these web portals.

Since so many people use these systems, it is important to guarantee a certain level of service quality and this raises lots of questions which are also mathematically interesting. How many people can find a job in an employment portal and how can we improve the matching rates? What is the probability that someone can find a match in a dating website? What is a good pricing mechanism for system controllers to regularise the traffic and to maximise the profit? In this work we answer many of these questions. To explore the nature of probabilistic matching systems, we use sophisticated mathematical techniques which are also common to very distinct areas of science, such as pricing financial products, projectile guidance systems, and so on.

Our research reveals many important properties these systems exhibit, some of which seem counter-intuitive at the beginning. Through stability analysis, we show that if the system is uncontrolled, many users accumulate in the system without finding a match. We then suggest four types of admission controls to remedy this problem and we study the performance measures of the system under these controls. One of the surprising results is that under one type of admission control policies, a higher probability for individuals to match leads to less users finding a match. Further, using fluid and diffusion approximations, we suggest that the mathematically intractable queue length processes can be replaced by some analytical diffusion processes. Finally, we develop optimal control and dynamic pricing for probabilistic matching systems for both unobservable and observable models.

Acknowledgements

I have dreamt so many times of writing, on a peaceful spring day, the acknowledgements of my doctoral thesis to mark its completion. Now this dream has come true. I am very grateful that I have completed this journey, during which I have received so much support and have never been alone.

First I would like to express my great gratitude to my supervisor Dr. Burak Büke, without whom none of this work would have been brought to fruition. Burak, when you drew the two-dimensional continuous time Markov chain on board and introduced me to this challenging and interesting research topic three and half years ago, my PhD adventure started. Your guidance, encouragement and patience lead me to grow in the academic field. Thank you for bringing me into the mathematical wonderland full of beauty, fun and challenges. I will never forget how you encouraged me to overcome my fear and to grow my strong capability in programming, which was once my weakest area. Thank you for respecting each of my research ideas, while in the mean time pointing out every imperfection and impreciseness in them and guiding them to the right direction. Thank you for supporting me to attend many academic conferences and communications, which matured my research, communication and presentation skills. Not only that have you taught me how to be a good mathematician, but more importantly, you have shown me how to be a good person. I will not forget the way you taught me to deliver good tutorials and to help students. There are too many thank you's I would like to say to you but above all, thank you for being my supervisor for my PhD, which has been one of the most important and valuable experiences in my life.

Next I would like to express my great gratitude to many people who provided valuable comments and discussions during my PhD research. I would like to thank my second supervisors, Dr Miklós Rásonyi and Dr Tibor Antal for their kindness and helpfulness in discussing my research and future development. I owe a big thank you to Mr Slavomír Takáč for the time and effort he spent on the discussion for solving the complex recursion. I must send my special thank you to Dr Martin Takáč and Dr Jin-Han Xie for their generosity in providing insightful research discussions whenever I needed one. Martin and Jin-Han, your kindness, hard work and intelligence were of great support for me and kept motivating me to make progress!

There are many people who have influenced me greatly and have shaped my academic path from an undergraduate to a PhD student. I am using this opportunity to say thank you to all of you here: Associate Professor Lei Zhang at Sun Yat-Sen University, Professor Gernot Akemann and Dr. Cormac Lucas at Brunel University, Professor Jacek Gondzio and Dr. Sotirios Sabanis at The University of Edinburgh.

Moreover, I am grateful to all the financial support I have received during my PhD. I have been fully supported by a PhD scholarship from the School of Mathematics of the University of Edinburgh. I am thankful to the PG travel funding from the School of Mathematics and EPSRC grant EP/I017127/1 (Mathematics for Vast Digital Resources), which supported me to attend research conferences and communications.

Finally and the most importantly, I would like to thank all of my family and my friends. Mum and Dad, thank you for your endless love and financial and spirit support; thank you for always being there for me no matter happens; and thank you for setting excellent examples for me on how to be a good person. Chris, thank you for finding me at the PhD commencement ceremony; thank you for your firm belief in me; thank you for being by my side sharing every happiness with me and reminding me of who I am whenever I feel lost. All of my friends in Edinburgh: Jin-Han, Martin, Andrea, Hatice, Sarah, Dee, Carmen Rovi, Carmen Li, Tim, Robert, Yiming, Zheng and all other friends who haven't been named here, I am so grateful that I have your companionship during this wonderful three and half years. Thank you so much for making the memory of my student times in Edinburgh, a most romantic city, complete!

Contents

Abstract iii

Lay Summary iv

Acknowledgements v

List of Tables ix

List of Figures x

1 Introduction 1

2 Stability Analysis and Stabilizing Policies 5

2.1 Introduction 5

2.2 Mathematical Model 8

2.3 Transient Behaviour of the System 10

2.4 Stability Analysis – Instability of the Uncontrolled System 12

2.5 Stabilizing Policies 15

2.5.1 The Simple Threshold Policy 17

2.5.2 The Accept-the-Shortest-Queue Policy 18

2.5.3 The Functional Threshold Policy 20

2.5.4 The One-Sided Threshold Policy 24

2.6 Numerical Results 26

2.6.1 Computational Experiments on Long Run Percentages of Rejected Users 27

2.6.2 Computational Experiments on Average Queue Lengths and Average Waiting Times 29

2.7 Conclusion 34

3 Fluid and Diffusion Approximations 36

3.1 Introduction 36

3.2 Mathematical Model 38

3.3 Fluid and Diffusion Approximations with Constant Matching Probabilities 39

3.3.1 Fluid Limits 39

3.3.2 Diffusion Limits 44
3.4 Fluid and Diffusion Limits for Systems with Small Matching Probabilities 47
3.4.1 Fluid Limits 47
3.4.2 Diffusion Limits 56
3.5 Numerical Experiments 59
3.6 Conclusion 62

4 Optimal Control and Dynamic Pricing 65
4.1 Introduction 65
4.2 Unobservable Model 68
4.2.1 Model Description 68
Assumptions 68
Formulation 69
Fluid Optimization Problem 70
4.2.2 Deterministic Finite Horizon Problem 71
4.2.3 Deterministic Infinite Horizon Problem 74
Existence of the Optimal Solution 75
Properties of an Optimal Solution 77
4.3 Observable Model 80
4.3.1 Model Description 80
4.3.2 Myopic Pricing 82
4.4 Conclusion 89

5 Conclusion and Future Work 91

A Proof of Theorem 2.1 94

B Stationary Probabilities under the ASQ Policy. 100

C Insensitivities of Functional Threshold Policies with $h(x) = x + d$ 102

Bibliography 105

List of Tables

2.1 Long run percentage of rejected users for functional threshold policies with various threshold functions when $\lambda_1 = \lambda_2 = 1$ 28
2.2 Average queue lengths and waiting times for functional threshold policies with various threshold functions when $\lambda_1 = \lambda_2 = 1$ 29
2.3 Long run average queue length and waiting times for functional threshold policies with various threshold functions when $\lambda_1 = 1$ and $\lambda_2 = 2$ 30
2.4 Long run average queue length under the one-sided threshold policy for varying q 31
2.5 Long run average queue length under one sided threshold policy for varying N_2 31

List of Figures

2.1 Matching probability q vs. long run percentage of rejected class-1 users for functional threshold policies with different threshold functions 28

2.2 State space for the CTMC under functional threshold policies with various threshold functions when $\lambda_1 = \lambda_2 = 1$ 29

2.3 Average queue lengths and waiting times vs. λ_2 for various q and $\lambda_1 = 1$ under the functional threshold policy with $h(x) = x + 5$ 32

2.4 Average queue lengths and waiting times vs. λ_2 for various q and $\lambda_1 = 1$ under the one-sided threshold policy with $N_1 = 2$ 33

3.1 Fluid Limits when $\lambda_1 < \lambda_2$ for various q 60

3.2 Fluid Limits when $\lambda_1 = \lambda_2$ for various q 61

3.3 Fluid Limits when $\lambda_1 < \lambda_2$ for various γ 61

3.4 Diffusion Limits when $\lambda_1 < \lambda_2$ for various γ 62

3.5 Diffusion Limits when $\lambda_1 < \lambda_2$ for various q 62

3.6 Diffusion Limits when $\lambda_1 = \lambda_2$ for various q and γ 63

Chapter 1

Introduction

Due to advances in technology in the last decade, society has widely adopted the internet for business and personal interactions. Web portals that serve as a meeting point for people who provide a specific service/product and people who demand the service/product are becoming increasingly popular. Employment and rental portals, dating and matrimonial sites are some examples of such systems. Since so many people use these systems, it is very important to guarantee a certain level of service quality and this raises many questions which are mathematically interesting. In an employment portal, how many people can find a job and how can we improve the matching rate? What is the probability that someone can find a satisfactory match from a dating website? What should be a good pricing mechanism for a system operator to regularise the traffic and to maximize the profit? In this work, we answer many of these questions.

Note that a common factor in each of these systems is that each pair of users has a probability to match with each other. We introduce a new queueing model, which we call a *probabilistic matching system*, to model the traffic in these web portals. To understand the dynamics of user behaviour in probabilistic matching systems, first consider the example of an employment portal. There are two classes of users, employers and employees, in an employment system. Employers arrive at the system at random times to hire an employee. Upon arrival, they first check the resumés of the employees in the portal. If they can find suitable candidates, they hire one of the candidates and close the position. If there are no suitable candidates, employers create a job posting on the portal website and wait for a suitable candidate to apply for the position. Similarly, when a potential employee arrives at the system, she applies for the existing postings and leaves the system if she gets hired. If she cannot find a suitable job, she posts her resumé on the portal website and waits until a suitable job becomes available. An important feature of this system is that each given employer-employee pair matches with some given probability. As a result, the operators of a probabilistic matching system do not have any control of who matches with whom.

There are several different types of matching systems studied in the literature. Adan and Weiss [1] and Caldentey et al. [14] consider a matching system where customers are matched with servers. Several different types of customers and servers arrive at the system each according to a stochastic process, and each customer type can be served by a subset of server types and each server type can serve a subset of customer types. Adan and Weiss [1] study the stability of the system under the first-come-first-serve policy and derive explicit product-form equations for the matching rate for some specific configurations. Bušić et al. [13] develop necessary and sufficient conditions for the stability of different matching policies for systems where customers and servers form a bipartite graph and study the computational complexity of deciding whether a given policy is stable. In a recent and different line of work, Gurvich and Ward [25] consider a queueing system with multiple types of jobs that can match with each other. Their goal is to minimize the holding cost by developing a dynamic policy to decide which types should match at any given time.

The concept of "type" plays an important role in the work presented above, and differs from the "class" concept used in this work. In [1], [13] and [14], customers and servers can be viewed as two classes of users of the matching system. These classes are further divided into types according to their special properties, e.g. there are several different types of customers and servers. A special subclass of these systems is the double-ended queue (see e.g. Kashyap [29] and Liu et al. [33]), where there is exactly one type of user from each class. In the aforementioned models, once the types of two users (a server and a customer) are known, they match or fail to match deterministically. In many real world systems, this assumption fails to hold and special attention is needed on an individual basis. For example, when a company is hiring for a position, it does not just hire any person who has the desired background, but rather the personal qualifications of the candidates play an important role in addition to their competences in the field. Hence, each candidate should be considered individually, rather than being classified into a specific type. In this work, we incorporate this individuality by assuming that matchings occur probabilistically.

We consider probabilistic matching systems with two classes (e.g. customers-servers, or employers-employees). Upon arrival each user scans the queue of the other class and may match with each individual in the queue independently with probability q. If there are more than one matching users in the other queue, one is chosen uniformly at random. Similar to the aforementioned work, we also assume that the matching procedure does not take time and happen instantaneously. The matching probability q is a key factor in the analysis of probabilistic matching systems. If the matching probability $q = 1$ and there are two users from different classes have arrived, they match and leave the system immediately and the users from different classes cannot co-exist in the system. Hence, the system can be modelled as a one-dimensional continuous time random walk. However, when the matching probability is less than 1, we need a two-dimensional process, as users from different classes can be present in the system at the same time.

In this work, we explore the nature of probabilistic matching systems from three aspects: stability, fluid and diffusion approximation and optimal control. We introduce a continuous time Markov chain model for probabilistic matching systems and analyse user behaviours in probability matching systems by an exact analysis in Chapter 2. We first provide an explicit formula for the distribution of the number of matched pairs up to time t which turns out to be complicated for further analysis. Then we study the stability of the system and show that if the system is uncontrolled, then many users accumulate in the system without finding a match. We proceed to suggest four types of admission controls to remedy this problem. Through analysis on the performance measures under these controls, we reveal many interesting properties the system exhibits, some of which seem counter-intuitive at the beginning.

The computational complexity and mathematical intractability of the matching process by an exact analysis motivates us to use approximation techniques to further characterize the system. We study diffusion approximations in Chapter 3 to replace to mathematically intractable queue length process by some appropriate diffusion process. We suggest two types of scaling to obtain fluid and diffusion limits. The fluid limit describes the tendency how the queue grows while the diffusion limit captures the fluctuations of the queue length process around its fluid limit. Under the first scaling, time and space are scaled while the matching probability is kept constant. We show that under this scaling both fluid and diffusion limits do not depend on the matching probability. The scaling is suitable for systems with a high matching probability. To provide tools to address a small matching probability, we propose a second scaling that also handles the phenomenon of abandonments by impatient users. In particular, we scale the matching probability and the abandonment rate along with time and space. In addition to a direct analysis on the fluid limits, we study the asymptotic behaviour of the fluid limits and reveal how the user abandonment rate affects the long run average of number of users in the system.

Finally we study the optimal pricing control and revenue management problem for probabilistic matching systems in Chapter 4. We formulate mathematical models for both unobservable and observable systems. For the unobservable system we suggest a deterministic optimal control for both finite and infinite horizon problem, based on fluid limits discussed in Chapter 3. For the observable model, we first derive the maximum price the system can gain from an arriving employer depending on the number of employees in the system by the analysis of employers' strategies. We then develop the optimal pricing mechanism under the assumption of myopic pricing, which maximizes the profit at each stage regardless of future stages. We finally identify under which conditions the optimal myopic pricing is the real optimal pricing and we leave the verification of these conditions as an open question for further research direction.

Chapter 2

Stability Analysis and Stabilizing Policies

2.1 Introduction

In this chapter, we provide a Markov chain model for analysing probabilistic matching systems. We start our analysis by studying the transient behaviour of the matching process, i.e., the counting process for the number of matched pairs up to a finite time t. We first ask the following question: "What is the probability that exactly k matchings have occurred, if we know that exactly m class-1 and n class-2 users have arrived?" To the best of our knowledge, this basic probability distribution has not been studied in the literature. In Section 2.3 we provide an explicit equation for the mass function of this distribution. Unfortunately, this equation is relatively complicated and this indicates the difficulty in completely characterizing the counting process for the number of matched pairs.

Next, we study the stability of the probabilistic matching systems by defining a system to be stable when it is ergodic. The earlier lines of work on "assembly-like queues" [27] and "queues with paired customers" [32] exhibit similarities to probabilistic matching systems in regards to stability. These systems operate similarly to the probabilistic matching systems where pairs from different classes match with probability $q = 1$. However, in these systems the matching procedure (or assembly) takes some time and requires a resource, whereas the matchings are assumed to occur instantaneously in a probabilistic matching system. Harrison [27] studied the waiting time processes for the assembly-like queues and showed that these systems are not stable, regardless of the balance between input and service rates. When the matching probability $q = 1$, a probabilistic matching system is modelled as a one-dimensional random walk on integers, which is known to be unstable. More specifically, it is null recurrent or transient depending on whether the arrival rates are equal or not, respectively. Using a coupling argument, we show that this implies instability of the matching systems with $q < 1$. We also show that when arrival rates are equal, a probabilistic matching system is null recurrent even when $q < 1$.

To stabilize probabilistic matching systems, we suggest four different classes of admission control policies: (i) the simple threshold policy, (ii) the accept-the-shortest-queue policy, (iii) the functional threshold policy and (iv) the one-sided threshold policy. The simple threshold policy places constant bounds on the number of users that can be present in the system from each class. As the resulting state space is finite, the simple threshold policy stabilizes the matching systems when $q = 1$. However, if the matching probability is less than 1, this policy yields absorbing states which indicates that some users experience an infinite waiting time. To avoid absorbing states, the accept-the-shortest-queue and the functional threshold policies use a "moving" threshold and try to "balance" the number of users from each class. The accept-the-shortest-queue admits users if they belong to a class with the minimum number of users in the system and the functional threshold policy admits users if the number of users from that class is less than a function of the number of users from the other class. With very mild conditions on the threshold function to prevent absorbing states, we show that both policies stabilize the system for any set of arrival rates and any positive matching probability. This result is closely related to the work of Latouche [32] on queues with paired customers. Latouche [32] studies the stability of these systems with state-dependent arrival rates, and characterizes the stationary distributions using matrix analytic methods when the system is stable. Latouche defines the "excess" as the difference between the numbers of the two classes of users, and concludes if the state-dependent arrival process keeps the excess bounded then the queues with paired customers can be stable. The stability of the accept-the-shortest-queue policy relies on the same idea and guarantees an excess with absolute value always less than one. However, the functional threshold policy allows the excess to be an increasing function of number of users in the system, and suggests that for the probabilistic matching systems the stability can be achieved by keeping the excess only "under control", instead of keeping it strictly bounded. Our last policy, one-sided threshold policy, relies on the assumption that one of the classes has a higher arrival rate than the other, and only rejects users from the class with higher arrival rate if they exceed a certain constant threshold. This policy also stabilizes probabilistic matching systems with any set of arrival rates that satisfies the above assumption, any positive matching probability and any non-negative threshold value.

Our results for the performance measures of probabilistic matching systems under these policies are even more intriguing. One may think that, as the matching probability increases, the users match more easily and hence, the departure rate of matched pairs should increase, or equivalently the long run rejection rate should decrease. However, contrary to this initial intuition, we prove that under the accept-the-shortest-queue policy and a subclass of the functional threshold policies, the long run rejection rate does not depend on the matching probability. Even more surprisingly, we observe that under most reasonable functional threshold policies *the long run rejection rate actually increases as the matching probability increases*! More specifically, when the threshold function is chosen so that the operators of the system become more eager to admit users of a class as users of the other class accumulate in the system, then the long run percentage of rejected users is an increasing function of the matching probability. In Section 2.6 we intuitively explain this phenomenon based on the geometry of the state space. This explanation is closely related to the "excess" (as defined by Latouche [32]) and clarifies why the long run rejection rate is independent of the matching probability under the accept-the-shortest-queue policy. We further show that the behaviour of the difference between average queue lengths of the two classes is closely related to the behaviour of the long run rejection rate. We prove that for the cases which the rejection rate is independent of the matching probability, the difference of queue lengths also does not depend on the matching probability. However, we observe that if the rejection rate is an increasing function of the matching probability, then the difference between average queue lengths is a decreasing function of the matching probability and vice versa. We finally show that under the one-sided threshold policy, the long run rejection rate is independent of both the matching probability and the threshold value, the explanation of which is however completely different from the above and relies on the well-known arrival-departure theorem.

This chapter is organized as follows. We introduce the mathematical model for probabilistic matching systems in Section 2.2 and we analyse the transient behaviour of the system by concentrating on the process counting the number of matches up to time t in Section 2.3. In Sections 2.4 and 2.5 we discuss the stability of probabilistic matching systems and introduce four stabilizing admission control policies. We present simulation results analysing the long run rejection rates, average queue lengths and average waiting times under the suggested policies in Section 2.6. Finally, we present open problems and future research directions in Section 2.7.

2.2 Mathematical Model

In this section we present a continuous-time Markov chain (CTMC) model for the probabilistic matching systems. The basic assumptions of our model are as follows:

1. The arrival processes of class-1 and class-2 users are independent Poisson processes with rates $0 < \lambda_1 < \infty$ and $0 < \lambda_2 < \infty$, respectively.

2. Each class-1 and class-2 user pair matches with probability q ($0 < q \leq 1$), independent of other users.

3. When a class-1 user arrives, she checks whether there are any class-2 users in the queue matching with her. If there are matching class-2 users, she chooses one of them uniformly at random and then they leave the system together. If there is no matching class-2 user in the system, she joins the queue. A similar mechanism applies when a class-2 user arrives.

4. Once a suitable match is found, the matched pair leaves the system immediately, i.e., the matching procedure does not take any time.

5. Users do not abandon the system without being matched.

Under these assumptions, the system can be modelled as a two-dimensional CTMC $\{X^q(t) = (X_1^q(t), X_2^q(t)), t \geq 0\}$, where $X_i^q(t)$ is the number of class-i users, $i = 1, 2$, in a system with matching probability q at time t. The state space is $\mathbb{S} = \mathbb{N}^2$, where $\mathbb{N}$ is the set of non-negative integers. Since the arrival processes are Poisson processes and the matchings of each pair of class-1 and class-2 users are independent, we have the Markov property.

The probability that a given class-1 and class-2 user pair does not match is $1 - q$. Hence, due to independence of matchings, a class-1 user finding j class-2 users waiting in the system upon arrival does not match with anyone and joins the queue with probability $(1-q)^j$ or she leaves the system with a matching class-2 user with probability $1 - (1 - q)^j$. For notational simplicity, we define $r = 1 - q$ as the probability of not matching for each pair and state the generator matrix as follows:

$$Q_{(n_1,n_2)(n_1',n_2')} = \begin{cases} \lambda_1 r^{n_2} & \text{if } n_1' = n_1 + 1 \text{ and } n_2' = n_2, \\ \lambda_2 r^{n_1} & \text{if } n_1' = n_1 \text{ and } n_2' = n_2 + 1, \\ \lambda_1(1 - r^{n_2}) & \text{if } n_1' = n_1 \text{ and } n_2' = n_2 - 1 \geq 0, \\ \lambda_2(1 - r^{n_1}) & \text{if } n_1' = n_1 - 1 \geq 0 \text{ and } n_2' = n_2, \\ -(\lambda_1 + \lambda_2) & \text{if } n_1' = n_1 \text{ and } n_2' = n_2, \\ 0 & \text{otherwise.} \end{cases}$$

As the generator matrix suggests, it is convenient to use the probability of not matching, r, in the equations, but on the other hand, the probability of matching q is a more intuitive quantity to refer to in the natural language. Hence, we use q and r notation together in the remainder of the thesis, assuming that the relation $q+r = 1$ as obvious.

When $q = 1$, class-1 users and class-2 users cannot co-exist in the system at the same time. Hence, the system can be modeled as a one-dimensional CTMC, where $X^1(t) = k \geq 0$ when there are k class-1 users and $X^1(t) = -k \leq 0$ when there are k class-2 users in the system at time t. This CTMC is a continuous time random walk on integers with rates $Q_{n,n+1} = \lambda_1$ and $Q_{n,n-1} = \lambda_2$.

2.3 Transient Behaviour of the System

We first study the transient behaviour of the system to gain probabilistic insight about the situation at time t. Let $A_i(t)$ and $M^q(t)$ be the number of arrivals from class-i and the number of matched pairs up to time t, respectively. Then, $X_i^q(t) = A_i(t) - M^q(t), i = 1, 2$, for all $t \geq 0$. We focus on the counting process $\{M^q(t), t \geq 0\}$, and provide an explicit equation to the probability of observing exactly k matched pairs up to time t.

When $q = 1$, the probability of having k matched pairs up to time t is trivial. As every class-1 (class-2) user matches upon arrival if there are any class-2 (class-1) users in the system, $M^1(t) = \min\{A_1(t), A_2(t)\}$. Hence, $\mathbb{P}[M^1(t) = 0] = e^{-\lambda_1 t} + e^{-\lambda_2 t} - e^{-(\lambda_1+\lambda_2)t}$, and for $k > 0$,

$$\mathbb{P}[M^1(t) = k] = \frac{e^{-\lambda_1 t}(\lambda_1 t)^k}{k!}\left(1 - \sum_{i=0}^{k-1} \frac{e^{-\lambda_2 t}(\lambda_2 t)^i}{i!}\right) + \frac{e^{-\lambda_2 t}(\lambda_2 t)^k}{k!}\left(1 - \sum_{i=0}^{k} \frac{e^{-\lambda_1 t}(\lambda_1 t)^i}{i!}\right).$$

However, when $0 < q < 1$, the problem is significantly more complicated.

We now define

$$P^q_{k,m,n} \equiv \mathbb{P}[M^q(t) = k | A_1(t) = m, A_2(t) = n], \tag{2.1}$$

for $0 < q \leq 1$, then using the law of total probability, we write

$$\mathbb{P}[M^q(t) = k] = e^{-(\lambda_1+\lambda_2)t} \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} P^q_{k,m,n} \frac{\lambda_1^m \lambda_2^n t^{m+n}}{m!\, n!}. \tag{2.2}$$

The quantity $P^q_{k,m,n}$ is of interest on its own right. For example, one may be interested in the probability of exactly k people getting hired, when there are m companies hiring and n employees looking for jobs. To find an explicit equation for $P^q_{k,m,n}$ we need to solve a three dimensional recursion. Unfortunately, generating function methods are not easy to use to solve this recursion as the coefficients involve powers. Hence, we resort to more direct methods for solving the recusion. Theorem 2.1 presents an explicit formula for the desired conditional probability.

Theorem 2.1. *Suppose $P^q_{k,m,n}$ is as defined in (2.1) and without loss of generality assume $m \leq n$. Then,*

(i) when $q = 1$, $P^1_{k,m,n} = 1$ if $k = m$ and 0 otherwise.

(ii) when $0 < q < 1$,

$$P^q_{k,m,n} = \begin{cases} r^{mn}, & k = 0, \\ a_{k,m} r^{(m-k)n} \prod_{i=0}^{k-1}(1 - r^{n-i}), & 1 \leq k \leq m, \\ 0, & \textit{otherwise}, \end{cases} \tag{2.3}$$

where

$$a_{k,m} = \sum_{l>0} \left(\sum_{d_1+d_2+\ldots+d_l=k} r^k \frac{(-1)^l}{\prod_{i=1}^{l} \prod_{j=1}^{d_i}(1 - r^j)} \right) + r^{-mk} \frac{r^{k^2}}{\prod_{j=1}^{k}(1 - r^j)} + \sum_{i=1}^{k-1} \frac{r^{-mi} r^{ki}}{\prod_{j=1}^{i}(1 - r^j)} \sum_{l>0} \left(\sum_{d_1+\ldots+d_l=k-i} r^{k-i} \frac{(-1)^l}{\prod_{w=1}^{l} \prod_{j=1}^{d_w}(1 - r^j)} \right),$$

with indexes $d_1, d_2, \ldots$ belonging to $\mathbb{N}_+ = \{1, 2, 3, \ldots\}$.

Proof. See Appendix A □

Theorem 2.1 reveals that the elementary conditional probability of k matches succeed given that m class-1 users and n class-2 users have arrived has a fairly complicated expression. This implies that even the one dimensional distribution for the matching process is far from trivial, which further indicates the difficulty in the calculation of finite dimensional distributions to completely characterize $\{M^q(t), t \geq 0\}$ and the transient behaviour of probabilistic matching systems. Next, we concentrate on the steady-state behaviour of probabilistic matching systems.

2.4 Stability Analysis – Instability of the Uncontrolled System

In this work, a system is defined to be stable if it is ergodic and unstable otherwise. In an unstable system, users accumulate in the system without finding a match. We give the formal definition as follows.

Definition 2.2. A probabilistic matching system is *stable* if it is ergodic, i.e., positive recurrent, and unstable if it is either null recurrent or transient.

In this section, we first show that an uncontrolled probabilistic matching system is unstable for any set of arrival rates, λ_1 and λ_2, and matching probability q. To study the stability of the uncontrolled system, we first prove that it is unstable when the matching probability $q = 1$. Then, using a coupling argument, we show that this also implies the instability of the systems where $0 < q < 1$.

Theorem 2.3. *An uncontrolled probabilistic matching system is unstable for any set of arrival rates* λ_1 *and* λ_2*, and matching probability* $0 < q \leq 1$*. More specifically, it is null recurrent if* $\lambda_1 = \lambda_2$ *and transient if* $\lambda_1 \neq \lambda_2$*.*

Proof. When $q = 1$, the system can be modelled as a one-dimensional continuous time random walk which is known to be null recurrent if $\lambda_1 = \lambda_2$ and transient if $\lambda_1 \neq \lambda_2$. To prove the case where $0 < q < 1$, we use a coupling argument. Let $0 < \theta_1 < \theta_2 < \cdots$ be the sequence of occurrence times of events following a homogeneous Poisson process with rate $\lambda_1 + \lambda_2$, and, to determine the class of arrivals, define a sequence of independent discrete random variables $\{\tau_n, n \in \mathbb{N}\}$ which takes the values 1 or 2 with probabilities $\frac{\lambda_1}{\lambda_1+\lambda_2}$ and $\frac{\lambda_2}{\lambda_1+\lambda_2}$, respectively. Let $\{U_n, n \in \mathbb{N}\}$ be a sequence of independent uniform(0,1) random variables, and $\mathbb{1}(A)$ be the indicator function for event A which takes value 1 if A occurs and 0 otherwise. Finally define $\tilde{X}^q_{0,i} = 0, i = 1, 2,$

and for $n \in \mathbb{N}$,

$$\tilde{A}_{n,i} = \sum_{j=1}^{n} \mathbb{1}(\{\tau_j = i\}), i = 1, 2,$$

$$\tilde{M}_n^q = \sum_{j=1}^{n} \left(\mathbb{1}(\{\tau_j = 1\})\mathbb{1}(\{U_j > r^{\tilde{X}^q_{j-1,2}}\}) + \mathbb{1}(\{\tau_j = 2\})\mathbb{1}(\{U_j > r^{\tilde{X}^q_{j-1,1}}\}) \right),$$

$$\tilde{X}^q_{n,i} = \tilde{A}_{n,i} - \tilde{M}^q_n,$$

$$\tilde{X}^q_i(t) = \tilde{X}^q_{n,i}, \forall t \in [\theta_n, \theta_{n+1}) \text{ and } i = 1, 2,$$

$$\tilde{X}^q(t) = (\tilde{X}^q_1(t), \tilde{X}^q_2(t)), \forall t \in [\theta_n, \theta_{n+1}).$$

The process $\{\tilde{X}^q(t), t \geq 0\}$ is stochastically equivalent to $\{X^q(t), t \geq 0\}$. If the matching probability is one, i.e., $r = 1 - q = 0$ and $\tilde{X}^1_{n,1} = 0$ ($\tilde{X}^1_{n,2} = 0$) then arriving class-2 (class-1) user cannot match upon arrival. Hence, we take the indeterminate form 0^0 to be 1 for notational convenience.

Let $r_1 > r_2$ (or equivalently $q_1 < q_2$) and define

$$n^* = \min \left\{ n : \tilde{X}^{q_1}_{n,1} < \tilde{X}^{q_2}_{n,1} \text{ or } \tilde{X}^{q_1}_{n,2} < \tilde{X}^{q_2}_{n,2} \right\}.$$

By definition $n^* > 0$, and without loss of generality, we can assume $\tilde{X}^{q_1}_{n^*,1} < \tilde{X}^{q_2}_{n^*,1}$, which implies $\tilde{X}^{q_1}_{n^*-1,1} = \tilde{X}^{q_2}_{n^*-1,1}$. The number of class-$i$ arrivals is independent of the matching probability q and the number of matched individuals (departures) is the same for $\tilde{X}^q_{n,1}$ and $\tilde{X}^q_{n,2}$. Hence, $\tilde{X}^{q_1}_{n^*-1,1} = \tilde{X}^{q_2}_{n^*-1,1}$ implies $\tilde{X}^{q_1}_{n^*-1,2} = \tilde{X}^{q_2}_{n^*-1,2}$. If $\tau_{n^*} = 1$, then $U_{n^*} > r_1^{\tilde{X}^{q_1}_{n^*-1,2}}$ and $U_{n^*} \leq r_2^{\tilde{X}^{q_2}_{n^*-1,2}}$, and if $\tau_{n^*} = 2$, then $U_{n^*} > r_1^{\tilde{X}^{q_1}_{n^*-1,1}}$ and $U_{n^*} \leq r_2^{\tilde{X}^{q_2}_{n^*-1,1}}$ which is not possible in either cases and leads to a contradiction. Hence, we conclude that such an n^* does not exist and $\tilde{X}^{q_1}_i(t) \geq \tilde{X}^{q_2}_i(t), i = 1, 2$, holds for every $t \geq 0$. This implies that for any $q < 1$, if $\{\tilde{X}^1_i(t), t \geq 0\}$ is transient, then $\{\tilde{X}^q_i(t), t \geq 0\}$ is also transient, and if $\{\tilde{X}^1_i(t), t \geq 0\}$ is null recurrent, then $\{\tilde{X}^q_i(t), t \geq 0\}$ is either null recurrent or transient.

Now we show that $\{X^q(t), t \geq 0\}$ is recurrent when $\lambda_1 = \lambda_2$ and $0 < q < 1$. Let $X^q_n = (X^q_{n,1}, X^q_{n,2})$ denote the corresponding embedded DTMC. Using Theorem 2.2.1 from Fayolle et al. [21] pg. 26, if we can find a finite set H and a positive function

$f(i, j)$ such that for $(i, j) \notin H$,

$$\mathbb{E}[f(X^q_{n+1,1}, X^q_{n+1,2})|(X^q_{n,1}, X^q_{n,2}) = (i, j)] - f(i, j) \leq 0$$

and $f(i, j) \to \infty$ as $i + j \to \infty$, the result follows. The transition probabilities for the embedded DTMC are

$$p_{(i,j)(k,l)} = \begin{cases} r^j/2 & \text{if } k = i + 1 \text{ and } l = j, \\ r^i/2 & \text{if } k = i \text{ and } l = j + 1, \\ (1 - r^j)/2 & \text{if } k = i \text{ and } l = j - 1 \geq 0, \\ (1 - r^i)/2 & \text{if } k = i - 1 \geq 0 \text{ and } l = j, \\ 0 & \text{otherwise.} \end{cases}$$

We choose $f(i, j)$ to be

$$f(i, j) = \begin{cases} 1 & \text{if } i = j = 0, \\ i + j - 2 & \text{if } i > 0, j > 0 \text{ and } i + j > 2, \\ i + j & \text{otherwise.} \end{cases}$$

Clearly, $f(i, j)$ is positive and $f(i, j) \to \infty$ as $i+j \to \infty$. Let $H_1 = \{(i, j) : 0 \leq i, j \leq 2\}$ and for any $(i, j) \notin H_1$ we have

$$\begin{aligned} &\mathbb{E}[f(X^q_{n+1,1}, X^q_{n+1,2})|(X^q_{n,1}, X^q_{n,2}) = (i, j)] - f(i, j) \\ &= \begin{cases} 0 & \text{if } i = 0, j > 2 \text{ or } i > 2, j = 0, \\ -r^j & \text{if } i = 1, j > 2, \\ -r^i & \text{if } i > 2, j = 1, \\ r^i + r^j - 1 & \text{if } i > 2, j > 2. \end{cases} \end{aligned}$$

Let $H_2 = \{(i, j) : 0 \leq i, j \leq \frac{\ln(1-r^2)}{\ln r}\}$, then for any $(i, j) \notin H = H_1 \cup H_2$, we have $\mathbb{E}[f(X^q_{n+1}) - f(X^q_n)|X^q_n = (i, j)] \leq 0$ and the result follows. □

Theorem 2.3 indicates that for any set of parameters, users of an uncontrolled probabilistic matching system experiences arbitrarily long waiting times. In the next section, we suggest admission control policies to stabilize probabilistic matching systems and analyse some performance measures for those policies.

2.5 Stabilizing Policies

In this section we analyse four different admission control policies to stabilize probabilistic matching systems. The policies specify the conditions under which an arriving user is allowed to enter the system and we define an admission control policy a *stabilizing policy* if it makes the probabilistic matching system ergodic.

Definition 2.4. A *stabilizing policy* is an admission control policy which makes the probabilistic matching system ergodic.

The first policy we study is the *simple threshold* policy which relies on limiting the number of users from each class in the system. The simple threshold policy stabilizes probabilistic matching systems with matching probability $q = 1$, whereas fails to provide a finite expected waiting time for users for systems with $0 < q < 1$. Analyzing the reasons behind this failure, we then suggest the *accept-the-shortest-queue* (ASQ) policy which relies on balancing the system by accepting only from the class with minimum number of users in the system. ASQ policy manages to stabilize systems with any set of arrival rates and matching probability, but may lead to very poor performance measures. Hence, we introduce the functional threshold (FT) policy by relaxing the definition of "balancing". The *one-sided threshold* (OST) policy is employed if a class of users has a higher arrival rate than the other, and imposes an upper bound on the number of users belonging to the class with higher arrival rate while always accepting the other users.

In addition to proving the stability of probabilistic matching systems under the suggested policies, we analyse how these policies affect some key performance measures. As the admission control policies are based on rejecting users under specific conditions, the long run percentage of rejected users is an important measure to assess the performance of a specific policy. A basic result in queueing theory states that the throughput of a stable system, which is defined to be the long run average rate at which users leave the system, is equal to the long run average rate at which users are admitted to the system. This implies that there is a one-to-one relationship between long run percentage of rejected users and throughput. For completeness, we prove this result for probabilistic matching systems in Theorem 2.5. We refer the reader to Asmussen [4] and El-Taha and Stidham [20] for more results of similar flavour.

Theorem 2.5. *Consider a probabilistic matching system under a stabilizing policy and let $A_i^e(t)$ and $M^q(t)$ to be the processes counting the number of type i users admitted to the system and the number of matched pairs up to time t respectively. Then,*

$$\lim_{t\to\infty}\frac{A_1^e(t)}{t}=\lim_{t\to\infty}\frac{A_2^e(t)}{t}=\lim_{t\to\infty}\frac{M^q(t)}{t},\ a.s.$$

Proof. Without loss of generality, we assume that $X^q(0)=(0,0)$. Define $\theta_0=0$,

$$\theta_j=\inf\{t\geq\theta_{j-1}:X^q(t)=(0,0)\text{ and }\exists s\text{ where }\theta_{j-1}<s<t,X^q(s)\neq(0,0)\}$$

for $j\in\mathbb{Z}_{>0}$ and $J(t)=\max\{j:\theta_j\leq t\}$. Then, for $i=1,2$

$$\begin{aligned}\frac{A_i^e(t)-M^q(t)}{t}&=\sum_{j=1}^{J(t)}\frac{A_i^e(\theta_j)-A_i^e(\theta_{j-1})-M^q(\theta_j)+M^q(\theta_{j-1})}{t}\\&\quad+\frac{A_i^e(t)-A_i^e(\theta_{J(t)})-M^q(t)+M^q(\theta_{J(t)})}{t}\\&=\frac{A_i^e(t)-A_i^e(\theta_{J(t)})-M^q(t)+M^q(\theta_{J(t)})}{t}\\&\leq\frac{A_i(\theta_{J(t)+1})-A_i(\theta_{J(t)})}{t}\\&=\frac{A_i(\theta_{J(t)+1})-A_i(\theta_{J(t)})}{J(t)+1}\frac{J(t)+1}{t}.\end{aligned}$$

The first equality above uses the fact that $A_i^e(\theta_{J(t)})=\sum_{j=1}^{J(t)}A_i^e(\theta_j)-A_i^e(\theta_{j-1})$ and $M^q(\theta_{J(t)})=\sum_{j=1}^{J(t)}M^q(\theta_j)+M^q(\theta_{j-1})$. The second equality follows as $A_i^e(\theta_j)=M^q(\theta_j)$ for all j by the definition of θ_j. Finally the inequality in the third step follows by using $M^q(\theta_{J(t)})-M^q(t)<0$ and then realizing that $A_i^e(t)-A_i^e(\theta_{J(t)})$ is the number of accepted users in $(\theta_{J(t)},t]$, whereas $A_i(\theta_{J(t)+1})-A_i(\theta_{J(t)})$ is the total number of accepted and rejected users in the same time window. As $t\to\infty$, the second term on the right hand side converges to a finite number by ergodicity and the elementary renewal theorem. Now, let $A_{i,j}=A_i(\theta_{j+1})-A_i(\theta_j)$ and $\mathbb{E}[A_{i,j}]=\lambda_i\mathbb{E}[\theta_{j+1}-\theta_j]<\infty$ (see e.g., Corollary V.6.7 in Çınlar [15]). Then for any $\epsilon>0$, $\limsup_{t\to\infty}\frac{A_i^e(t)-M^q(t)}{t}\leq\epsilon$ if and only if $\mathbb{P}[A_{i,j}>\epsilon j\text{ infinitly often}]=0$. Since we have $\sum_{j=1}^{\infty}\mathbb{P}[\frac{A_{i,j}}{\epsilon}>j]=\mathbb{E}[\frac{A_{i,j}}{\epsilon}]<\infty$,

$\mathbb{P}[A_{i,j} > \epsilon j$ infinitly often$] = 0$ follows from Borel-Cantelli lemma and the result as desired holds. □

Theorem 2.5 implies that if c_i is the long run proportion of rejected users from class-i, then

$$\lim_{t\to\infty} \frac{A_1^e(t)}{t} = (1-c_1)\lambda_1 = (1-c_2)\lambda_2 = \lim_{t\to\infty} \frac{A_2^e(t)}{t} \tag{2.4}$$

which is also equal to the throughput of the system. Moreover, using Poisson-arrivals-see-time-averages (PASTA) property (see El-Taha and Stidham [20]), c_i also equals to the total stationary probability of being at a state where class-i users are rejected.

2.5.1 The Simple Threshold Policy

The simple threshold policy imposes a constant upper bound on the number of users in the system for each class. The aim is to reduce the process $\{X^q(t), t \geq 0\}$ to a finite state space irreducible CTMC which would be always ergodic.

Definition 2.6. A simple threshold policy is an admission control policy which admits a class-i user arriving at time t if and only if $X_i^q(t-) \leq N_i$, where $i = 1, 2$ and $0 \leq N_i < \infty$.

Theorem 2.7. *When $q = 1$, the simple threshold policy stabilizes a probabilistic matching system for any $0 < \lambda_i < \infty$, and $0 \leq N_i < \infty$ where $i = 1, 2$.*

Proof. The system with $q = 1$ is reduced to an irreducible one-dimensional CTMC with finite state space $\mathbb{S} = \{-N_2 - 1, -N_2, ..., -1, 0, 1, ..., N_1, N_1 + 1\}$ under the simple threshold policy and hence ergodic (see Kulkarni [31], pg. 82, Theorem 3.7 and pg. 285, Theorem 6.10). □

The admitted users leave the system only when there is a continuing intake of users from the other class. When $q = 1$, users of different classes cannot co-exist in the system. Hence, when the number of users of a certain class reaches its maximum, the arrivals from the other class are always admitted. However, when $0 < q < 1$, the resulting CTMC is no longer irreducible. Once the system reaches state (N_1+1, N_2+1), no users are admitted under the simple threshold policy and hence no users leave the system resulting in infinite waiting times. Therefore, better control mechanisms are needed when $0 < q < 1$.

Now, we investigate the long run average proportion of rejected users when $q = 1$ and the simple threshold policy is employed. Even though, the simple threshold policy is not applicable when $q < 1$, it provides us with an interesting benchmark to compare other control policies.

Theorem 2.8. *Let c_i be the long run average proportion of rejected class i users for $i = 1, 2$. If $q = 1$ and the simple threshold policy is employed, then*

$$c_1 = \begin{cases} \frac{1}{N_1+N_2+3}, & \text{for } \lambda_1 = \lambda_2, \\ \frac{1-\frac{\lambda_2}{\lambda_1}}{1-(\frac{\lambda_2}{\lambda_1})^{N_1+N_2+3}}, & \text{for } \lambda_1 \neq \lambda_2, \end{cases} \tag{2.5}$$

and c_2 is given by interchanging λ_1 and λ_2 in (2.5).

Proof. Let $\{X^{1,ST}(t), t \geq 0\}$ be the resulting stochastic process, then c_1 and c_2 correspond to the stationary probabilities for states $-N_1 - 1$ and $N_2 + 1$, respectively. The process $\{X^{1,ST}(t) + N_1 + 1, t \geq 0\}$ is stochastically identical to an $M/M/1/N_1 + N_2 + 2$ queueing process and the result follows (see, Gross and Harris [24] pg. 77). □

By considering the extreme cases $N_1 = N_2 = 0$ and $N_1, N_2 \to \infty$ in equation (2.5), we obtain bounds on c_1. In particular,

$$\max(1 - \frac{\lambda_2}{\lambda_1}, 0) \leq c_1 \leq \frac{\lambda_1^2}{\lambda_1^2 + \lambda_1\lambda_2 + \lambda_2^2}. \tag{2.6}$$

These bounds act as a benchmark when evaluating other policies.

2.5.2 The Accept-the-Shortest-Queue Policy

In the previous section we have seen that assuring continuing intakes of both classes of users plays a key role in avoiding any absorbing states. Therefore, rather than imposing strict bounds, the ASQ policy tries to maintain a balance between different classes by only admitting users belonging to the shorter queue. As the system size increases, both classes of users get more likely to find a match upon arrivals.

Definition 2.9. The accept-the-shortest-queue policy is an admission control policy which admits a class-i user arriving at time t if and only if $X_i^q(t-) = \min_{j\in\{1,2\}} X_j^q(t-), i = 1, 2$.

Let $\{X^{q,ASQ}(t), t \geq 0\}$ be the CTMC representing the probabilistic matching system under the ASQ policy. Then the state space is $\mathbb{S} = \{-1, 0, 1\}$ when $q = 1$, and $\mathbb{S} = \{(i, j) \in \mathbb{N}^2 : |i - j| \leq 1\}$ when $0 < q < 1$. We now prove the stability of the system under the ASQ policy using Foster's criterion (see e.g., Fayolle et al. [21], pg. 29, Theorem 2.2.3 or Kulkarni [31], pg. 95. Theorem 3.10).

Theorem 2.10. *A probabilistic matching system is stable for any set of arrival rates* λ_1 *and* λ_2 *and matching probability* $0 < q \leq 1$ *under the ASQ policy.*

Proof. When $q = 1$, the state space is finite and the result follows. When $0 < q < 1$, define $\{X_n^{q,ASQ}, n \in \mathbb{N}\}$ to be the corresponding embedded DTMC and $f(i, j) = i+j+1$. The transition probabilities for the embedded DTMC are

$$p_{(i,j)(k,l)} = \begin{cases} \lambda_1 r^j/(\lambda_1+\lambda_2) & \text{if } k = i+1 \text{ and } l = j = i, \\ \lambda_2 r^i/(\lambda_1+\lambda_2) & \text{if } k = i = j \text{ and } l = j+1, \\ \lambda_1(1-r^j)/(\lambda_1+\lambda_2) & \text{if } k = i = j \text{ and } l = j-1 \geq 0, \\ \lambda_2(1-r^i)/(\lambda_1+\lambda_2) & \text{if } k = i-1 \geq 0 \text{ and } l = j = i, \\ r^j & \text{if } k = l = j = i+1, \\ 1-r^j & \text{if } k = l = i = j-1 \geq 0, \\ r^i & \text{if } k = l = i = j+1, \\ 1-r^i & \text{if } k = l = j = i-1 \geq 0. \\ 0 & \text{otherwise.} \end{cases}$$

Then, $f(i, j)$ is positive for all states and

$$\mathbb{E}[f(X_{n+1,1}^{q,ASQ}, X_{n+1,2}^{q,ASQ})|X_n^{q,ASQ} = (i,j)] - f(i,j) = \begin{cases} 2r^i - 1, & \text{if } i > j, \\ 2r^j - 1, & \text{if } i \leq j. \end{cases}$$

Let $m = \min\{i \in \mathbb{N} : r^i < \frac{1}{2}\}$. Then, for the finite set $\mathbb{H} = \{(i, j) \in \mathbb{N}^2, 0 \leq i, j \leq m\} \subset \mathbb{S}$ and $\epsilon = \frac{1}{2} - r^m > 0$, we have $\mathbb{E}[f(X_{n+1,1}^{q,ASQ}, X_{n+1,2}^{q,ASQ})|X_n^{q,ASQ} = (i,j)] - f(i,j)| < 2$, if $(i,j) \in \mathbb{H}$, and $\mathbb{E}[f(X_{n+1,1}^{q,ASQ}, X_{n+1,2}^{q,ASQ}) | X_n^{q,ASQ} = (i,j)] - f(i,j) < -2\epsilon$ if $(i,j) \notin \mathbb{H}$. Thus, all conditions in Foster's Criterion are satisfied, and so the system is ergodic. □

The ergodicity of the CTMC guarantees the existence of stationary probabilities. Theorem 2.11 provides an explicit representation for the stationary probabilities.

Theorem 2.11. *The stationary probabilities under the ASQ when* $0 < q < 1$ *are*

$$p_{i,j} = a_{ij} p_{0,0}, \ \textit{for} \ (i,j) \in \mathbb{S},$$

where

$$a_{ij} = \begin{cases} 1 & \textit{if } i = j = 0 \\ \dfrac{r^{i^2}}{\left[\prod_{k=1}^{i}(1-r^k)\right]^2} & \textit{if } i \geq 1, j = i \\ \dfrac{\lambda_1}{\lambda_2(1-r)} & \textit{if } i = 1, j = 0 \\ \dfrac{\lambda_2}{\lambda_1(1-r)} & \textit{if } i = 0, j = 1 \\ \dfrac{\lambda_2 r^{i(i+1)} p_{0,0}}{\lambda_1 \prod_{k=1}^{i}(1-r^k) \prod_{k=1}^{i+1}(1-r^k)} & \textit{if } i \geq 1, j = i+1 \\ \dfrac{\lambda_1 r^{i(i-1)} p_{0,0}}{\lambda_2 \prod_{k=1}^{i}(1-r^k) \prod_{k=1}^{i-1}(1-r^k)} & \textit{if } j \geq 1, i = j+1 \end{cases}$$

and $p_{0,0} = \dfrac{1}{1 + \sum_{i=1}^{\infty} \sum_{j=i-1}^{i+1} a_{ij}}$.

Proof. See Appendix B. □

In principle, the performance measures of a probabilistic matching system under the ASQ policy can be calculated using the stationary probabilities presented in Theorem 2.11. In the next section we generalize the ASQ policy and use another method to calculate the long run percentage of rejected users and present some insights about average queue lengths and waiting times for a more general set of policies including the ASQ policy.

2.5.3 The Functional Threshold Policy

In this section we generalize the idea of applying a "moving" threshold behind the ASQ policy. Instead of applying the threshold as the number of users in the other queue, the functional threshold (FT) policy sets the threshold to be a *function* of the number of users in the other queue.

Definition 2.12. An admission control policy is called a functional threshold policy if it admits a class-i user arriving at time t when $X_i^q(t-) \leq h(X_j^q(t-))$, $i, j = 1, 2, i \neq j$, where $h(\cdot) : \mathbb{N} \to \mathbb{R}$ is a non-decreasing function and satisfies $x \leq h(x) < \infty$, for all $x \geq 0$.

The threshold function $h(\cdot)$ makes the FT policy more flexible compared to the ASQ Policy. When the threshold function is set to be $h(x) = x$, the functional threshold policy is equivalent to the ASQ policy. The condition $x \leq h(x) < \infty$, for all $x \geq 0$ prevents selecting inappropriate threshold functions (e.g., $h(x) = N$, as seen in the ST policy), and implies that $\min_{j\in\{1,2\}} X_j^q(t-) \leq X_i^q(t-) \leq h(X_i^q(t-)), i = 1, 2$. Hence, we always accept the users from the class with the shortest queue and assure a continuing intake to avoid absorbing states.

Theorem 2.13. *A functional threshold policy is a stabilizing policy for probabilistic matching systems with any set of arrival rates $\lambda_1 > 0$ and $\lambda_2 > 0$, and matching probability $0 < q \leq 1$.*

Proof. When $q = 1$, the resulting CTMC is irreducible with a finite state space and hence stable. When $0 < q < 1$, the state space of the CTMC is $\mathbb{S} = \{(i, j), i \leq \max(h(j + 1), h(j) + 1), j \leq \max(h(i + 1), h(i) + 1)\}$ and we apply Foster's criterion on the embedded DTMC. To write down the transition probabilities for the embedded DTMC, we assume $i \leq j$ and consider the following cases:

When $h(i) + 1 \leq j \leq \max\{h(i) + 1, h(i + 1)\}$,

$$p_{(i,j)(k,l)} = \begin{cases} r^j & \text{if } k = i + 1 \text{ and } j = l, \\ 1 - r^j & \text{if } k = i \text{ and } l = j - 1, \\ 0 & \text{otherwise.} \end{cases}$$

When $i \leq j \leq h(i)$,

$$p_{(i,j)(k,l)} = \begin{cases} \frac{1}{\lambda_1+\lambda_2}\lambda_2 r^i & \text{if } k=i \text{ and } j=l+1, \\ \frac{1}{\lambda_1+\lambda_2}\lambda_2(1-r^i) & \text{if } k=i-1 \text{ and } l=j, \\ \frac{1}{\lambda_1+\lambda_2}\lambda_1 r^j & \text{if } k=i+1 \text{ and } l=j, \\ \frac{1}{\lambda_1+\lambda_2}\lambda_1(1-r^j) & \text{if } k=i \text{ and } l=j-1, \\ 0 & \text{otherwise.} \end{cases}$$

The transition probabilities for the states where $i > j$ can be obtained by interchanging i and j. Next, we define $f(i,j) = i+j+1$, then $f(i,j)$ is positive for all states and

$$\begin{aligned} &\mathbb{E}[f(X^{q,FT}_{n+1,1}, X^{q,FT}_{n+1,2})|X^{q,FT}_n = (i,j)] - f(i,j) \\ &= \begin{cases} 2r^i - 1 & \text{if } h(i)+1 \leq j \leq \max\{h(i)+1, h(i+1)\}, \\ 2r^j - 1 & \text{if } h(j)+1 \leq i \leq \max\{h(j)+1, h(j+1)\}, \\ \dfrac{\lambda_2(2r^i-1)+\lambda_1(2r^j-1)}{\lambda_1+\lambda_2} & \text{if } i \leq j \leq h(i) \text{ or } j \leq i \leq h(j). \end{cases} \end{aligned}$$

Let $m = \min\{i : r^i < \frac{1}{2}\}$ and $\epsilon = 1 - 2r^m > 0$. Then when $(i,j) \notin \mathbb{H} = \{(k,l) \in \mathbb{N}^2 : k \leq m \text{ or } l \leq m\}$, we have $\mathbb{E}[f(X^{q,FT}_{n+1,1}, X^{q,FT}_{n+1,2})|X^{q,FT}_n = (i,j)] - f(i,j) < -\epsilon$. Hence, the result follows. □

Under the functional threshold policy, stating and solving the global balance equations is fairly difficult. However, it is still possible to obtain insights about some key performance measures of the system by imposing some mild restrictions on the threshold function $h(x)$.

Theorem 2.14. *Suppose that the functional threshold policy is employed with the threshold function $h(x) = x + d$, where $d \geq 0$ is an arbitrary constant. Then, the long run percentage of rejected users, c_1 and c_2, are independent of the matching probability $0 < q \leq 1$ and*

$$c_1 = \begin{cases} \frac{1}{2\lceil d\rceil+3}, & \text{for } \lambda_1 = \lambda_2, \\ \frac{1-\frac{\lambda_2}{\lambda_1}}{1-(\frac{\lambda_2}{\lambda_1})^{2\lceil d\rceil+3}}, & \text{for } \lambda_1 \neq \lambda_2, \end{cases} \tag{2.7}$$

and c_2 can be obtained interchanging λ_1 and λ_2.

Proof. See Appendix C. □

Corollary 2.15. *Under the ASQ policy, the long run percentage of rejected users is independent of the matching probability and*

$$c_1 = \frac{\lambda_1^2}{\lambda_1^2 + \lambda_1\lambda_2 + \lambda_2^2}, \tag{2.8}$$

c_2 is obtained interchanging λ_1 and λ_2.

Proof. The result follows by replacing $d = 0$ in (2.7) and cancelling as appropriate. □

One may expect that as the matching probability q increases, the users match more quickly and this yields a better throughput and hence smaller long run percentage of rejected users. However, contrary to this initial intuition, Theorem 2.14 indicates that under the specified subclass of functional threshold policies, the long run average percentage of rejected users does not depend q. In Section 2.6, we numerically show that for many reasonable threshold functions the behaviour of the system is even more counter-intuitive, i.e., the long run rejection probabilities actually increase as the matching probability increases! Further discussions about the reasons behind this unexpected phenomenon are provided in Section 2.6.

Unlike the throughput, the long run average number of users in the system and the average waiting times depend on the matching probability q, even when $h(x) = x + d$. However, we are able to prove similar insensitivity results for the difference between the long run average numbers of users from different classes for the same class of threshold functions.

Theorem 2.16. *Let L_i^q denote the long run average numbers of class-i user, $i = 1, 2$ in the system and $\rho = \lambda_2/\lambda_1$. If the functional threshold policy is employed with $h(x) = x + d$, where $d \geq 0$ is a constant, then for any $0 < q \leq 1$, the difference between average queue lengths of classes, $L_1^q - L_2^q$, does not depend on the matching probability q, and we have $L_1^q - L_2^q = 0$ if $\lambda_1 = \lambda_2$ and*

$$L_1^q - L_2^q = \frac{(d+2)\rho^{2d+3} + d + 1}{1 - \rho^{2d+3}} + \frac{(1-\rho)\rho^{d+2}(\rho^{d+2} - \rho^{-d-1})}{(1-\rho^{2d+3})(\rho-1)^2} \tag{2.9}$$

if $\lambda_1 \neq \lambda_2$.

Proof. See Appendix C. □

Corollary 2.17. *Suppose the functional threshold policy is employed with $h(x) = x+d$, where $d \geq 0$ is a constant. Let W_1^q and W_2^q denote the long run average waiting times for users, $\rho = \lambda_2/\lambda_1$ and $\lambda^e \equiv (1-c_1)\lambda_1$, where c_1 is as in (2.7), then for any $0 < q \leq 1$, the difference between average waiting times of classes, $W_1^q - W_2^q$, does not depend on the matching probability q, and we have $W_1^q - W_2^q = 0$ if $\lambda_1 = \lambda_2$ and*

$$W_1^q - W_2^q = \frac{1}{\lambda^e}\left(\frac{(d+2)\rho^{2d+3} + d + 1}{1 - \rho^{2d+3}} + \frac{(1-\rho)\rho^{d+2}(\rho^{d+2} - \rho^{-d-1})}{(1-\rho^{2d+3})(\rho-1)^2}\right)$$

if $\lambda_1 \neq \lambda_2$.

Proof. Using PASTA property (see El-Taha and Stidham [20], Corollary 1.10 and Theorem 3.23), Little's Law and (2.4), $W_i^q = L_i^q/\lambda^e$ for $i = 1, 2$, and hence the result follows from Theorem 2.16. □

Functional threshold policy relies on rejecting both types of users in a similar fashion. We next introduce a policy which rejects only one type of user.

2.5.4 The One-Sided Threshold Policy

The policies we discuss so far reject both classes of users when their numbers reach certain limits. If it is known that the arrival rate of a class is less than the arrival rate of the other (e.g., $\lambda_1 < \lambda_2$), it may not be reasonable to ever reject that class of users. For example, in general, the rate of employers arriving at an employment portal is significantly less than the arrival rate of employees. Thus, each job posting is deemed valuable and the employment portal would not want to lose any employer who wishes to subscribe. In such a matching system, it is more reasonable to reject only employees when the number of them reaches a certain threshold N_2, but to always accept employers .

Definition 2.18. When $\lambda_1 < \lambda_2$, a one-sided threshold (OST) policy admits users of class-2 at time t if and only if $X_2^q(t-) \leq N_2$, whereas users of class-1 are always admitted.

Theorem 2.19. *A probabilistic matching system with arrival rates $\lambda_1 < \lambda_2$ and matching probability $0 < q \le 1$ is ergodic under a one-sided threshold policy which applies a finite threshold $N_2 \ge 0$ to class-2 users.*

Proof. When $q = 1$, the one-dimensional CTMC has a state space the set of integers from $-\infty$ to $N_2 + 1$. Since $\frac{\lambda_1}{\lambda_2} < 1$, the system is ergodic. When $0 < q < 1$, we have the state space $\mathbb{S} = \{(i,j) \in \mathbb{N}^2 : i \ge 0, 0 \le j \le N_2 + 1\}$. The transition probabilities for the embedded DTMC under one-sided threshold policy are

$$p_{(i,j)(k,l)} = \begin{cases} \lambda_1 r^j/(\lambda_1+\lambda_2) & \text{if } k = i+1 \text{ and } l = j \le N_2, \\ \lambda_2 r^i/(\lambda_1+\lambda_2) & \text{if } k = i \text{ and } l = j+1 \le N_2+1, \\ \lambda_1(1-r^j)/(\lambda_1+\lambda_2) & \text{if } k = i \text{ and } 0 \le l = j-1 \le N_2 - 1, \\ \lambda_2(1-r^i)/(\lambda_1+\lambda_2) & \text{if } k = i-1 \ge 0 \text{ and } l = j \le N_2, \\ r^j & \text{if } k = i+1 \text{ and } l = j = N_2+1, \\ 1 - r^j & \text{if } k = i \text{ and } l = j-1 = N_2, \\ 0 & \text{otherwise.} \end{cases}$$

For $0 < q < 1$, we can always find a positive number a, such that, $r < a(1-r)$ (recall that $r = 1 - q$). Thus, the inequality $r^j < a(1-r^j)$ holds for all j and in particular, for some $\epsilon_0 > 0$, $r^{N_2+1} - a(1-r^{N_2+1}) < -\epsilon_0$. For any state $(i,j) \in \mathbb{S}$, define $f(i,j) = i + aj + 1$ and $d(i,j) = \mathbb{E}[f(X^{q,OST}_{n+1,1}, X^{q,OST}_{n+1,2}) | X^{q,OST}_n = (i,j)] - f(i,j)$. Then, for all $i \ge 0$, $d(i, N_2+1) = r^{N_2+1} - a(1-r^{N_2+1}) < -\epsilon_0$. Also, for all $i \ge 0$, $d(i,0) = \frac{1}{\lambda_1+\lambda_2}(\lambda_1 - \lambda_2 + a\lambda_2 r^i + \lambda_2 r^i)$. Since, $\lambda_1 < \lambda_2$, there exists an m_1 and $\epsilon_1 > 0$ such that $d(i,0) < -\epsilon_1$. If $N_2 \ge 1$, then for $i \ge 1$ and $1 \le j \le N_2$

$$d(i,j) = \frac{1}{\lambda_1+\lambda_2}(\lambda_1 a r^i - \lambda_1(1-r^i) + \lambda_2 r^j - \lambda_2(1-r^j)a) < \frac{\lambda_1}{\lambda_1+\lambda_2}(r^i(a+1) - 1).$$

Thus, there exists an $m_2 > 0$ and $\epsilon_2 > 0$, such that, when $i > m_2$, $d(i,j) = -\epsilon_2 < 0$. Take $m_2 = 0$ when $N_2 = 0$, and let $m = \max\{m_1, m_2\}$ and $\epsilon = \min\{\epsilon_0, \epsilon_1, \epsilon_2\}$. Then, for the finite set $\mathbb{H} = \{(i,j) \in \mathbb{N}^2, 0 \le i \le m, 0 \le j \le N_2\}$ we have, $d(i,j) < -\epsilon, (i,j) \notin \mathbb{H}$. The conditions of Foster's criterion are satisfied and hence the system is stable. □

Theorem 2.19 states that the stability of the matching system neither depends on the matching probability nor the threshold N_2. As a consequence of Theorem 2.5, we prove that the long run average percentage of rejected users of a probabilistic matching system under a one-sided threshold policy is also independent of the matching probability q and the threshold N_2.

Theorem 2.20. *Suppose a one-sided threshold policy with threshold N_2 is employed on a probabilistic matching system with arrival rates $0 < \lambda_1 < \lambda_2 < \infty$ and matching probability $0 < q \leq 1$. Then, the long run proportion of rejected users is independent of both the matching probability q and the threshold N_2 and is given by*

$$c_2 = 1 - \frac{\lambda_1}{\lambda_2}.$$

Proof. Since, no class-1 user is rejected, using Theorem 2.5,

$$\lim_{t\to\infty} \frac{A_1^e(t)}{t} = \lim_{t\to\infty} \frac{A_2^e(t)}{t} = \lambda_1, \ a.s.$$

The result follows from (2.4). □

Unlike the percentage of rejected users, most other performance measures, such as the average waiting time or the average queue length, depend on both the matching probability and the threshold. We analyse these quantities numerically in Section 2.6.

The one-sided threshold policy achieves the lower bound in (2.6) which is the best rejection rate possible as rejecting less users will definitely yield an unstable system. As there are always rejected users from both classes in functional threshold policy, the same performance cannot be attained under any threshold function. On the other hand, the rejection percentage under the ASQ policy is equal to the upper bound in (2.6), which is the worst rejection percentage possible under the simple threshold policy when $q = 1$.

2.6 Numerical Results

We have seen that it is rather difficult to derive explicit equations for the performance measures of the probabilistic matching systems under the suggested stabilizing policies.

This is partly due to the transition rates involving powers of the matching probability. In this section, we present a numerical analysis of the performance of probabilistic matching systems under different policies. Some of our results appear to be quite counter-intuitive and we present explanations for these results.

2.6.1 Computational Experiments on Long Run Percentages of Rejected Users

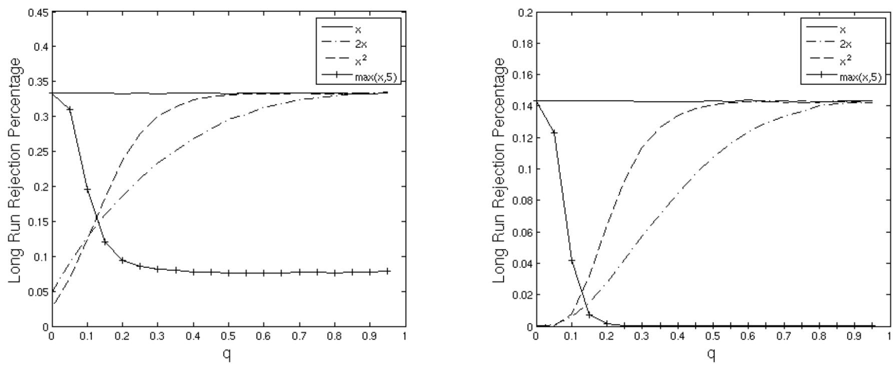
Theorem 2.20 shows that if one-sided threshold policy is employed the long run rejection rate is insensitive to both the matching probability and the threshold value. Similarly, Theorem 2.14 shows that if the functional threshold policy is employed with a specific type of threshold function, the long run rejection rate of a probabilistic matching system is insensitive to the matching probability. In this section, we present simulation results to see how the rejection rate behaves under various functional threshold policies.

We consider two probabilistic matching systems, where the first system has arrival rates $\lambda_1 = \lambda_2 = 1$ and the second system has $\lambda_1 = 1$ and $\lambda_2 = 2$, to test how the matching probability affects the long run percentages of rejected users. We simulate 10 replications in each experiment, where each replication covers 1,000,000 time units. We compare four different threshold functions: $h_1(x) = x, h_2(x) = 2x, h_3(x) = x^2$ and $h_4(x) = \max\{5, x\}$. Note that the first threshold function $h_1(x) = x$ corresponds to the ASQ policy.

Table 2.1 and Figure 2.1 summarize the corresponding results. As proven in Theorem 2.14, the first column of Table 2.1 demonstrates the insensitivity of the long run rejection probabilities to the matching probability q under the ASQ policy. The situation for general threshold functions is even more surprising. One may intuitively guess that as the matching probability increases, the users match faster and as a result better performance for rejection rates can be achieved. Contrary to this initial intuition that we would observe lower rejection percentages for higher matching probabilities, we discover that for $h_2(x) = 2x$ and $h_3(x) = x^2$ the long run rejection percentages actually *increase* as the matching probability q increases. The rejection percentages are very close to 0 when q is close to 0, and as q increases to 1, they converge to that of the ASQ policy. However, for the threshold function $h_4(x) = \max\{5, x\}$, we notice an opposite

Table 2.1: Long run percentage of rejected users for functional threshold policies with various threshold functions when $\lambda_1 = \lambda_2 = 1$

q	x	$2x$	x^2	$\max\{x, 5\}$
0.10	0.333	0.091	0.069	0.310
0.20	0.333	0.159	0.183	0.120
0.30	0.333	0.211	0.275	0.086
0.40	0.333	0.252	0.315	0.080
0.50	0.332	0.282	0.329	0.077
0.60	0.333	0.303	0.332	0.076
0.70	0.334	0.319	0.333	0.077
0.80	0.333	0.326	0.333	0.077
0.90	0.333	0.332	0.333	0.078
1.00	0.333	0.334	0.334	0.078

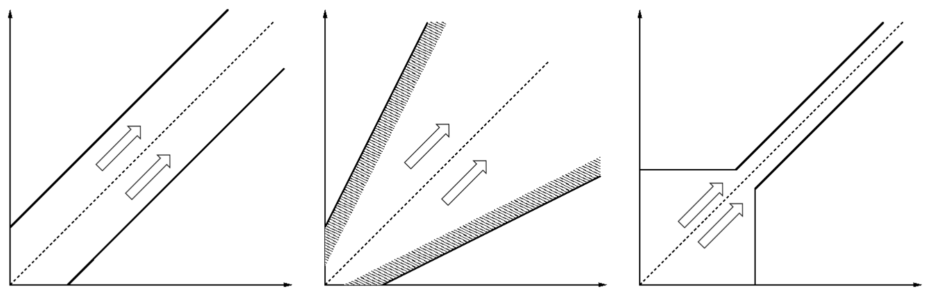


(a) $\lambda_1 = \lambda_2 = 1$ (b) $\lambda_1 = 1, \lambda_2 = 2$

Figure 2.1: Matching probability q vs. long run percentage of rejected class-1 users for functional threshold policies with different threshold functions

behaviour which matches our initial intuition, i.e., the rejection probabilities are very close to that of the ASQ policy when q is close to 0 and it *decreases* as q increases.

We explain this surprising behaviour using Figure 2.2 which illustrates the state space of the CTMC under various functional threshold policies. The boundaries in Figures 2.2(a) and 2.2(c) shown with bold lines correspond to the regions where a class of users are rejected. For $h_2(x) = 2x$, the shaded region corresponds to the states where users are rejected. As q decreases, we see that the probability mass of the stationary distribution moves in the direction of the arrows shown in the figures, somewhat parallel to the diagonal illustrated by the dashed line. For the threshold function $h_2(x) = 2x$, the random walk is pushed towards a wider region as q decreases, and the proportion of time spent in rejection region decreases. For $h_4(x) = \max\{x, 5\}$, the situation is the



(a) $h_1(x) = x + d$ (b) $h_2(x) = 2x$ (c) $h_4(x) = \max\{x, 5\}$

Figure 2.2: State space for the CTMC under functional threshold policies with various threshold functions when $\lambda_1 = \lambda_2 = 1$

Table 2.2: Average queue lengths and waiting times for functional threshold policies with various threshold functions when $\lambda_1 = \lambda_2 = 1$

	$h_1(x) = x$		$h_2(x) = 2x$		$h_3(x) = x^2$		$h_4(x) = \max\{x, 5\}$	
q	L_1	W_1	L_1	W_1	L_1	W_1	L_1	W_1
0.1	6.458	9.680	6.532	7.188	6.655	7.148	6.655	9.512
0.2	2.998	4.497	3.009	3.575	2.945	3.607	3.311	3.766
0.3	1.842	2.766	1.831	2.318	1.767	2.440	2.370	2.591
0.4	1.265	1.898	1.235	1.650	1.228	1.794	2.000	2.174
0.5	0.924	1.387	0.889	1.239	0.911	1.358	1.813	1.965
0.6	0.702	1.053	0.671	0.964	0.699	1.048	1.727	1.870
0.7	0.550	0.826	0.530	0.778	0.550	0.826	1.681	1.822
0.8	0.446	0.669	0.437	0.651	0.447	0.670	1.653	1.791
0.9	0.377	0.566	0.375	0.561	0.377	0.566	1.629	1.766
1.0	0.333	0.500	0.333	0.500	0.334	0.501	1.613	1.747

opposite, i.e., the walk is pushed to narrower areas and the proportion of time spent in the rejection region increases as q decreases. This explanation is also in accordance with Theorem 2.14, as the width of the state space is constant for $h_1(x) = x + d$ as seen in Figure 2.2(a).

2.6.2 Computational Experiments on Average Queue Lengths and Average Waiting Times

We now turn to the study of long run average queue lengths (L_i) and waiting times (W_i). Our simulations use the same structure described in Section 2.6.1. Our first set of experiments analyse how changing the matching probability q affects our parameters under the functional threshold policy. The results are presented in Table 2.2 and Table 2.3.

As expected, we see that the long run average queue lengths decrease as q increases.

Table 2.3: Long run average queue length and waiting times for functional threshold policies with various threshold functions when $\lambda_1 = 1$ and $\lambda_2 = 2$

	$h_1(x) = x$						$h_2(x) = 2x$					
q	L_1	L_2	$L_2 - L_1$	W_1	W_2	$W_2 - W_1$	L_1	L_2	$L_2 - L_1$	W_1	W_2	$W_2 - W_1$
0.1	6.249	6.678	0.429	7.297	7.798	0.501	4.531	8.759	4.228	4.537	8.771	4.234
0.2	2.776	3.206	0.430	3.241	3.743	0.502	2.096	4.006	1.910	2.129	4.070	1.941
0.3	1.630	2.058	0.428	1.902	2.401	0.499	1.250	2.458	1.208	1.305	2.565	1.260
0.4	1.053	1.482	0.429	1.228	1.727	0.499	0.812	1.690	0.878	0.874	1.819	0.945
0.5	0.715	1.143	0.428	0.834	1.333	0.499	0.557	1.243	0.686	0.616	1.376	0.760
0.6	0.496	0.924	0.428	0.578	1.077	0.499	0.397	0.965	0.568	0.450	1.092	0.642
0.7	0.348	0.776	0.428	0.406	0.906	0.500	0.293	0.788	0.495	0.336	0.904	0.568
0.8	0.248	0.676	0.428	0.289	0.789	0.500	0.224	0.677	0.453	0.260	0.786	0.526
0.9	0.182	0.611	0.429	0.213	0.714	0.501	0.176	0.611	0.435	0.206	0.712	0.506
1.0	0.143	0.572	0.429	0.167	0.667	0.500	0.143	0.572	0.429	0.167	0.668	0.501
	$h_3(x) = x^2$						$h_4(x) = \max\{x, 5\}$					
q	L_1	L_2	$L_2 - L_1$	W_1	W_2	$W_2 - W_1$	L_1	L_2	$L_2 - L_1$	W_1	W_2	$W_2 - W_1$
0.1	3.475	10.467	6.992	3.473	10.458	6.985	6.243	6.851	0.608	7.118	7.811	0.693
0.2	2.063	3.928	1.865	2.131	4.058	1.927	1.898	5.158	3.260	1.911	5.195	3.284
0.3	1.367	2.199	0.832	1.507	2.422	0.915	0.604	5.084	4.480	0.603	5.079	4.476
0.4	0.969	1.502	0.533	1.109	1.719	0.610	0.214	5.061	4.847	0.214	5.060	4.846
0.5	0.689	1.143	0.454	0.801	1.328	0.527	0.085	5.041	4.956	0.085	5.038	4.953
0.6	0.490	0.923	0.433	0.571	1.075	0.504	0.039	5.032	4.993	0.040	5.037	4.997
0.7	0.345	0.775	0.430	0.403	0.905	0.502	0.024	5.023	4.999	0.024	5.018	4.994
0.8	0.248	0.676	0.428	0.289	0.789	0.500	0.018	5.021	5.003	0.018	5.020	5.002
0.9	0.183	0.610	0.427	0.213	0.711	0.498	0.016	5.017	5.001	0.016	5.017	5.001
1.0	0.143	0.571	0.428	0.167	0.667	0.500	0.015	5.013	4.998	0.015	5.009	4.994

The average queue lengths under all threshold functions behave similarly. The only exception is that, if q is close to 1, the average queue lengths are significantly higher under $h_4(x)$, which is expected because when $q = 1$, the number of users is bounded by 5 for this threshold function, whereas the others are bounded by 1. We also observe that the average waiting times for $h_1(x)$ and $h_4(x)$ are quite high for small q due to the poor throughput. When q is close to 1, we observe that the average waiting times under $h_4(x)$ is still higher, because even though the throughput for this threshold function is higher than the others, the average queue length is still relatively higher.

Theorem 2.16 proves that under specific functional threshold policies, the difference between average queue lengths of different types of customers is constant with respect to q. When $\lambda_1 = \lambda_2$, this is trivially true for any functional threshold policy, as due to symmetry $L_1 - L_2 = 0$. When $\lambda_1 = 1$ and $\lambda_2 = 2$, the observations as presented in Table 2.3, are parallel to those related to the rejection probabilities. For the threshold functions where the rejection probabilities are insensitive to q, we see that the difference between average queue lengths is also insensitive to q. For the threshold functions where the rejection probabilities are increasing (decreasing) with respect to q, the difference $L_2 - L_1$ is decreasing (increasing).

Table 2.4: Long run average queue length under the one-sided threshold policy for varying q

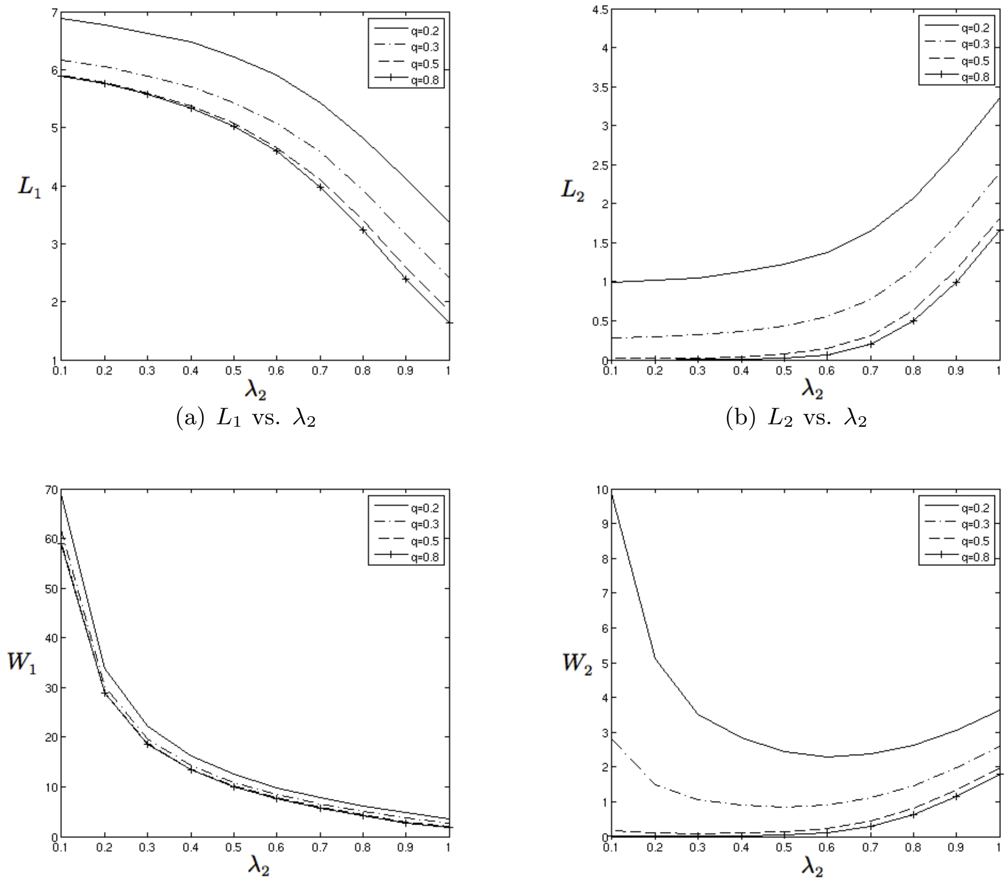
	$N_2 = 3$		$N_2 = 5$	
q	L_1	L_2	L_1	L_2
0.1	12.847	3.219	8.851	5.152
0.2	3.429	3.192	1.933	5.116
0.3	1.327	3.167	0.604	5.086
0.4	0.597	3.139	0.213	5.066
0.5	0.291	3.121	0.086	5.043
0.6	0.161	3.100	0.041	5.032
0.7	0.102	3.085	0.025	5.024
0.8	0.077	3.075	0.019	5.019
0.9	0.067	3.067	0.017	5.016
1	0.063	3.060	0.016	5.013

Table 2.5: Long run average queue length under one sided threshold policy for varying N_2

	$q = 0.1$		$q = 0.4$	
N_2	L_1	L_2	L_1	L_2
0	33.726	0.499	4.128	0.500
1	21.551	1.350	1.875	1.320
2	16.149	2.270	1.023	2.211
3	12.822	3.220	0.596	3.140
4	10.557	4.182	0.354	4.096
5	8.852	5.155	0.215	5.060
6	7.537	6.128	0.131	6.039
7	6.468	7.109	0.080	7.021
8	5.599	8.102	0.048	8.018
9	4.856	9.087	0.029	9.015
10	4.262	10.075	0.018	10.007

Next, we study the average queue lengths and the average waiting times under the one-sided threshold policy and see how they depend on N_2 and q. We assume that $\lambda_1 = 1$ and $\lambda_2 = 2$ and vary q and N_2. Since the average waiting times can be calculated as the products of the average queue lengths and the throughput (which is 1 in our case), we present the results only for the average queue lengths in Tables 2.4 and 2.5. We see that the average queue length (L_1) of the user class with the lower arrival rate is highly sensitive to the changes in matching probability and the threshold, and decreases as these quantities increase. On the other hand, the average queue length for the class with higher arrival rate (L_2) is less sensitive to the changes in the matching probability and the threshold. Under the one-sided threshold policy, the number of class-2 users is bounded by $N_2 + 1$ and we observe that average queue length is in general very close to this upper bound and increases almost linearly as N_2 increases.

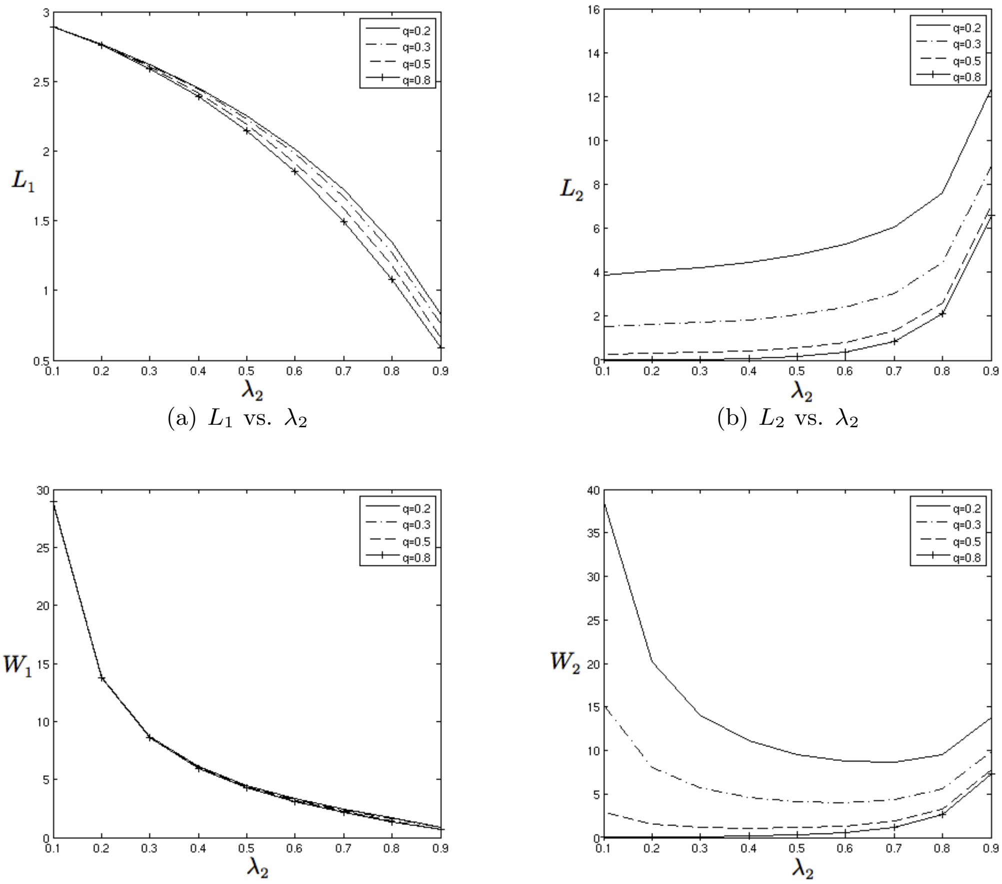
The one-sided threshold policy requires one of the arrival rates to be strictly greater than the other, e.g., $\lambda_1 > \lambda_2$. Next, we investigate how the ratio of arrival rates λ_2/λ_1 affects the average queue lengths and waiting times under different control policies. In these experiments, we fix $\lambda_1 = 1$ and vary λ_2, while ensuring that $\lambda_1 > \lambda_2$. We present our results for functional threshold and one-sided threshold policies in Figures 2.3 and 2.4, respectively. We see that as λ_2 increases to λ_1, the average queue lengths and average waiting times for class-1 users decrease monotonically for both control policies. Similarly, we observe that the average queue lengths for class-2 users increase as λ_2 increases. Surprisingly, the average waiting times of class-2 users do not exhibit the



(a) L_1 vs. λ_2 (b) L_2 vs. λ_2

(c) W_1 vs. λ_2 (d) W_2 vs. λ_2

Figure 2.3: Average queue lengths and waiting times vs. λ_2 for various q and $\lambda_1 = 1$ under the functional threshold policy with $h(x) = x + 5$



(a) L_1 vs. λ_2 (b) L_2 vs. λ_2

(c) W_1 vs. λ_2 (d) W_2 vs. λ_2

Figure 2.4: Average queue lengths and waiting times vs. λ_2 for various q and $\lambda_1 = 1$ under the one-sided threshold policy with $N_1 = 2$

same monotonic behaviour. Under both policies, the average waiting times of class-2 users first decrease as λ_2 increases and then start increasing and this non-monotonic behaviour is observed especially when the matching probability is low. To understand the reasons behind this unexpected behaviour, we should remember two basic properties of the probabilistic matching systems: (i) if a user does not match with users in the system upon arrival, she should wait for new users to arrive at the system, and (ii) if too many users from the same class accumulate in the system, it becomes less likely for those users to be picked by a new arrival. When q and λ_2 are very small, an increase in λ_2 causes class-1 users to leave faster and new class-1 users can be admitted, which in return decreases the waiting time of class-2 users due to (i). However, when λ_2 is above a critical value, the negative effects of (ii) dominate the benefits of refreshing the class-1 queue, and the average waiting times increase as λ_2 increases.

2.7 Conclusion

In this chapter we present a Markov chain model to conduct an exact analysis of probabilistic matching systems. We derive an explicit formula for the probability distribution of the matching process to characterize the transient behaviour of the probabilistic matching systems. We show that if no control mechanism is applied, a probabilistic matching system is not ergodic for any set of parameters. We suggest admission control policies to ensure stability and analysed some performance measures. The simple threshold policy and one-sided threshold policy employ constant threshold values to admit users in the system. The simple threshold policy stabilizes the system for $q = 1$, but fails to stabilize the system when $q < 1$. The one-sided threshold policy stabilizes the system when one of the classes has a higher arrival rate. The ASQ and functional threshold policies rely on balancing the number of users in the system and stabilizes matching systems for any set of parameters. We prove that under a subset of functional threshold policies, the long run proportion of rejected users and the difference between average queue lengths is insensitive to the matching probability. Even more surprisingly, we show that the long run proportion of rejected users is an increasing function of the matching probability for a wide subset of functional threshold policies and we have the following conjecture:

Conjecture 2.21. Suppose the functional threshold policy is employed with a threshold function $h(x)$ which satisfies $h(x+1) \geq h(x)+1$ (or equivalently, $\frac{h(x+1)-h(x)}{x+1-x} \geq 1$). Then, the long run proportion of rejected users is a non-decreasing function of the matching probability q.

The analysis on probabilistic matching systems provide us with many interesting results. One research direction for further study is to analyse how a probabilistic matching system performs under user abandonments. The computational complexity and mathematically intractability of the matching process as shown in Section 2.3 suggest that an exact analysis of these systems for further characterization of the system is very difficult. We suggest to study heavy traffic limits to gain further insight in Chapter 3.

Chapter 3

Fluid and Diffusion Approximations

3.1 Introduction

In Chapter 2, Markov chain models are proposed for an exact analysis on probabilistic matching systems. As shown in Section 2.3, the probabilistic matching behaviour complicates the analysis of these systems and renders a complete exact analysis intractable. This motivates us to study diffusion approximation methods, in which the mathematically intractable queue length processes are replaced by some appropriate diffusion processes, to further characterize the system performance. In this chapter we propose approximation methods based on fluid and diffusion limits under two different scalings. A fluid limit is a deterministic process that approximates the stochastic queue length process. It applies the idea of functional law of large numbers and describes the tendency how the queue grows. On the other hand, the diffusion limit captures the fluctuations of the queue length process around its fluid limit and applies the idea of functional central limit theorem. Under the first scaling, we scale time and space while keeping the matching probability constant to obtain the limiting processes. We show that under this scaling both fluid and diffusion limits do not depend on the matching probability, which implies that the users from at most one class accumulate in the system and the probability of a user finding a match upon arrival approaches either zero or one.

To provide tools to address the matching probability explicitly, we propose a second scaling that also handles the abandonment of impatient users and scales the matching probability and the abandonment rate along with time and space. The resulting fluid and diffusion limits under this scaling involve differential equations which are not analytically tractable in the general case, although we can derive an analytical formula for the fluid limit when there are no abandonments. In Chapter 2 we show that some performance measures including the difference between the average queue lengths of different classes are insensitive to the matching probability under certain control policies. In this chapter, we show that despite not imposing any control policy, the difference between queue lengths for different classes is still insensitive to the matching probability in the fluid limit.

In addition, we analyze the asymptotic behaviour of the fluid limits. We first compare the fluid limits under both scalings, i.e., limits with and without scaling the matching probability, and show that when the abandonment rate is zero, the fluid limits in both scaling regimes agree with each other as time goes to infinity. Further, we show that for non-zero abandonment rates, the fluid limits converge to a unique fixed point, which is representative of the long run average number of users in the system. We prove that as the abandonment rate increases, the fixed point component for the class with lower arrival rate first experiences an increase and then decrease, while for the class with higher arrival rate it decreases monotonically. Finally we present numerical results of the fluid and diffusion limits in the second scaling regime.

There exists an extensive literature on fluid and diffusion approximations for Markovian systems with abandonments. Ward and Glynn [39] suggest diffusion approximations for the $M/M/1$ queue with exponential abandonments. They generalize these results to arrival, service and abandonment times with general distributions in [40]. Garnett et al. [23] consider $M/M/N$ queue with exponential abandonments and suggest diffusion approximations under Halfin-Whitt regime (see Halfin and Whitt [26]). Generalizing these results, Dai and He [18] and Mandelbaum and Momčilović [35] suggest diffusion approximations for many-server queues with general arrival, service and abandonment times. A recent work by Liu et al. [33] suggests diffusion approximations for the double sided queue where arrivals are renewal processes and customers abandon the system if they cannot find a match after an exponential time. This paper is closest to our work in nature and even though we restrict ourselves to Poisson arrival processes, our work extends [33] by assuming probabilistic matching structure.

3.2 Mathematical Model

The model we study in this chapter is similar to that introduced in Chapter 2, except that we incorporate the factor of user abandonments by assuming that each user abandons the system without being matched after waiting an exponentially distributed time with rate $\gamma \geq 0$.

Recall that in Chapter 2, we define $X_i^q(t)$ to be the number of class-i users in the system at time t when the matching probability is q and $M^q(t)$ to be the number of matched pairs up to time t. For notational simplification, we drop the notation of q for the rest of the work and denote $\{X_i(t)\}_{t\geq 0}$ and $\{M(t)\}_{t\geq 0}$ to be the queue length process of user-i and the matching process respectively.

With the introduction of the phenomenon of user abandonments, the continuous time Markov chain (CTMC) on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ has the generator matrix

$$Q_{(n_1,n_2)(n_1',n_2')} = \begin{cases} \lambda_1(1-q)^{n_2} & \text{if } n_1' = n_1 + 1 \text{ and } n_2' = n_2, \\ \lambda_2(1-q)^{n_1} & \text{if } n_1' = n_1 \text{ and } n_2' = n_2 + 1, \\ \lambda_1(1-(1-q)^{n_2}) + \gamma n_2 & \text{if } n_1' = n_1 \text{ and } n_2' = n_2 - 1 \geq 0, \\ \lambda_2(1-(1-q)^{n_1}) + \gamma n_1 & \text{if } n_1' = n_1 - 1 \geq 0 \text{ and } n_2' = n_2, \\ -(\lambda_1 + \lambda_2 + \gamma(n_1+n_2)) & \text{if } n_1' = n_1 \text{ and } n_2' = n_2, \\ 0 & \text{otherwise.} \end{cases}$$

The above model reduces to the one introduced in Chapter 2 when $\gamma = 0$.

It is sometimes useful in our analysis to express the queue length processes, $X_i(t)$, as the difference of counting processes. As $A_i(t)$ is defined to be the number of arrivals from class-i in Chapter 2, we further define $R_i(t)$ to be the number of user abandonments from class-i up to time t. Then we have the basic relation

$$X_i(t) = A_i(t) - M(t) - R_i(t) \text{ for all } t \geq 0 \text{ and } i = 1, 2.$$

As we have seen in Chapter 2, the essential element distinguishing a probabilistic matching system from a conventional queuing system is the matching probability q. In particular, for a system without abandonment, i.e., $\gamma = 0$, when the matching probability $q = 1$, class-1 and class-2 users cannot co-exist in the system at any time. Therefore the number of matched pairs up to time t is equal to the minimum of class-1 and class-2 arrivals, that is,

$$X_i(t) = A_i(t) - M(t) = A_i(t) - \min\{A_1(t), A_2(t)\}, \text{ for all } t \geq 0 \text{ and } i = 1, 2.$$

However, when $0 < q < 1$, analyzing the matching process $M(t)$ is far more difficult. The one dimensional distribution of the matching process, $\mathbb{P}(M(t) = k)$ for a given $t \geq 0$ and $k \in \mathbb{N}$ provided in Theorem 2.1 in Chapter 2 presents its complicated nature which further indicates the difficulty in fully characterizing the law of the matching process. Hence, in this chapter we propose fluid and diffusion approximations for probabilistic matching systems.

3.3 Fluid and Diffusion Approximations with Constant Matching Probabilities

In this section we focus on fluid and diffusion approximations for probabilistic matching systems obtained by only scaling time (or equivalently the arrival rates) and space while keeping the matching probability constant. This approach is especially useful in approximating systems where the probability that a given pair of users matches is high. For scalings with a constant matching probability, we assume that the users do not abandon the system without being matched, i.e., $\gamma = 0$.

3.3.1 Fluid Limits

We start by defining the scaled process $\{(\bar{X}_1^n(t), \bar{X}_2^n(t)), t \geq 0\}$ as

$$\bar{X}_i^n(t) = \frac{X_i(nt)}{n}, \; i = 1, 2.$$

We derive the limiting process of $\{\bar{X}_i^n(t), t \geq 0\}$ as $n \to \infty$. Note that in the rest of this chapter, we use the notation $X(\omega, t)$ when we need to specify the sample path of the stochastic process $X(t)$ corresponding to a scenario $\omega \in \Omega$. In this chapter, when we say a sequence of stochastic processes $\{X^n(t)\}$ convergences almost surely (a.s.) to a stochastic process $\{X(t)\}$ uniformly on compact sets (u.o.c.), it means that

$$\mathbb{P}(\omega, \bar{X}^n(\omega, t) \to \bar{X}(\omega, t), u.o.c) = 1$$

Below we give the definition for Converge Uniformly on Compact Sets which is introduced in Chen and Yao [17].

Definition 3.1. (Converge Uniformly on Compact Sets (See Chen and Yao [17])) For any $\omega \in \Omega$, we say that $\bar{X}^n(\omega, t)$ converges uniformly on compact sets (u.o.c.) to $\bar{X}(\omega, t)$, if $\sup_{0 \leq t \leq T} |\bar{X}^n(\omega, t) - \bar{X}(\omega, t)|$ converges to 0 for all $T > 0$ as $n \to \infty$.

A direct application of the functional strong law of large numbers (see e.g. [8], [17] and [41]) to Poisson arrival processes yields

$$\bar{A}_i^n(t) := \frac{A_i(nt)}{n} \xrightarrow{\text{a.s.}} \lambda_i t \text{ u.o.c. as } n \to \infty, i = 1, 2, \tag{3.1}$$

where a.s. indicates that the convergence is almost surely.

As users from a class accumulate in the system, the users from the other class are more likely to match upon their arrival. This implies that class-1 and class-2 users are unlikely to accumulate in the system at the same time. Lemma 3.2 formalizes this argument.

Lemma 3.2. *For any fixed $k > 0$, $\min\{\frac{X_1(nt)}{n^k}, \frac{X_2(nt)}{n^k}\} \xrightarrow{a.s.} 0$ u.o.c. as $n \to \infty$.*

Proof. If $q = 1$, since class-1 and class-2 do not co-exist in the system, for any $t \geq 0$, $\min\{X_1^n(t), X_2^n(t)\} = 0$, and hence the desired conclusion follows trivially. If $0 < q < 1$, to simplify the notation, define $I^{n,k}(t) := \min(\frac{X_1(nt)}{n^k}, \frac{X_2(nt)}{n^k})$, choose an $a \in (0, k)$ and let $\lambda = \lambda_1 + \lambda_2$. Then for $m \leq n^2 - 1 \in \mathbb{N}$, we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{0 \leq t \leq \frac{m+1}{n^2}} I^{n,k}(t) \geq n^{-a} | \sup_{0 \leq t \leq \frac{m}{n^2}} I^{n,k}(t) < n^{-a}\right) \\
&= \mathbb{P}\left(\sup_{\frac{m}{n^2} \leq t \leq \frac{m+1}{n^2}} I^{n,k}(t) \geq n^{-a} | \sup_{0 \leq t \leq \frac{m}{n^2}} I^{n,k}(t) < n^{-a}\right) \\
&= \mathbb{P}\left(\sup_{\frac{m}{n^2} \leq t \leq \frac{m+1}{n^2}} \min(X_1(nt), X_2(nt)) \geq n^{k-a} | \sup_{0 \leq t \leq \frac{m}{n^2}} \min(X_1(nt), X_2(nt)) < n^{k-a}\right) \\
&\leq \sum_{j=0}^{\infty} \frac{e^{-\frac{\lambda}{n^2}} (\frac{\lambda}{n^2})^j}{j!} j r^{n^{k-a}} \qquad (3.2) \\
&= \frac{\lambda}{n^2} r^{n^{k-a}}.
\end{aligned}$$

We see that the inequality (3.2) holds using the following argument. For both $X_1(nt)$ and $X_2(nt)$ to reach a level above n^{k-a} at some point during $[\frac{m}{n^2}, \frac{m+1}{n^2}]$, at least one of the arrivals occurring during $[\frac{m}{n^2}, \frac{m+1}{n^2}]$ should fail to match and stay in the system upon arrival when facing at least $\lfloor n^{k-a} \rfloor$ users from the other user queue (where $\lfloor x \rfloor$ is the smallest interger no smaller then x). If we observe j arrivals during this time frame, the probability of this event is bounded by $jr^{n^{k-a}}$. Then, for any fixed $T > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{0\le t\le T} I^{n,k}(t) \ge n^{-a}\right)\\
&= \mathbb{P}\left(\sup_{0\le t\le T} I^{n,k}(t) \ge n^{-a} | \sup_{0\le t\le T-\frac{1}{n^2}} I^{n,k}(t) < n^{-a}\right) \mathbb{P}\left(\sup_{0\le t\le T-\frac{1}{n^2}} I^{n,k}(t) < n^{-a}\right)\\
&\quad + \mathbb{P}\left(\sup_{0\le t\le T} I^{n,k}(t) \ge n^{-a} | \sup_{0\le t\le T-\frac{1}{n^2}} I^{n,k}(t) \ge n^{-a}\right) \mathbb{P}\left(\sup_{0\le t\le T-\frac{1}{n^2}} I^{n,k}(t) \ge n^{-a}\right)\\
&\le \mathbb{P}\left(\sup_{0\le t\le T} I^{n,k}(t) \ge n^{-a} | \sup_{0\le t\le T-\frac{1}{n^2}} I^{n,k}(t) < n^{-a}\right)\\
&\quad + \mathbb{P}\left(\sup_{0\le t\le T-\frac{1}{n^2}} I^{n,k}(t) \ge n^{-a}\right)\\
&\le \sum_{m=0}^{Tn^2} \mathbb{P}\left(\sup_{0\le t\le \frac{m+1}{n^2}} I^{n,k}(t) \ge n^{-a} | \sup_{0\le t\le \frac{m}{n^2}} I^{n,k}(t) < n^{-a}\right)\\
&\le \sum_{m=0}^{Tn^2} \frac{\lambda}{n^2} r^{n^{k-a}} = T\lambda r^{n^{k-a}}
\end{aligned}$$

For any $\epsilon > 0$, there exists an N_ϵ, such that for $n \ge N_\epsilon$ such that

$$\mathbb{P}\left(\sup_{0\le t\le T} I^{n,k}(t) \ge \epsilon\right) \le \epsilon,$$

which implies $I^{n,k}(t) \xrightarrow{\mathbb{P}} 0$ u.o.c. Furthermore,

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\sup_{0\le t\le T} I^{n,k}(t) \ge n^{-a}\right) \le T\lambda \sum_{n=0}^{\infty} r^{n^{k-a}} < \infty.$$

For any $\epsilon > 0$ choosing $N \ge 1$, such that for $N^{-a} < \epsilon$, we obtain

$$\begin{aligned}
&\sum_{n=1}^{\infty} \mathbb{P}\left(\sup_{0\le t\le T} I^{n,k}(t) > \epsilon\right)\\
&= \sum_{n=1}^{N-1} \mathbb{P}\left(\sup_{0\le t\le T} I^{n,k}(t) > \epsilon\right) + \sum_{n=N}^{\infty} \mathbb{P}\left(\sup_{0\le t\le T} I^{n,k}(t) > \epsilon\right)\\
&\le \sum_{n=1}^{N-1} \mathbb{P}\left(\sup_{0\le t\le T} I^{n,k}(t) > \epsilon\right) + \sum_{n=N}^{\infty} \mathbb{P}\left(\sup_{0\le t\le T} I^{n,k}(t) \ge n^{-a}\right) < \infty
\end{aligned}$$

Using Borel-Cantelli lemma we get

$$\mathbb{P}\left(\sup_{0\le t\le T} I^{n,k}(t) > \epsilon \text{ infinitely often}\right) = 0,$$

and $I^{n,k}(t) = \min\left(\frac{X_1(nt)}{n^k}, \frac{X_2(nt)}{n^k}\right) \xrightarrow{\text{a.s.}} 0$ u.o.c. □

Theorem 3.3. *$\bar{X}_i^n(t) \xrightarrow{a.s.} \bar{X}_i(t)$ u.o.c. as $n \to \infty$, where*

$$\bar{X}_i(t) = \lambda_i t - \min(\lambda_1, \lambda_2)t, i = 1, 2.$$

Proof. Equation (3.1) and Lemma 3.2 imply that there exists a $\Omega' \subset \Omega$ with $\mathbb{P}(\Omega') = 1$ where for every $\omega \in \Omega'$,

$$\frac{A_i(\omega, nt)}{n} \to \lambda_i t \text{ u.o.c.}$$
$$\min\left(\frac{X_1(\omega, nt)}{n}, \frac{X_2(\omega, nt)}{n}\right) \to 0 \text{ u.o.c.}$$

for all $t \ge 0$ and $i = 1, 2$. Our first goal is to show $\bar{M}^n(\omega, t) := \frac{M(nt)}{n} \to \min(\lambda_1 t, \lambda_2 t)$ u.o.c. as $n \to \infty$ for all $t \ge 0$ and $\omega \in \Omega'$. Suppose that there exists some $\omega' \in \Omega'$ which this statement does not hold and without loss of generality assume $\lambda_1 \ge \lambda_2$. Also, we know that the number of matchings is always bounded by the number of arrivals as $M(\omega, t) < \min(A_1(t), A_2(t))$ for all $t \ge 0$. These imply that there exists a $\delta > 0$, $N_\delta > 0$ sequences $n_j \to \infty$ as $j \to \infty$ and $0 \le t_j \le T$ such that $\lambda_2 t_j - \bar{M}^{n_j}(\omega, t_j) > \delta$ for all $j > N_\delta$. Boundedness of t_j and $M(\omega, 0) = 0$ also implies that there exists a subsequence $t_{j_k} \to t' > 0$. For any $\epsilon > 0$, we can choose N_ϵ such that for every $k > N_\epsilon$ we have $\left|\frac{A_i(n_{j_k} t_{j_k})}{n_{j_k}} - \lambda_i t_{j_k}\right| < \frac{\epsilon}{2}$ for $i = 1, 2$ and $|t_{j_k} - t'| < \frac{\epsilon}{2(\lambda_1 - \lambda_2)}$, which in turn implies

$$\begin{aligned}
\frac{A_1(n_{j_k} t_{j_k})}{n_{j_k}} - \frac{M(n_{j_k} t_{j_k})}{n_{j_k}} &= \frac{A_1(n_{j_k} t_{j_k})}{n_{j_k}} - \frac{M(n_{j_k} t_{j_k})}{n_{j_k}} - (\lambda_1 - \lambda_2)(t_{j_k} - t') \\
&\quad + (\lambda_1 - \lambda_2)(t_{j_k} - t') \\
&> \lambda_2 t_{j_k} - \frac{M(n_{j_k} t_{j_k})}{n_{j_k}} + (\lambda_1 - \lambda_2)t' - \epsilon \\
&> \delta - \epsilon + (\lambda_1 - \lambda_2)t'.
\end{aligned}$$

Similarly, we also get

$$\frac{A_2(n_{j_k}t_{j_k})}{n_{j_k}} - \frac{M(n_{j_k}t_{j_k})}{n_{j_k}} > \delta - \epsilon.$$

Letting $\epsilon \to 0$, we get

$$\min\{\frac{X_1(\omega, n_{j_k}t_{j_k})}{n_{j_k}}, \frac{X_2(\omega, n_{j_k}t_{j_k})}{n_{j_k}}\} > \delta,$$

which contradicts with Lemma 3.2 and proves $\bar{M}^n(t) \xrightarrow{\text{a.s.}} \min\{\lambda_1 t, \lambda_2 t\}$ u.o.c. as n goes to ∞ for all $t \geq 0$. Then the result follows using continuous mapping theorem ([17], Theorem 5.2). □

3.3.2 Diffusion Limits

Fluid limits provide useful approximations to determine how queue lengths grow, however they fail to represent the stochastic fluctuations. To understand the fluctuations of sample paths around the fluid limit, we now focus on diffusion approximations. A direct application of functional central limit theorem (see e.g. Theorem 5.7 in [17]) on Poisson arrival streams we get

$$\hat{A}_i^n(t) := \frac{A_i(nt) - n\bar{A}_i(t)}{\sqrt{n}} \Rightarrow \hat{A}_i(t), i = 1, 2, \tag{3.3}$$

where $\hat{A}_i = \sqrt{\lambda_i}B_i$, $B_i(t), i = 1, 2$, is independent one-dimensional standard Brownian motions and "$\Rightarrow$"denotes weak convergence with respect to the Skorokhod topology. We define the process

$$\hat{X}_i^n(t) = \frac{X_i(nt) - \bar{X}_i(nt)}{\sqrt{n}}.$$

Now we state the result of diffusion limits for probabilistic matching systems when the matching probability is kept constant, which shows that the diffusion limits are independent of the matching probability q.

Theorem 3.4. *As $n \to \infty, \hat{X}_i^n \Rightarrow \hat{X}_i, i = 1, 2,$ where $\hat{X}_i$ is defined as:*

1. *If $\lambda_1 = \lambda_2$, $\hat{X}_i = \hat{A}_i - \min(\hat{A}_1, \hat{A}_2), i = 1, 2.$*
2. *If $\lambda_1 > \lambda_2, \hat{X}_1 = \hat{A}_1 - \hat{A}_2, \hat{X}_2 = 0.$*

Proof. We first consider the case when $\lambda_1 = \lambda_2 = \lambda$. Define $\hat{M}^n(t) := \frac{M(nt)-\lambda nt}{\sqrt{n}}$. Using Skorokhod representation theorem (Theorem 5.1 in [17]) there exists versions of $A_i(t), \hat{A}_i(t)$ and $B_i(t), i = 1, 2$, which we denote $A'_i(t), \hat{A}'_i(t)$ and $B'_i(t), i = 1, 2$, and matching and scaled processes $M(t)$ and $\hat{A}_i^{n'}(t), i = 1, 2$ associated with these versions such that $\hat{A}_i^{n'}(t) \xrightarrow{\text{a.s.}} \hat{A}'_i(t) = \sqrt{\lambda}B'_i(t), i = 1, 2$. Lemma 3.2 implies $\min(\hat{A}_1^{n'}(t) - \hat{M}^{n'}(t), \hat{A}_2^{n'}(t) - \hat{M}^{n'}(t)) \xrightarrow{\text{a.s.}} 0$ u.o.c. Proceeding in the same manner as in the proof of Theorem 3.3, we get $\hat{M}^{n'} \xrightarrow{\text{a.s.}} \min(\hat{A}'_1, \hat{A}'_2)$. Applying the continuous mapping theorem (Theorem 5.2 in [17]) the result follows for $\lambda_1 = \lambda_2 = \lambda$.

When $\lambda_1 > \lambda_2$, let $\tau_m = \inf\{t \geq 0 : A_2(t) \geq m\}$ and define a sequence of random variables $\{\xi_m\}_{m\geq 1}$ such that

$$\xi_m = \begin{cases} 1, & \text{the } m\text{-th arriving user-2 finds a match successfully upon her arrival,} \\ 0, & \text{otherwise.} \end{cases}$$

We have $\tau_m \to \infty$, as $m \to \infty$, and for any $m \geq 1$, $\sum_{m=1}^{A_2(t)} \xi_m \leq M(t)$. Generate a sequence of a uniform random variables $\{U_m\}_{m\geq 1}$ such that $U_m \sim U(0, 1)$, then assuming $0^0 = 1$, we have

$$\begin{aligned} \mathbb{P}(\xi_m = 0) &= \mathbb{P}(U_m < (1-q)^{X_1(\tau_m)}) \\ &\leq \mathbb{P}(U_m < (1-q)^{A_1(\tau_m)-m}) \\ &= \mathbb{E}[\mathbb{P}(U_m < (1-q)^{A_1(\tau_m)-m} | A_1(\tau_m))] \\ &= \mathbb{E}[((1-q)^{A_1(\tau_m)-m}) \wedge 1]. \end{aligned}$$

Next we show that there exists an $M > 0$ and $c > 0$ such that for any $m \geq M$,

$$\mathbb{E}[(1-q)^{A_1(\tau_m)-m}] < (1-q)^{cm}.$$

For any c_1 such that $1 < c_1 < \frac{\lambda_1}{\lambda_2}$ we have, as $t \to \infty$, $\frac{A_1(t)}{t} - c_1\frac{A_2(t)}{t} \xrightarrow{\text{a.s.}} \lambda_1 - c_1\lambda_2$, i.e., there exists a $T > 0$, such that for any $t > T$, $A_1(t) - c_1A_2(t) > \frac{(\lambda_1 - c_1\lambda_2)t}{2}$ a.s. Since

$\tau_m \to \infty$, there exists an $M > 0$ such that for any $m \geq M$, we have $\tau_m > T$ and

$$A_1(\tau_m) - c_1 A_2(\tau_m) = A_1(\tau_m) - c_1 m > \frac{(\lambda_1 - c_1\lambda_2)}{2}\tau_m > 0 \text{ a.s.}$$

Choosing $c = c_1 - 1$ we have $\mathbb{E}[(1-q)^{A_1(\tau_m)-m}] < (1-q)^{cm}$ and

$$\sum_{m=0}^{\infty} \mathbb{P}(\xi_m = 0) = \sum_{m=0}^{\infty} \mathbb{P}(U_m < r^{X_1(T_2(m))}) < \sum_{m=0}^{\infty} r^{cm} < \infty.$$

Using Borel-Cantelli Lemma, $\mathbb{P}(\xi_m = 0 \text{ infinitely often}) = 0$ which in turn implies

$$\hat{X}_2^n(t) = \frac{A_2(nt) - M(nt)}{\sqrt{n}} \xrightarrow{\text{a.s.}} 0.$$

Finally, we have

$$\begin{aligned}\hat{X}_1^n(t) &= \frac{A_1(nt) - M(nt)}{\sqrt{n}} - \frac{(\lambda_1 - \lambda_2)nt}{\sqrt{n}} \\ &= \frac{A_1(nt) - \lambda_1 nt}{\sqrt{n}} - \frac{A_2(nt) - \lambda_2 nt}{\sqrt{n}} - \frac{A_2(nt) - M(nt)}{\sqrt{n}}.\end{aligned}$$

Hence, the result follows from the continuous mapping theorem. □

We conclude that when the matching probability q is kept as a constant in the diffusion approximation, it is absent in both the fluid limits and the diffusion limits. Moreover, we can compare our results with those of an $M/M/1$ queue. When the arrival rates in probabilistic matching systems are not equal, the fluid and diffusion limits of queue length process i behaves in accordance with that of an M/M/1 queue with arrival rate λ_i and service rate λ_j (see Chen and Yao (2001) [17] for more details). When the arrival rates are identical, the diffusion limits are distinct from those of an M/M/1 queue, due to the fact that in a probabilistic matching system, the next arriving user i is possible to be matched immediately upon arrival which indicates that the accumulation of user j when no user i is at present would not be a “waste” unlike the service time generated in an empty M/M/1 queue. As a result, rather than having the one-sided reflection mapping of the net-input process, we only have the positive sign of the difference between the arrival processes. We suggest that this diffusion approximation would fit the system which has a relatively high matching probability of each pair of users and thus the probability of an arriving user getting matched increases significantly as the number of users from the other queue grows. However, the underlying assumption above does not hold in those systems which have a very small matching probability for each pair of users, because if q is very close to 0, a user is not so likely to find a match upon arrival even when there are many users in the other queue.

3.4 Fluid and Diffusion Limits for Systems with Small Matching Probabilities

The matching probability disappears in the fluid and diffusion limits presented in Section 3.3 and this indicates that the systems with matching probability $0 < q < 1$ behave very similar to the systems with matching probability 1. However, in many real world problems the matching probability q is very small and we need tools that explicitly addresses the probabilistic nature of the matchings. In this section, we suggest a second type of diffusion approximation which scales q together with the space and time to get a better description of the dynamics of those systems with small matching probabilities.

We often observe that the users are impatient and may leave the system without being matched if they cannot match after waiting for sometime. We include this factor in the discussion of the queue length process in the new asymptotic regime, adopting a similar approach to that of Ward and Glynn [39], in which the diffusion limit of an M/M/1+M queue with small abandonment rate is provided. We assume that each user has an exponentially distributed abandonment time with rate γ, $0 \leq \gamma < \infty$, independent of others, where $\gamma \ll \lambda_i, i = 1, 2$. Hence, as we scale space, time and the matching probability, we also let abandonment rate approach to zero.

3.4.1 Fluid Limits

Let $X_i^n(t)$ to be number of class-i users in a probabilistic matching system where class-i users arrive according to a Poisson process with λ_i, users abandon the system if they do not match after waiting an exponential time $\gamma^{(n)} = \frac{\gamma}{n}$, $(0 \leq \gamma < \infty)$, the matching probability is $q^{(n)} = \frac{q}{n}$, $0 < q < 1$. Then, we define

$$\bar{X}^{s,n}(t) := \frac{X_i^n(nt)}{n}$$

to be the scaled system in this regime. Note that we use the small notation s in the scaled process to indicate that this is the situation of systems with *small* matching probability. Now our goal is to show that as $n \to \infty$, the scaled system approaches to the fluid limit $\bar{X}^s(t)$, which is the unique solution to the following ordinary differential equations (ODE):

$$\bar{X}_1^s(0) = \bar{X}_2^s(0) = 0, \tag{3.4}$$

$$\frac{d\bar{X}_1^s(t)}{dt} = \lambda_1 e^{-q\bar{X}_2^s(t)} - \lambda_2(1 - e^{-q\bar{X}_1^s(t)}) - \gamma\bar{X}_1^s(t), \tag{3.5}$$

$$\frac{d\bar{X}_2^s(t)}{dt} = \lambda_2 e^{-q\bar{X}_1^s(t)} - \lambda_1(1 - e^{-q\bar{X}_2^s(t)}) - \gamma\bar{X}_2^s(t). \tag{3.6}$$

Define

$$F(x) = \begin{pmatrix} \lambda_1 e^{-qx_2} - \lambda_2(1 - e^{-qx_1}) - \gamma x_1 \\ \lambda_2 e^{-qx_1} - \lambda_1(1 - e^{-qx_2}) - \gamma x_1 \end{pmatrix}. \tag{3.7}$$

Then the equations (3.5)-(3.6) are in the form $\frac{dx}{dt} = F(x) = (F_1(x), F_2(x))^\intercal$, where $F(\cdot)$ is Lipschitz and hence the initial value problem admits a unique solution. We first show that the solution $\bar{X}^s(t)$ is bounded when $\gamma > 0$.

Lemma 3.5. *Let* $\bar{X}^s(t) = (\bar{X}_1^s(t), \bar{X}_2^s(t))$ *be the unique solution to* (3.4)-(3.6) *and* $\gamma > 0$*, then*

$$\sup_{0 \le t < \infty} \bar{X}_i^s(t) < \lambda_i/\gamma, \quad i = 1, 2.$$

Proof. For any (x_1, x_2) such that $x_1 \ge \lambda_1/\gamma$, we have

$$F_1(x_1, x_2) = \lambda_1 e^{-qx_2} - \lambda_2(1 - e^{-qx_1}) - \gamma x_1 < \lambda_1 - \gamma x_1 \le 0.$$

Using (3.4) this implies that $\bar{X}_1^s(t) \le \lambda_1/\gamma$, for all t. Similar argument also holds for $\bar{X}_2^s(t)$. □

When the matching probability is scaled in a way that $q^n \to 0$, the techniques we use to derive fluid and diffusion limits differ from the ones used in Section 3.3. In particular, we appeal to the Laplace transform methods where a limiting kernel with the corresponding Laplace transform is identified (see e.g. [19] for a brief review of these methods). For this purpose, we need the Lévy kernel for the Markov process $\bar{X}^{s,n}(t)$ defined as follows:

$$\begin{aligned}K^n(x,dy) := &\lambda_1 n(1-\frac{q}{n})^{nx_2}\delta\left((y-x)-(\frac{1}{n},0)\right)dy\\ &+\lambda_2 n(1-\frac{q}{n})^{nx_1}\delta\left((y-x)-(0,\frac{1}{n})\right)dy\\ &+(\lambda_1 n(1-(1-\frac{q}{n})^{nx_2})+\gamma n x_2)\delta\left((y-x)+(0,\frac{1}{n})\right)dy\\ &+(\lambda_2(1-(1-\frac{q}{n})^{nx_1})+\gamma n x_1)\delta\left((y-x)+(\frac{1,}{n},0)\right)dy,\end{aligned}$$

where $\delta(y)$ is the Dirac delta function. Then, we can define the Laplace transform of operator $K^n(x,dy)$ as

$$\begin{aligned}m^n(x,\theta) &= \int_{(0,\infty)\times(0,\infty)} e^{\langle\theta,y\rangle}K^n(x,dy)\\ &= \lambda_1 n(1-\frac{q}{n})^{nx_2}e^{\frac{\theta_1}{n}}+\lambda_2 n(1-\frac{q}{n})^{nx_1}e^{\frac{\theta_2}{n}}\\ &\quad+(\lambda_1 n(1-(1-\frac{q}{n})^{nx_2})+\gamma n x_2)e^{-\frac{\theta_2}{n}}\\ &\quad+(\lambda_2 n(1-(1-\frac{q}{n})^{nx_1}+\gamma n x_1)e^{-\frac{\theta_1}{n}}. \end{aligned} \tag{3.8}$$

Now, we are ready to state our result for convergence to the fluid limit.

Theorem 3.6. *For any $\delta > 0$ and $T > 0$,*

$$\limsup_{n\to\infty} n^{-1}\log\mathbb{P}\left(\sup_{0\le t\le T}|\bar{X}_i^{s,n}(t)-\bar{X}_i^s(t)|>\delta\right)<0 \tag{3.9}$$

and as $n\to\infty$,

$$\bar{X}_i^{s,n}(t) \xrightarrow{a.s.} \bar{X}_i^s(t) \ u.o.c.,$$

where $\bar{X}_i^s(t), i=1,2$ is the unique solution to the system of ODE given by (3.4)-(3.6).

Proof. If $\gamma = 0$, set $\mathbb{S} = \mathbb{R}_{\ge 0}\times\mathbb{R}_{\ge 0}$ and $T^n = T$, otherwise choose $C_i > \lambda_i/\gamma$ for $i = 1,2$, and set $\mathbb{S} = [0,C_1]\times[0,C_2]$ and $T^n = \inf\{t\ge 0 : \bar{X}^{s,n}(t)\notin\mathbb{S}\}\wedge T$. Then, Proposition 5.1 in [19] implies

$$\limsup_{n\to\infty} n^{-1}\log \mathbb{P}\left(\sup_{0\le t\le T^n}|\bar{X}_i^{s,n}(t)-\bar{X}_i^{s}(t)|>\delta\right)<0 \tag{3.10}$$

if we can show that the following three conditions hold:

(i) There exists a $\eta_0 > 0$ such that

$$\sup_n \sup_{x\in\mathbb{S}} \sup_{|\theta|\le\eta_0} \frac{m^n(x,n\theta)}{n} < \infty$$

.

(ii) $\sup_{x\in\mathbb{S}}\left|\frac{\partial m^n(x,\theta)}{\partial\theta}|_{\theta=0} - F(x)\right| \to 0.$

(iii) $\limsup_{n\to\infty} n^{-1}\log\mathbb{P}(|\bar{X}_i^{s,n}(0)-\bar{X}_i^{s}(0)|>\delta)<0.$

The third condition is trivially satisfied as we assume that a probabilistic matching system is initially empty and we have $\bar{X}^{s,n}(0) = 0$ for all n. When $\gamma > 0$ the first condition follows as when $x\in\mathbb{S}$ for any $\eta_0 > 0$ and $\theta\le\eta_0$ we have

$$\begin{aligned}\frac{m^n(x,n\theta)}{n} &= (\lambda_1(1-(1-\frac{q}{n})^{nx_2})+\gamma x_2)e^{-\theta_2} + (\lambda_2(1-(1-\frac{q}{n})^{nx_1})+\gamma x_1)e^{-\theta_1}\\ &+\lambda_1(1-\frac{q}{n})^{nx_2}e^{\theta_1}+\lambda_2(1-\frac{q}{n})^{nx_1}e^{\theta_2}\\ &\le(\lambda_1+\lambda_2)e^{\eta_0}+\lambda_1+\lambda_2+\gamma(C_1+C_2).\end{aligned}$$

Similarly, when $\gamma = 0$, the supremum can be bounded by $(\lambda_1+\lambda_2)e^{\eta_0}+\lambda_1+\lambda_2$. To prove the second condition we write

$$\begin{aligned}\left.\frac{\partial m^n(x,\theta)}{\partial\theta_1}\right|_{\theta_1=0} &= \lambda_1(1-\frac{q}{n})^{nx_2} - (\lambda_2(1-(1-\frac{q}{n})^{nx_1})+\gamma x_1),\\ \left.\frac{\partial m^n(x,\theta)}{\partial\theta_2}\right|_{\theta_2=0} &= \lambda_2(1-\frac{q}{n})^{nx_1} - (\lambda_1(1-(1-\frac{q}{n})^{nx_1})+\gamma x_2).\end{aligned}$$

Then it is easy to see pointwise convergence $\frac{\partial m^n(x,\theta)}{\partial\theta}|_{\theta=0}\to F(x)$ and when $\gamma > 0$ the uniform convergence follows from continuity of the functions and compactness of the underlying set. When $\gamma = 0$, to show the uniform convergence we use the definition of uniform convergence. In particular, we need to show that for any $\epsilon > 0$, there exists N

such that when $n > N$, we have for any $x \in \mathbb{R}_{\geq 0}$, $|(1-\frac{q}{n})^{nx} - e^{-qx}| < \epsilon$. First we show that for any $\epsilon > 0$, there exists N_1 and c such that when $n > N_1$ and $x > c$, we have $|(1-\frac{q}{n})^{nx} - e^{-qx}| < \epsilon$. We know that $\ln(1-\frac{q}{n})^n \to -q$ as $n \to \infty$. For any δ_1 such that $0 < \delta_1 < q$, we can find an N_1, s.t., for $n > N_1$, we have $x \ln(1-\frac{q}{n})^n < x(-q+\delta_1)$. As a result, letting $c_1 = \frac{\ln \frac{\epsilon}{2}}{-q+\delta_1}$, when $x > c_1$, we have $\ln(1-\frac{q}{n})^{nx} < x(-q+\delta_1) < \ln \frac{\epsilon}{2}$, or equivalently, $(1-\frac{q}{n})^{nx} < \frac{\epsilon}{2}$. Moreover, we know that as $x \to \infty$, $e^{-qx} \to 0$. We can find a c_2 such that when $x > c_2$, $e^{-qx} < \frac{\epsilon}{2}$. Letting $c = \max(c_1, c_2)$, the statement follows. Next, due to compactness and pointwise convergence, we know that for $x \in [0, c]$, there exists an N_2 such that for $n > N_2$, we have $|(1-\frac{q}{n})^{nx} - e^{-qx}| < \epsilon$. Therefore choosing $N = \max(N_1, N_2)$ we have the uniform convergence for any $x \in \mathbb{R}_{\geq 0}$. As a result, the uniform convergence result for our system when $\gamma = 0$ follows. Therefore (3.10) follows from Proposition 5.1 in [19]. When $\gamma = 0$, $T_n = T$ a.s., and when $\gamma > 0$ from (3.10), Lemma 3.5 and $C_i > \lambda_i/\gamma$ we conclude that $T^n \xrightarrow{\mathbb{P}} T$, which implies (3.9). The almost sure convergence is a simple application of Borel-Cantelli lemma. □

When there are abandonments ($\gamma > 0$), the right hand sides of (3.5) and (3.6) involve both e^{-qx} and x terms which makes it difficult to obtain an analytical solution. However, when the customers do not abandon the system, the ODE can be solved analytically. Corollary 3.7 presents this special case.

Corollary 3.7. *When $\gamma = 0$, as $n \to \infty$,*

$$\bar{X}_i^{s,n}(t) \xrightarrow{a.s.} \frac{1}{q} \ln(e^{\lambda_1 qt} + e^{\lambda_2 qt} - 1) - \lambda_j t \ \ u.o.c., i, j \in \{1, 2\}, i \neq j. \tag{3.11}$$

Proof. Setting $\gamma = 0$ and taking the integral of (3.5) and (3.6), we see that

$$\bar{X}_1^s(t) + \lambda_2 t = \bar{X}_2^s(t) + \lambda_1 t =: y(t).$$

Then, we have

$$\frac{dy(t)}{dt} = e^{-qy(t)}(\lambda_1 e^{\lambda_1 qt} + \lambda_2 e^{\lambda_2 qt})$$

and $y(0) = 0$ which has the unique solution $y(t) = \frac{1}{q}\ln(e^{\lambda_1 qt} + e^{\lambda_2 qt} - 1)$ and the result follows. □

In [12], certain performance measures are proven to be independent of the matching probability q under some additional control policies. Specifically, Theorem 14 in [12] states that under some admission control policies where the difference between long run average queue lengths of class-1 and class-2 users does not depend on the matching probability q. The following corollary also indicates a similar property even under the presence of user abandonments.

Corollary 3.8. *When $\gamma > 0$, as $n \to \infty$,*

$$\bar{X}_1^{s,n}(t) - \bar{X}_2^{s,n}(t) \xrightarrow{a.s.} \frac{\lambda_2 - \lambda_1}{\gamma} e^{-\gamma t} + \frac{\lambda_1 - \lambda_2}{\gamma}, \ u.o.c.$$

Proof. Applying Theorem 3.6 and the continuous mapping theorem, $\bar{X}_1^{s,n}(t) - \bar{X}_2^{s,n}(t)$ converges to the unique solution of

$$\frac{dx(t)}{dt} = \lambda_1 - \lambda_2 - \gamma x(t) \tag{3.12}$$

with initial condition $x(0) = 0$. Using integrating factors, the solution of this first order ODE can be obtained as $\bar{X}_1^s(t) - \bar{X}_2^s(t) = \frac{\lambda_2 - \lambda_1}{\gamma} e^{-\gamma t} + \frac{\lambda_1 - \lambda_2}{\gamma}$. □

Corollary 3.8 implies that when $\gamma > 0$, the matching probability q does not affect the difference between the numbers of class-1 and class-2 users in the system. As $t \to \infty$, this difference converges to $\frac{\lambda_1 - \lambda_2}{\gamma}$, which coincides with the results of [39] for M/M/1+M queue with has arrival rate λ_1, service rate λ_2 and abandonment rate $\gamma > 0$.

Next, we analyze the asymptotic behaviour of the fluid limits as time goes to infinity. Corollary 3.7 assumes γ to be 0 and allows us to compare $\bar{X}^s(t)$ with fluid limits $\bar{X}(t)$, given in Theorem 3.3. Different from $\bar{X}(t)$ which does not carry any information on the matching probability q, the fluid limits in Corollary 3.7 depends on q. When t is small, $\bar{X}_i^s(t)$ grows for both $i = 1$ and 2 as q increases. However, as t becomes larger, the influence of the matching probability becomes weaker. Proposition 3.1 shows that the fluid limits $\bar{X}^s(t)$ converges to $\bar{X}(t)$ as $t \to \infty$.

Proposition 3.1. *Suppose $\gamma = 0$, then as $t \to \infty$, $|\bar{X}_i(t) - \bar{X}_i^s(t)| \to 0, t \geq 0, i = 1, 2$.*

Proof. Without lost of generality, we assume that $\lambda_1 \geq \lambda_2$. Then applying using

Corollary 3.7 and Theorem 3.3, we have

$$\begin{aligned}\bar{X}_1^s(t) - \bar{X}_1(t) &= \frac{1}{q}\ln(e^{\lambda_1 qt} + e^{\lambda_2 qt} - 1) - \lambda_1 t \\ &= \ln(e^{\lambda_1 qt} + e^{\lambda_2 qt} - 1)^{\frac{1}{q}} - \lambda_1 t \\ &= \ln \frac{\sqrt[q]{e^{\lambda_1 qt} + e^{\lambda_2 qt} - 1}}{\sqrt[q]{e^{\lambda_1 qt}}}\end{aligned}$$

Since $\lambda_1 > \lambda_2$, we can see that as $t \to \infty$, $\left|\frac{\sqrt[q]{e^{\lambda_1 qt}+e^{\lambda_2 qt}-1}}{\sqrt[q]{e^{\lambda_1 qt}}}\right| \to 1$ and this implies $|\bar{X}_1^s(t) - \bar{X}_1(t)| \to 0$. □

In other words, we can explain the dynamics of a probabilistic matching system in the following way: without considering the effect of user abandonments, if each pair of users gets harder to match with each other, we observe more users waiting in the system. However if we run the system long enough, the average of numbers of users in the system only depends on the arrival rates. Next we show that for general abandonment rate $\gamma \geq 0$, the fluid limits of the queue length processes converge to a fixed point as $t \to \infty$.

Proposition 3.2. *If $\gamma > 0$, the fluid limit $\bar{X}_i^s(t) \to x_i^*, i = 1, 2$ as $t \to \infty$, where $x_i^* \in \mathbb{R}$ satisfies the following set of equations*

$$\lambda_1 e^{-qx_2^*} - \lambda_2(1 - e^{-qx_1^*}) - \gamma x_1^* = 0, \tag{3.13}$$

$$\lambda_2 e^{-qx_1^*} - \lambda_1(1 - e^{-qx_2^*}) - \gamma x_2^* = 0. \tag{3.14}$$

Proof. First, we prove that Equations (3.13) and (3.14) have a unique solution. Subtracting the second equation from the first one $x_2^* = x_1^* + \frac{\lambda_2 - \lambda_1}{\gamma}$ and replacing this into (3.13) we get

$$\lambda_1 e^{-\frac{q(\lambda_2-\lambda_1)}{\gamma}} e^{-qx_1^*} - \lambda_2(1 - e^{-qx_1^*}) - \gamma x_1^* = 0.$$

The left hand side of the equation is decreasing in x_1^*, equals to $\lambda_1 e^{-\frac{q(\lambda_2-\lambda_1)}{\gamma}} > 0$ if $x_1^* = 0$ and goes to $-\infty$ as $x_1^* \to \infty$. Hence, using the intermediate value theorem we conclude that (3.13) and (3.14) have a unique solution and $x^* = (x_1^*, x_2^*)$ is the unique fixed point of the system of equations (3.4)-(3.6).

When $\lambda_1 \neq \lambda_2$, $\bar{X}^s(t)$ solving (3.4)-(3.6) converges to x^* as $t \to \infty$, if we can find a Lyapunov function $V(x)$ with the following properties (see e.g. Strogatz [37]):

1. $V(x) > 0$ for all $x \neq x^*$ and $V(x^*) = 0$.

2. $\frac{dV(\bar{X}^s(t))}{dt} < 0$ for all $x \neq x^*$.

Without loss of generality, we assume that $\lambda_1 > \lambda_2$ and define $V(x) = \lambda_1 - \lambda_2 + \gamma(x_2 - x_1)$. Writing $V(x)$ as $V(x) = \lambda_1 e^{-qx_2} - \lambda_2(1 - e^{-qx_1}) - \gamma x_1 - (\lambda_2 e^{-qx_1} - \lambda_1(1 - e^{-qx_2}) - \gamma x_2)$, we have $V(x^*) = 0$ and $V(x) \neq 0$ for all $x \neq x^*$. Applying Corollary 3.8 we have $x_1 - x_2 < \frac{\lambda_1 - \lambda_2}{\gamma}$ and hence, $V(x) > 0$. The second condition follows as

$$\frac{dV(\bar{X}^s(t))}{dt} = \gamma\left(\frac{d\bar{X}_2^s(t)}{dt} - \frac{d\bar{X}_2^s(t)}{dt}\right) = \lambda_2 - \lambda_1 + \gamma(\bar{X}_1^s(t) - \bar{X}_2^s(t)) = -V(\bar{X}^s(t)),$$

which is negative. Therefore, x^* is globally asymptotically stable: for all initial conditions, $\bar{X}^s(t) \to x^*$ as $t \to \infty$.

When $\lambda_1 = \lambda_2 = \lambda$, Corollary 3.8 implies that $\bar{X}_1^s(t) = \bar{X}_2^s(t)$. Denoting $\tilde{X}(t) = \bar{X}_1^s(t) = \bar{X}_2^s(t)$ and $\tilde{x}^* = x_1^* = x_2^*$ we need to show that $\tilde{X}(t) \to \tilde{x}^*, t \to \infty$, where $\tilde{X}(t)$ and $\tilde{x}^*$ satisfy the following equations:

$$\frac{\tilde{X}(t)}{dt} = 2\lambda e^{-q\tilde{X}(t)} - \lambda - \gamma\tilde{X}(t), \tag{3.15}$$

$$0 = 2\lambda e^{-q\tilde{x}^*} - \lambda - \gamma\tilde{x}^* \tag{3.16}$$

The righthand side of (3.16) is a decreasing function of $\tilde{x}^*$ and can be seen to have a unique solution. Equation (3.15) defines a gradient system with potential function $U(x) = \lambda x + \frac{1}{2}\gamma x^2 + \frac{2\lambda}{q}e^{-qx}$, i.e., it can be written as $\frac{\tilde{X}(t)}{dt} = -\bigtriangledown U(\tilde{X}(t))$ where $U(x)$ is a continuously differentiable, single valued scalar function. Hence, using Theorem 7.2.1 in Strogatz [37] $\tilde{X}(t) \to \tilde{x}^*, t \to \infty$. □

The fixed point x^* in Proposition 3.2 can be thought of as the long run average numbers of users in the system. Now, we analyze how x^* behaves for different values of the abandonment rate γ. It is reasonable to expect that x^* should decrease as abandonment rate increases, which always holds for the user class with the higher arrival rate. However, Proposition 3.3 shows that for the class with lower arrival rate x^* first increases and then decreases as γ increases.

Proposition 3.3. *Suppose $\lambda_1 \geq \lambda_2$. Then the long run average number of user-1 x_1^* decreases as the abandonment rate γ increases, while the long run average number of user-2 x_2^* increases when $\frac{\lambda_1-\lambda_2}{\gamma} > \frac{\gamma x_1^*}{q\lambda_1(1-e^{-qx_2^*})+q\gamma x_2^*}$ and decrease when $\frac{\lambda_1-\lambda_2}{\gamma} < \frac{\gamma x_1^*}{q\lambda_1(1-e^{-qx_2^*})+q\gamma x_2^*}$.*

Proof. Manipulating Equation (3.13) to obtain x_2^*, substituting in Equation (3.14) and doing cancellations, we get

$$\ln(\lambda_2(1-e^{-qx_1^*})+\gamma x_1^*) = -qx_1^* - \frac{q(\lambda_2-\lambda_1)}{\gamma} + \ln\lambda_1 \tag{3.17}$$

Taking the implicit derivative of x_1^* with respect to γ, we obtain

$$x_1^* + \gamma\frac{dx_1^*}{d\gamma} + \frac{\gamma}{q}\frac{d}{d\gamma}[\ln(\lambda_2(1-e^{-qx_1^*})+\gamma x_1^*)] + \frac{\ln(\lambda_2(1-e^{-qx_1^*})+\gamma x_1^*)}{q} - \frac{\ln\lambda_1}{q} = 0. \tag{3.18}$$

Letting $D_1 = \lambda_2(1-e^{-qx_1^*})+\gamma x_1^*, D_2 = \gamma\lambda_2 + \gamma^2 x_1^* + \frac{\gamma^2}{q}$ and substituting Equation (3.17) into Equation (3.18) to get rid of the logarithm terms, we get

$$\frac{dx_1^*}{d\gamma} = \frac{D_1}{D_2}(\frac{\lambda_2-\lambda_1}{\gamma} - \frac{\gamma x_1^*}{qD_1}) \tag{3.19}$$

Since D_1 and D_2 are always positive, when $\lambda_1 \geq \lambda_2$, the right hand side of Equation (3.19) is always negative, and hence as γ increases x_1^* increases. Interchanging x_1^* and λ_1 with x_2^* and λ_2 the right hand side of Equation (3.19) is positive when $\frac{\lambda_1-\lambda_2}{\gamma} > \frac{\gamma x_2^*}{q\lambda_1(1-e^{-qx_2^*})+q\gamma x_2^*}$ and negative when $\frac{\lambda_1-\lambda_2}{\gamma} < \frac{\gamma x_2^*}{q\lambda_1(1-e^{-qx_2^*})+q\gamma x_2^*}$. Hence, the conclusion for x_2^* follows. □

Proposition 3.3 shows that as γ increases, the limiting number of users for the class with lower arrival rate first increases and then decreases and the limiting number of users for the class with higher arrival rate decreases monotonically, which coincides with the observation in Figure 3.3. This behaviour can be explained as follows. As the abandonment rate increases, users from both classes tend to abandon the system a lot faster and hence the arriving users from the class with lower arrival rate are less likely find a match. The decrease in the number of matches is higher than the increase in the abandonments and as a result we observe a certain level of accumulation in the limit for users from the class with lower arrival rates.

3.4.2 Diffusion Limits

Now, we move to the discussion on the diffusion limits when the matching probability and the abandonment rate are both scaled to study the fluctuations of the queue lengths around the fluid limit $\bar{X}^s(t)$. We define

$$\hat{X}_i^{s,n}(t) = \frac{X_i^{s,n}(nt) - \bar{X}_i^s(nt)}{\sqrt{n}},\ t \geq 0.$$

To prove weak convergence we again use convergence of generators utilizing the techniques in [19].

Theorem 3.9. *Suppose $\bar{X}^s(t) = (\bar{X}_1^s(t), \bar{X}_2^s(t))^\intercal$ is the unique solution to the system of ODEs given by (3.4)-(3.6). Denote*

$$\begin{aligned}
a_1(t) &= q\lambda_2 e^{-q\bar{X}_1^s(t)},\\
a_2(t) &= q\lambda_1 e^{-q\bar{X}_2^s(t)},\\
\sigma_1(t) &= \sqrt{\lambda_1 e^{-q\bar{X}_2^s(t)} + \lambda_2(1 - e^{-q\bar{X}_1^s(t)}) + \gamma\bar{X}_1^s(t)},\\
\sigma_2(t) &= \sqrt{\lambda_2 e^{-q\bar{X}_1^s(t)} + \lambda_1(1 - e^{-q\bar{X}_2^s(t)}) + \gamma\bar{X}_2^s(t)},
\end{aligned}$$

and further define $z(t) = \int_0^t e^{\gamma s}\sigma_2(s)dB_2(s) - \int e^{\gamma s}\sigma_1(s)dB_1(s)$, $z_3(t) = e^{\int_0^t a_1(s)+a_2(s)ds}$ and $z_1(t) = e^{-\int_0^t a_1(s)+a_2(s)ds}\left(-\int_0^t z_3(s)a_2(s)z(s)ds + \int_0^t z_3(s)e^{\gamma s}\sigma_1(s)dB_1(s)\right)$. Then we have $\hat{X}^{s,n}(t) \Rightarrow \hat{X}^s(t)$, where $\hat{X}^s = (\hat{X}_1^s(t), \hat{X}_2^s(t))$

$$\hat{X}_1^s(t) = e^{-\gamma t}z_1(t) \tag{3.20}$$

$$\hat{X}_2^s(t) = e^{-\gamma t}(z_1(t) + z(t)). \tag{3.21}$$

Proof. Let $\nabla F(x) = \begin{pmatrix} \frac{\partial F_1}{x_1} & \frac{\partial F_1}{x_2} \\ \frac{\partial F_2}{x_1} & \frac{\partial F_2}{x_2} \end{pmatrix} = \begin{pmatrix} -q\lambda_2 e^{-qx_1} - \gamma & -q\lambda_1 e^{-qx_2} \\ -q\lambda_2 e^{-qx_1} & -q\lambda_1 e^{-qx_2} - \gamma \end{pmatrix}$,

$$\sigma(x) = \begin{pmatrix} \sqrt{\lambda_1 e^{-qx_2} + \lambda_2(1-e^{-qx_1}) + \gamma x_1} & 0 \\ 0 & \sqrt{\lambda_2 e^{-qx_1} + \lambda_1(1-e^{-qx_2}) + \gamma x_2} \end{pmatrix},$$

and $\bar{X}^s(t) = (\bar{X}_1^s(t), \bar{X}_2^s(t))^\intercal$ is the unique solution to system of ODE given by (3.4)-(3.6). We first show that $\hat{X}^{s,n}(t) \Rightarrow \hat{X}^s(t)$, where $\hat{X}^s(t)$ is the unique solution to the stochastic differential equation

$$d\hat{X}^s(t) = \sigma(\bar{X}^s(t))dB_t + \nabla F(\bar{X}^s(t))\hat{X}^s(t)dt, \tag{3.22}$$

starting from $\hat{X}^s(0) = (0,0)^\intercal$, where $B = (B_1, B_2)^\intercal$ is a two-dimensional standard Brownian motion.

Defining $\mathbb{S}$ as in the proof of Theorem 3.6, the weak convergence follows from Lemma 5.5 in [19], if we can show the conditions below hold: (remember that $F(x)$ is defined as in Equation (3.7)).

(a) $F(x)$ is continuously differentiable on $\mathbb{S}$,

(b) $\sup_{x\in\mathbb{S}} \sqrt{n} \left| \frac{\partial m^n(x,\theta)}{\partial\theta}|_{\theta=0} - F(x) \right| \to 0$,

(c) $\frac{\partial^2 m(x,\theta)}{\partial\theta^2}|_{\theta=0}$ is Lipschitz continuous in x on $\mathbb{S}$, where $m(x,\theta)$ is defined by

$$\begin{aligned} m(x,\theta) = {} & (\lambda_1(1-e^{-qx_2}) + \gamma x_2)e^{-\theta_2} + (\lambda_2(1-e^{-qx_1}) + \gamma x_1)e^{-\theta_1} \\ & + \lambda_1 e^{-qx_2} e^{\theta_1} + \lambda_2 e^{-qx_1} e^{\theta_2}. \end{aligned}$$

Condition (a) is trivial and condition (b) reduces to showing $\sqrt{n}\left(\left(1-\frac{q}{n}\right)^{nx} - e^{-qx}\right)$ converges to 0, which is elementary calculus and hence (b) holds as well. Finally

$$\left.\frac{\partial^2 m(x,\theta)}{\partial\theta^2}\right|_{\theta=0} = \begin{pmatrix} \lambda_1 e^{-qx_2} + \lambda_2(1-e^{-qx_1}) + \gamma x_1 & 0 \\ 0 & \lambda_2 e^{-qx_1} + \lambda_1(1-e^{-qx_2}) + \gamma x_2 \end{pmatrix}$$

which is Lipschitz on $\mathbb{R}^2_{\geq 0}$. Using Lemma 5.5 in [19], $\hat{X}^n \Rightarrow \hat{X}^s$ as $n \to \infty$, where

$\hat{X}^s(t)$ is the unique solution to the stochastic differential equation (3.22).

Next we show that (3.20) and (3.21) together is the unique solution to (3.22) which can be expressed as:

$$d\hat{X}_1^s(t) = (-a_1(t) - \gamma)\hat{X}_1^s(t)dt - a_2(t)\hat{X}_2^s(t)dt + \sigma_1(t)dB_1(t)$$
$$d\hat{X}_2^s(t) = -a_1(t)\hat{X}_1^s(t)dt - (a_2(t) + \gamma)\hat{X}_2^s(t)dt + \sigma_2(t)dB_2(t).$$

Defining $z_i(t) = e^{\gamma t}\hat{X}_i^s(t), i = 1, 2$, we obtain

$$\begin{aligned} dz_1(t) &= e^{\gamma t} d\hat{X}_1^s(t) + \gamma e^{\gamma t}\hat{X}_1^s(t)dt \\ &= (-a_1(t) - \gamma)e^{\gamma t}\hat{X}_1^s(t)dt - e^{\gamma t}a_2(t)\hat{X}_2^s(t)dt \\ &\quad + \gamma e^{\gamma t}\hat{X}_1^s(t)dt + e^{\gamma t}\sigma_1(t)dB_1(t) \\ &= -a_1(t)z_1(t)dt - a_2 z_2(t)dt + e^{\gamma t}\sigma_1(t)dB_1(t), \end{aligned} \tag{3.23}$$

and similarly $dz_2(t) = -a_1(t)z_1dt - a_2(t)z_2(t)dt + e^{\gamma t}\sigma_2(t)dB_2(t)$. Furthermore, letting $z(t) = z_2(t) - z_1(t)$ we have

$$dz(t) = e^{\gamma t}(\sigma_2(t)dB_2(t) - \sigma_1(t)dB_1(t)). \tag{3.24}$$

Solving Equation (3.24) directly, we obtain that

$$z(t) = \int_0^t e^{\gamma s}\sigma_2(s)dB_2(s) - \int e^{\gamma s}\sigma_1(s)dB_1(s).$$

Substituting that $z_2(t) = z(t) + z_1(t)$ into the Equation (3.23), we have

$$dz_1(t) = (-a_1(t) - a_2(t))z_1(t)dt - a_2(t)z(t)dt + e^{\gamma t}\sigma_1(t)dB_1(t).$$

Moving the term of $z_1(t)$ to the right hand side, we have

$$dz_1(t) + (a_1(t) + a_2(t))z_1(t)dt = -a_2(t)z(t)dt + e^{\gamma t}\sigma_1(t)dB_1(t).$$

Multiplying both sides by $z_3(t) = e^{\int_0^t a_1(s)+a_2(s)ds}$, we have

$$\begin{aligned} &e^{\int_0^t a_1(s)+a_2(s)ds}dz_1(t) + (a_1(t) + a_2(t))e^{\int_0^t a_1(s)+a_2(s)ds}z_1(t)dt \\ &= z_3(t)(-a_2(t)z(t)dt + e^{\gamma t}\sigma_1(t)dB_1(t)), \end{aligned}$$

and thus,

$$d(z_1(t)e^{\int_0^t a_1(s)+a_2(s)ds}) = z_3(t)(-a_2(t)z(t)dt + e^{\gamma t}\sigma_1(t)dB_1(t)).$$

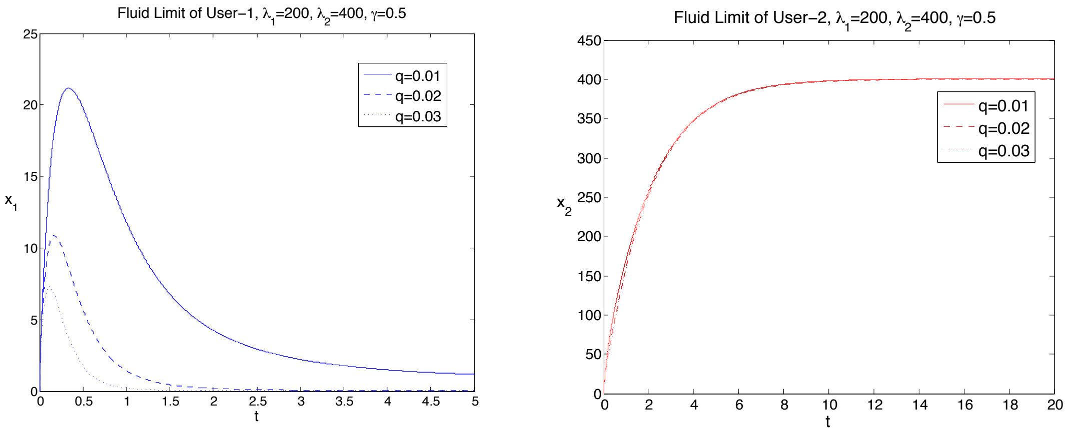
As a result $z_1(t) = e^{-\int_0^t a_1(s)+a_2(s)ds}\left(-\int_0^t z_3(s)a_2(s)z(s)ds + \int_0^t z_3(s)e^{\gamma s}\sigma_1(s)dB_1(s)\right)$, and $X_1^s(t) = e^{-\gamma t}z_1(t)$ and $X_2^s(t) = e^{-\gamma t}(z_1(t) + z(t))$ follow. □

Theorem 3.9 indicates that if the fluid limit $\bar{X}^s(t)$ is given the diffusion limit can be fully characterized analytically. However, as we have seen in Section 3.4.1, it is not always possible to analytically solve the ODEs for the fluid limit. In the next section, we present numerical experiments to study fluid and diffusion limits presented in this section.

3.5 Numerical Experiments

In Section 3.4, we show that when the matching probability and abandonment rate are scaled to go to zero along with the time and space, the fluid and diffusion limits can be expressed as the unique solutions to some systems of ODEs and SDEs which do not have explicit solutions in general. To gain some insight into the solutions, we study numerical approximations in this section. We use Euler and Euler-Maruyama method to obtain numerical solutions of ODEs (3.4)-(3.6) and SDEs (3.22), respectively. (See Kloeden and Platen [30] for more details.)

To study the fluid limit which is the unique solution to the system of ODEs (3.4)-(3.6), we apply Euler method with step size $h = 10^{-6}$. In our first experiment, we test the effect of the matching probability q on the fluid limits. First we consider the case $\lambda_1 < \lambda_2$ by setting $\lambda_1 = 200, \lambda_2 = 400, \gamma = 0.5$ and compute the fluid limits for $q = 0.01, 0.02, 0.03$. The results are given in Figure 3.1. We observe that for the



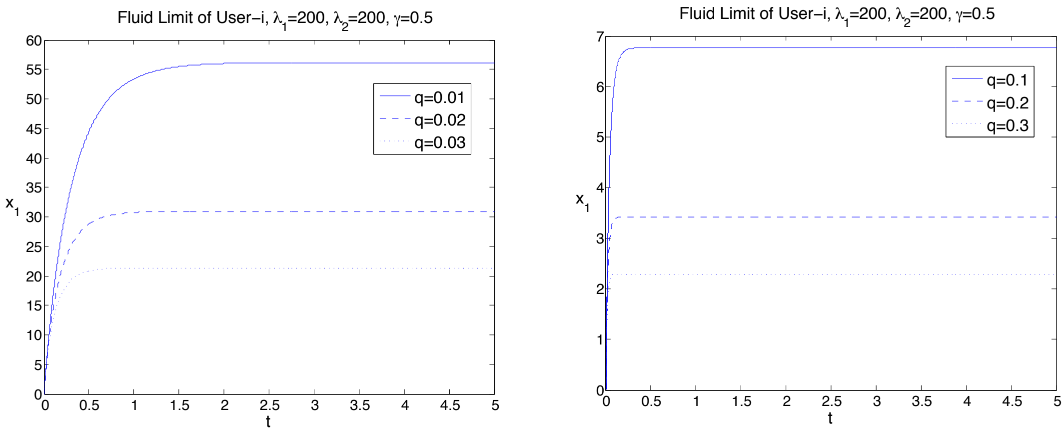
(a) Fluid Limit of User 1 for different q (b) Fluid Limit of User 2 for different q

Figure 3.1: Fluid Limits when $\lambda_1 < \lambda_2$ for various q

class with lower arrival rate, the number of users in the system demonstrates a very sharp increase at the beginning and then decreases approaching a limit as t goes to infinity. We see that there is a considerable difference between the number of users corresponding to different matcing probabilities for this class. On the other hand, the number of users for the class with higher arrival rate grows monotonically converging to its supremum as t goes to infinity. Surprisingly, the matching probability does not play a significant role for this class and the fluid limits corresponding to different matching probabilities are very close.

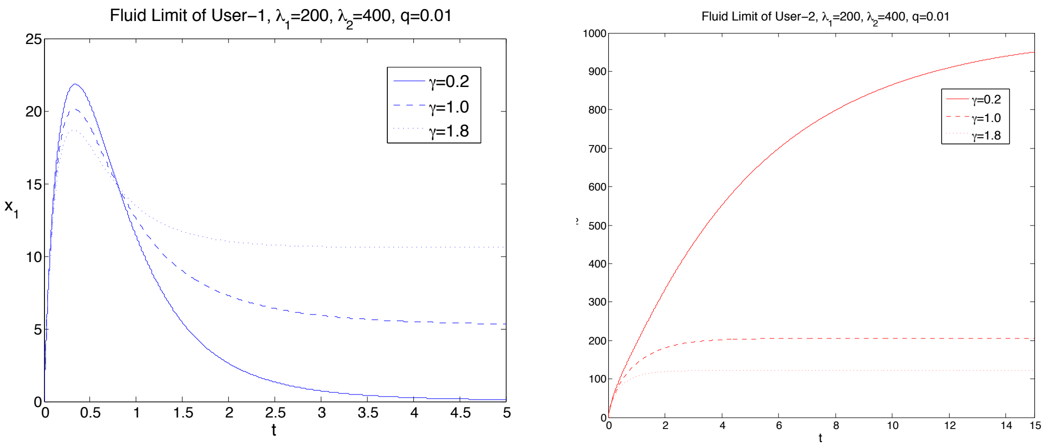
To test the case where $\lambda_1 = \lambda_2$, we performed the same experiment by taking $\lambda_1 = \lambda_2 = 200$. Figure 3.2 demonstrates that the number of users for both classes increase monotonically as t goes to infinity approaching to the supremum, which is very similar to the behaviour of the class with higher arrival rate when the rates are not equal. However, in this case the matching probability has a major effect on the limiting number of users and as q increases the number of users in the system decreases. Also as q gets larger we see that the number of users increases to its supremum faster and the fluid limit is steeper.

Next we study how the effect of the abandonment rate γ on the number of users in the system. In this set of experiments, we set the arrival rates $\lambda_1 = 200, \lambda_2 = 400$ and the matching probability $q = 0.01$ and vary the abandonment rate. Figure 3.3 shows that the shape of fluid limits are not affected by the changes in the abandonment



(a) Fluid Limit of User 1 and 2 for different q (b) Fluid Limit of User 1 and 2 for different q

Figure 3.2: Fluid Limits when $\lambda_1 = \lambda_2$ for various q

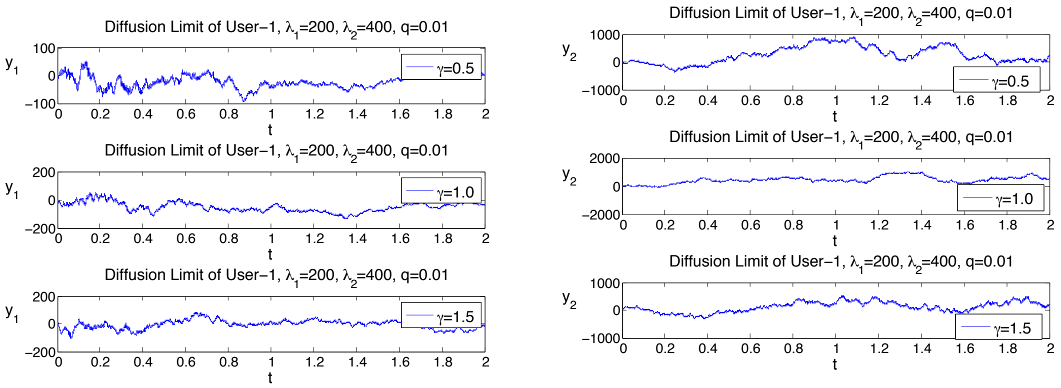


(a) Fluid Limit of User 1 for various γ (b) Fluid Limit of User 2 for various γ

Figure 3.3: Fluid Limits when $\lambda_1 < \lambda_2$ for various γ

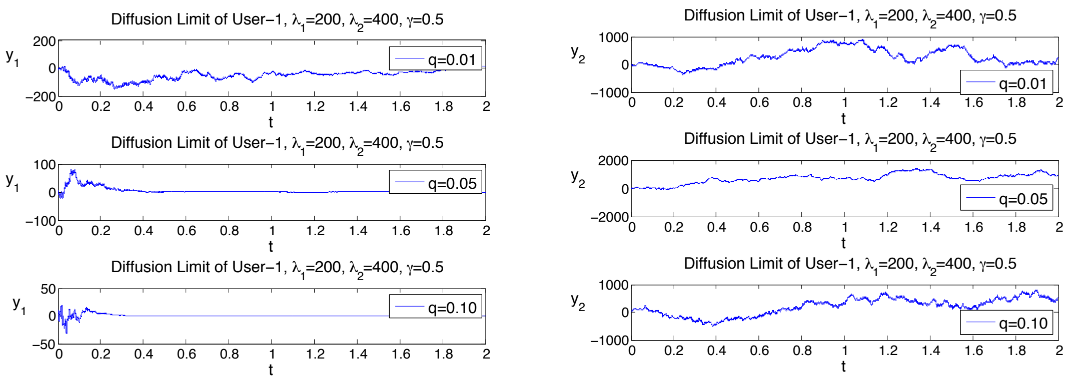
rate, i.e., the number of users for the class with lower arrival rate first increases and then decreases and the number of users for the class with higher arrival rate decreases monotonically. We also see that when there are abandonments the number of users for the class with lower arrival rate does not converge to 0 as t goes to infinity. In agreement with Proposition 3.3, we see that the limiting number of users for the class with lower arrival rate increases in our experiments as the abandonment rate increases.

Now, we discuss numerical approximation to diffusion limit, which is the unique solution to the system of SDEs (3.22). In our experiments, we apply the Euler- Maruyama method with the step size $h = 10^{-6}$. We again start with the case when the arrival rates are not equal and set $\lambda_1 = 200, \lambda_2 = 400$. Figures 3.4 and 3.5 demonstrate some



(a) Diffusion Limit of User 1 for various γ (b) Diffusion Limit of User 2 for various γ

Figure 3.4: Diffusion Limits when $\lambda_1 < \lambda_2$ for various γ



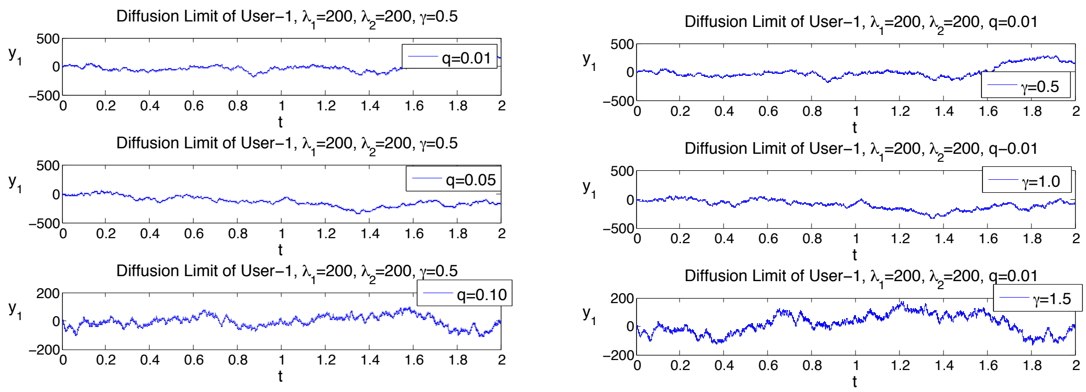
(a) Diffusion Limit of User 1 for various q (b) Diffusion Limit of User 2 for various q

Figure 3.5: Diffusion Limits when $\lambda_1 < \lambda_2$ for various q

sample paths. We see that the fluctuations for the class with higher arrival rates are always bigger. When q is fixed we see that the changes in γ does not have a major effect on fluctuations. We also see that the fluctuations for the class with lower arrival rate diminish as t increases. As q increases the fluctuations diminish a lot faster. Finally we observe in Figure 3.6 that when the arrival rates are equal and set to be $\lambda_1 = \lambda_2 = 200$, both queue length processes keep fluctuating as usual.

3.6 Conclusion

In this chapter, we propose two different scalings to obtain fluid and diffusion approximations to the queue length processes of probabilistic matching systems. For the first approach, the space and time are scaled while the matching probability is kept fixed.



(a) Diffusion Limit of User 1 for various q (b) Diffusion Limit of User 1 for various q

Figure 3.6: Diffusion Limits when $\lambda_1 = \lambda_2$ for various q and γ

Under this scaling, the matching probability q does not play any role in the fluid limit and the minimum of the queue lengths converges to zero. We suggest that this scaling should be used when the matching probability is considerably high.

The second scaling considers the systems in which the probability to match for each pair of users is small. The effect of abandonments is also taken into account and the matching probability and the departure rate are scaled along with time and space in this regime. The limiting processes enable us to address the matching probability explicitly. Unfortunately, the resulting system of ODEs cannot be solved analytically in general, although when there are no abandonments it is possible to obtain an analytical solution. In Chapter 2, some performance measures were shown to be insensitive to the matching probability under certain admission control policies. Using fluid limits, we show that the difference between the average queue lengths of different classes of users is also independent of the matching probability. We also analyse the asymptotic behaviour of the fluid limits in this scaling. First we show that when abandonment rate is zero, the two fluid limits, obtained with and without scaling the matching probability, converges to each other with time. We further show that when there are abandonments, the fluid limits converge to a unique fixed point, which represents the long run average number of users in the system. Conducting analysis on the fixed point, we reveal that as the abandonment rate increases, the number of users for the class with lower arrival rate first experiences an increase and then decrease while the number of users for the class with higher arrival rate decreases monotonically.

As analytical expressions are not available for fluid and diffusion limits, we resort to numerical methods to study the corresponding ODEs and SDEs. We observe that for the class with higher arrival rate, the number of users in the system increases monotonically. On the other hand, the users from the class with lower arrival rate first tend to accumulate in the system and then decrease to a limit as time goes to infinity. This limit is different from zero and increases as the abandonment rate increases agreeing with our theoretical analysis. This indicates that there are always a significant number of users waiting in the system from both classes.

We believe the fluid and diffusion limits introduced in this chapter will be helpful in many research directions. For example, the approximations introduced here can be used to study the performance of admission control policies which are intractable using exact methods. Another promising research direction which we introduce in the next chapter is to identify optimal and asymptotically optimal policies to maximize profit generated by charging users admission fees.

Chapter 4

Optimal Control and Dynamic Pricing

4.1 Introduction

In this chapter we focus on optimal control and dynamic pricing of probabilistic matching systems. These ideas are presented through an employment portal, an important example of probabilistic matching systems. The users of an employment portal are employees and employers and we assume that employees arrive with a larger rate than employers, i.e., $\lambda_1 > \lambda_2$. When an employer arrives at the system, she is charged a fee to enter the system and she makes a decision on whether to join the system based on the price, the information of the system given to her and her own willingness-to-pay function. As a consequence, the entry price affects the arrival process of employers. On the other hand, employees are accepted to the system without paying a fee. After employees and employers join the system, the dynamics of user behaviour is the same as described in Chapter 3. In particular we assume that users are impatient and leave the system after waiting an exponentially distributed time with a strictly positive rate $\gamma > 0$. When an employer abandons the system, the system operator pays a compensation fee $a > 0$ to the employer. We further assume that employers are not strategic, i.e., they do not abandon the system simply to earn the compensation fee.

To maximize the profit, the system operator may change the price p at any time based on the current system state. The entry of an employer brings a profit p to the system while a compensation fee of a is paid to the employer if she abandons the system without being matched. In general, the system operator maximizes the profit by attracting employers to join the system when there is a high probability that an arriving employer will find a match before she abandons the system. Intuitively speaking, when there are a large of number of employees in the system it is desirable to accept an arriving employer. However, if employers accumulate in the system, it might not be so beneficial to encourage the arriving employer to join the system. As we know that each employer has a maximum price that she is willing to pay to join the system, we control the entry of an arriving employer by dynamically setting the entry fee p. In this chapter, we study the natures of the pricing mechanisms with the objectives to maximize the expected total profit and the long run average profit respectively.

To define the objective function to maximize the expected total profit, it is crucial to specify the time horizon under consideration. When the system operator is looking at the performance of the system in a specific finite time period, the number of users remaining in the system at the end of the period is sometimes important. We include this factor by introducing a function based on the number of users left in the system at terminal time to the objective function. On the other hand, if we are considering the total profit in the infinite time horizon, it is desirable to bring in a discount factor $\rho > 0$ or to study the long run average profit. We introduce the formulation to the pricing problem with respect to different time horizon setting as follows. Let $\{p(t)\}_{t\geq 0}$ be the pricing process, i.e., $p(t)$ is the entry price for employers at time t, and $\{A_2^e(t)\}_{t\geq 0}$ is the counting process of employers accepted to the system influenced by the pricing process, which is called effective arrival process of employers. As defined in Chapter 3, $\{R_2(t)\}_{t\geq 0}$ is the abandonment process of employers. Given a pricing policy $\{p(t)\}_{t\geq 0}$ and a time horizon $T < \infty$, the finite horizon problem is to maximize the expected total profit in $[0, T]$:

$$J(T) = \mathbb{E}\left[\int_0^T (p(t)dA_2^e(t) - adR_2(t)) + h(X_1(T), X_2(T))\right], \tag{4.1}$$

where the exact form of $h(\cdot, \cdot)$ is the terminal cost function to be specified later. On the other hand, for the infinite horizon problem, we want to maximize the expected discounted total profit:

$$J(\infty) = \mathbb{E}\left[\int_0^\infty e^{-\rho t}(p(t)dA_2^e(t) - adR_2(t))\right]. \tag{4.2}$$

To study the long run average profit maximization problem, we want to maximize

$$J(\text{average}) = \lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\int_0^T (p(t)dA_2^e(t) - adR_2(t))\right]. \tag{4.3}$$

Our goal is to find the optimal pricing policy i.e., $\{p(t)\}_{t\geq 0}$ to maximize (4.1), (4.2) and (4.3) respectively.

Optimal pricing for queueing systems has been extensively studied in the last decades. We classify related literature into two streams: optimal pricing for unobservable and observable models. Hassin and Haviv [28] provides a detailed discussion for unobservable and observable models on queues with incentives. In an observable model, users see how many people are in the system, while in an unobservable model this information is not revealed.

In the unobservable queueing literature, Bäuerle [6] consider a general control problem for networks with linear dynamics by using fluid approximation to provide a lower bound to the stochastic network problem and show the existence of an average-cost optimal decision rule for scheduling problems in multiclass queueing networks. Maglaras [34] studies the profit maximizing problem for a multiclass $M_n/M/1$ queue with controllable arrival rates, general demand curves, and linear holding costs using a fluid model relaxation. Ward and Kumar [38] use the solution of an approximating singular diffusion control problem to construct an admission control policy for an $GI/GI/1$ queue with impatient customers. Atar and Reiman [5] study a dynamic pricing problem in finite time horizon in diffusion scale and proves diffusion-scale asymptotic optimality of a dynamic pricing policy that mimics the behaviour of the Brownian bridge. Afeche [2] looks at a multi-product $M/M/1$ queue under a market of heterogeneous price- and delay-sensitive users with a profit-maximizing objective.

On the other hand, Naor [36] is a pioneer in the area of pricing observable queues. Chen and Frank [16] introduce the idea of state dependent pricing to Naor's model and show that there is a threshold at which users are prevented to join in the system. Yildirim and Hasenbein [42] investigate a similar model as Chen and Frank [16] but with batch arrivals. After solving the pricing problem, Yildirim and Hasenbein formulate the control problem and determine the optimal admission thresholds for groups of different sizes implicitly. Borgs et al. [9] is another work which considers Chen and Frank's model and provides the first derivation of the optimal threshold in closed form and a formula for the maximum revenue under the optimal threshold.

In this chapter, we first study the optimal pricing problem for the unobservable model in Section 4.2 using the idea motivated by the fluid limit discussed in Chapter 3. Then we analyse the state dependent optimal pricing for the observable model in Section 4.3.

4.2 Unobservable Model

4.2.1 Model Description

Assumptions

In an unobservable model, employers are not informed about the number of employers and employees in the system upon arrival. They make decisions on whether to join the system based on the price and their willingness-to-pay function. We start our discussion by reviewing some standard economic assumptions on how the entry price influences the arrival rate of employers as introduced by Gallego and Van Ryzin [22].

We assume that the employment portal is operated in a market with imperfect competition and the effective arrival process of employers is a non-homogeneous Poisson process whose arrival rate at time t, i.e., $\tilde{\lambda}_2(t)$, is influenced by the price $p(t)$ through a function $\tilde{\lambda}_2(t) = \tau(p(t))$. We assume that $\tau(p)$ satisfies the following properties:

(i) $\tau(p)$ is non-increasing in p,

(ii) there is a one-to-one mapping between p and $\tau(p)$ and

(iii) $\tau(p)$ has an inverse $\tau^{-1}(\tilde{\lambda}_2)$.

Following the definition of price set in Gallego and Van Ryzin [22], we assume that the price $p(t)$ is selected from from the set of allowable price $\mathcal{P} \subseteq \mathbb{R}_{\geq 0}$. Further we restrict that $\tau(p) \leq \lambda_2$ holds for all possible entry prices chosen from $\mathcal{P}$, where $\lambda_2 < \lambda_1$ is the largest arrival rate of employers achieved when $p = 0$. Then the allowable set of arrival rates is denoted by $\Lambda = \{\tilde{\lambda}_2 : \tilde{\lambda}_2 = \tau(p), p \in \mathcal{P}\} \subseteq [0, \lambda_2]$. Further, the income rate of the system at time t can be expressed through a function of the arrival rate,

$$I(\tilde{\lambda}_2(t)) = p(t)\tilde{\lambda}_2(t) = \tau^{-1}(\tilde{\lambda}_2(t))\tilde{\lambda}_2(t).$$

We assume that the income rate function is continuous, bounded, concave and satisfies $I(\tilde{\lambda}_2)$ satisfies $\lim_{\tilde{\lambda}_2 \to 0} I(\tilde{\lambda}_2) = 0$, so that $\tilde{\lambda}_2^* = \min\{\tilde{\lambda}_2 \in \Lambda : I(\tilde{\lambda}_2) = \max_{\tilde{\lambda}_2 \in \Lambda} I(\tilde{\lambda}_2)\}$ is well-defined. As a consequence, we are changing the control variable from the pricing process $\{p(t)\}_{t \geq 0}$ to be the arrival rate of employers $\{\tilde{\lambda}_2(t)\}_{t \geq 0}$.

Formulation

Recall that the queue length processes of employees and employers are denoted as $\{X_1(t)\}_{t \geq 0}$ and $\{X_2(t)\}_{t \geq 0}$, respectively. As each employer abandons the system with rate $\gamma > 0$, we have $\mathbb{E}[dR_2(t)] = \mathbb{E}[\gamma X_2(t)dt]$ holds for any $t \geq 0$. Therefore, the expected total profit of the system operator under a pricing policy $\{p(t)\}_{t \geq 0}$ in the interval $[0, T]$ can be expressed as follows:

$$\begin{aligned}\mathbb{E}\left[\int_0^T (p(s)dA_2^e(s) - adR_2(s))\right] &= \mathbb{E}\left[\int_0^T (p(s)\tilde{\lambda}_2(s) - a\gamma X_2(s))ds\right] \\ &= \mathbb{E}\left[\int_0^T (I(\tilde{\lambda}_2(s)) - a\gamma X_2(s))ds\right].\end{aligned}$$

Define the terminal cost function as $h(x_1, x_2) = cx_2$, where $c \leq 0$. The finite horizon problem is to find a set of solutions $\{\tilde{\lambda}_2(t)\}_{0 \leq t \leq T}$ to achieve the maximum expected profit $J(T)$, where

$$J(T) = \mathbb{E}\left[\int_0^T (I(\tilde{\lambda}_2(t)) - a\gamma X_2(t))dt + cX_2(T)\right]. \tag{4.4}$$

For the infinite horizon problem, the discounted cost objective function can be expressed as

$$J(\infty) = \mathbb{E}\left[\int_0^\infty e^{-\rho t}(I(\tilde{\lambda}_2(t)) - a\gamma X_2(t))dt\right]. \tag{4.5}$$

Fluid Optimization Problem

An exact study of probabilistic matching systems as introduced in Chapter 2 implies that the queue length process is intractable and so are the control problems 4.4 and 4.5 defined above. Therefore we look for the optimal control in the limiting regime.

In many real world problems the matching probability for each pair of employees and employers is fairly small. Motivated by the idea of fluid approximations for systems with small matching probability that is introduced in Theorem 3.6 in Chapter 3, we therefore suggest to approximate the queue length processes $\{X_1(t)\}_{t\geq 0}$ and $\{X_2(t)\}_{t\geq 0}$ by the deterministic processes $\{x_1(t)\}_{t\geq 0}$ and $\{x_2(t)\}_{t\geq 0}$, respectively, where $x_1(t)$ and $x_2(t)$ are the unique solution of the following set of ODEs:

$$x_1(0) = x_2(0) = 0, \tag{4.6}$$

$$\frac{dx_1(t)}{dt} = \lambda_1 e^{-qx_2(t)} + \tilde{\lambda}_2(t)e^{-qx_1(t)} - \tilde{\lambda}_2(t) - \gamma x_1(t), \tag{4.7}$$

$$\frac{dx_2(t)}{dt} = \lambda_1 e^{-qx_2(t)} + \tilde{\lambda}_2(t)e^{-qx_1(t)} - \lambda_1 - \gamma x_2(t). \tag{4.8}$$

In Section 4.2.2 Section 4.2.3, we solve the finite horizon problem (4.4) and the infinite horizon problem (4.5) respectively subject to the system equations (4.6)-(4.8). For notational convenience in later sections, here we denote

$$\begin{aligned}
x(t) &= (x_1(t), x_2(t))^\intercal, \\
h(x(T)) &= cx_2(T), c \leq 0, \\
g(x(t), \tilde{\lambda}_2(t)) &= I(\tilde{\lambda}_2(t)) - a\gamma x_2(t), \\
f_1(x(t), \tilde{\lambda}_2(t)) &= \lambda_1 e^{-qx_2(t)} + \tilde{\lambda}_2(t)e^{-qx_1(t)} - \tilde{\lambda}_2(t) - \gamma x_1(t), \\
f_2(x(t), \tilde{\lambda}_2(t)) &= \lambda_1 e^{-qx_2(t)} + \tilde{\lambda}_2(t)e^{-qx_1(t)} - \lambda_1 - \gamma x_2(t) \\
f(x(t), \tilde{\lambda}_2(t)) &= (f_1(x(t), \tilde{\lambda}_2(t)), f_2(x(t), \tilde{\lambda}_2(t))^\intercal.
\end{aligned}$$

4.2.2 Deterministic Finite Horizon Problem

In this section, we solve the deterministic finite horizon optimal control problem. Specifically, we look for an admissible control trajectory $\{\tilde{\lambda}_2(t)|t \in [0,T]\}$ which together with its corresponding state trajectory $\{x^{\tilde{\lambda}_2}(t)|t \in [0,T]\}$, solves the following problem:

$$\begin{aligned} \max_{\tilde{\lambda}_2} \quad & J^D(T) = \int_0^T g(x(t), \tilde{\lambda}_2(t))dt + h(x(T)) \\ \text{s.t.} \quad & \begin{cases} x(0) & = (0,0), \\ \frac{dx_i(t)}{dt} & = f_i(x(t), \tilde{\lambda}_2(t)), i = 1,2, 0 \leq t \leq T. \end{cases} \end{aligned} \tag{4.9}$$

This is a standard continuous-time optimal control problem, which can be solved by the method of The Pontryagin Minimum Principle. Now we introduce the concepts of Hamiltonian function, adjoint equation and Pontryagin Minimum Principle.

The Hamiltonian function that maps triplets $(x, \tilde{\lambda}_2, \psi) \in \mathbb{R}^2 \times \mathbb{R} \times \mathbb{R}^2$ to real numbers is given by

$$H(x, \tilde{\lambda}_2, \psi) := g(x, \tilde{\lambda}_2) + \psi^{\mathsf{T}} f(x, \tilde{\lambda}_2), \tag{4.10}$$

where ψ, which is called adjoint variable. We suppose that ψ is a solution to the following system of linear differential equations:

$$\frac{\psi_i(t)}{dt} = -\nabla_{x_i} H(x^*(t), \tilde{\lambda}_2^*(t), \psi(t)), i = 1,2, \tag{4.11}$$

where $\{(x^*(t), \tilde{\lambda}_2^*(t))|t \in [0,T]\}$ is an optimal state and control trajectory. The Minimum principle is stated as follows.

Theorem 4.1 (Minimum Principle (Bertsekas [7])). *Let $\{\tilde{\lambda}_2^*(t)|t \in [0,T]\}$ be an optimal control trajectory and let $\{x^*(t)|t \in [0,T]\}$ be the corresponding state trajectory. Let $\psi(t)$ be the solution of the adjoint equation defined by* (4.11) *with boundary condition*

$$\psi(T) = \triangledown h(x^*(T)),$$

where $h(\cdot)$ is the terminal cost function. Then for all $t \in [0, T]$,

$$\tilde{\lambda}_2^*(t) = \arg\max_{\tilde{\lambda}_2 \in \Lambda} H(x^*(t), \tilde{\lambda}_2, \psi(t)). \tag{4.12}$$

In our optimal control problem (4.9), the objective function is concave and the system equations satisfy Lipschitz and continuous conditions, we can easily show the existence of an optimal solution. (See Proposition 2.1 in Bertsekas [7] for more details.) Next we state our main result of the optimal solution to the deterministic control problem.

Theorem 4.2. *Let $\{\tilde{\lambda}_2^*(t)\}_{t\in[0,T]}$ with its corresponding state trajectory $\{x^*(t)\}_{t\in[0,T]}$ be an optimal solution for the problem* (4.9). *Then $\{\tilde{\lambda}_2^*(t)\}_{t\in[0,T]}$ and $\{x^*(t)\}_{t\in[0,T]}$ satisfies*

$$\tilde{\lambda}_2^*(t) = \arg\max_{\tilde{\lambda}_2 \in \Lambda}(I(\tilde{\lambda}_2) + \tilde{\lambda}_2((\psi_1(t) + \psi_2(t))e^{-qx_1^*(t)} - \psi_1(t))), \tag{4.13}$$

where $\psi(t) = (\psi_1(t), \psi_2(t))^\intercal$ solves the systems of ODE:

$$\begin{cases} \frac{\psi_1(t)}{dt} & = \gamma\psi_1(t) + (\psi_1(t) + \psi_2(t))\tilde{\lambda}_2^*(t)qe^{-qx_1^*(t)} \\ \frac{\psi_2(t)}{dt} & = \gamma(a + \psi_2(t)) + (\psi_1(t) + \psi_2(t))\lambda_1 qe^{-qx_2^*(t)} \\ \psi_1(T) & = 0, \\ \psi_2(T) & = c. \end{cases} \tag{4.14}$$

Proof. Let $\psi(t) = (\psi_1(t), \psi_2(t))^T$ be the adjoint variable corresponding the optimal solution $\{\lambda_2^*(t)\}_{t\in[0,T]}$ and $\{x^*(t)\}_{t\in[0,T]}$, then the Hamiltonian function and the adjoint system are as follows:

$$\begin{aligned} H(x, \tilde{\lambda}_2, \psi) = I(\tilde{\lambda}_2) - a\gamma x_2 + \psi_1(\lambda_1 e^{-qx_2} + \tilde{\lambda}_2 e^{-qx_1} - \tilde{\lambda}_2 - \gamma x_1) \\ + \psi_2(\lambda_1 e^{-qx_2} + \tilde{\lambda}_2 e^{-qx_1} - \lambda_1 - \gamma x_2), \end{aligned} \tag{4.15}$$

$$\begin{cases} \frac{\psi_1(t)}{dt} &= -\nabla_{x_1} H(x^*(t), \tilde{\lambda}_2^*(t), \psi(t)) = \gamma\psi_1(t) + (\psi_1(t)+\psi_2(t))\tilde{\lambda}_2 q e^{-qx_1^*(t)} \\ \frac{\psi_2(t)}{dt} &= -\nabla_{x_2} H(x^*(t), \tilde{\lambda}_2^*(t), \psi(t)) = \gamma(a+\psi_2(t)) + (\psi_1(t)+\psi_2(t))\lambda_1 q e^{-qx_2^*(t)} \\ \psi_1(T) &= \nabla_{x_1} h(x^*(T)) = 0, \\ \psi_2(T) &= \nabla_{x_2} h(x^*(T)) = c. \end{cases}$$

Using the Minimum Principle as stated in Theorem 4.1, we achieve

$$\tilde{\lambda}_2^*(t) = \arg \max_{\tilde{\lambda}_2(t) \in \Lambda} (I(\tilde{\lambda}_2(t)) + \tilde{\lambda}_2(t)((\psi_1(t)+\psi_2(t))e^{-qx_1^*(t)} - \psi_1(t))). \tag{4.16}$$

□

Theorem 4.2 provides a way of finding an optimal solution to the deterministic finite horizon problem. Looking at Equation (4.13), we notice that the function to be maximized consists of two parts, a concave function $I(\tilde{\lambda}_2)$ and a linear term, and hence is a concave function. Therefore it has a unique solution depending on the definition of the rate income function $I(\tilde{\lambda}_2)$ as introduced in Section 4.2.1. We thus can express the optimal control $\tilde{\lambda}_2(t)$ in terms of the adjoint variable $\psi(t)$ for a certain income rate function. Substituting this expression into the adjoint system defined by (4.14) and the system function in (4.9), we can then solve the system and then obtain the optimal solution. Corollary 4.3 presents the solution for a special case when the terminal function is defined to be 0 and the income rate function is defined as $I(\tilde{\lambda}_2(t)) = (\lambda_2 - \tilde{\lambda}_2(t))\tilde{\lambda}_2(t)$.

Corollary 4.3. *When the income rate function is defined as* $I(\tilde{\lambda}_2(t)) = (\lambda_2 - \tilde{\lambda}_2(t))\tilde{\lambda}_2(t)$, *the optimal solution* $\{\tilde{\lambda}_2^*(t)\}_{t\in[0,T]}$ *and its corresponding state trajectory* $\{x^*(t)\}_{t\in[0,T]}$ *satisfy*

$$\tilde{\lambda}_2^*(t) = \begin{cases} \lambda_2, & \text{if } b(t) > \lambda_2, \\ \frac{1}{2}b(t), & \text{if } 0 \le \frac{b(t)}{2} \le \lambda_2, \\ 0, & \text{if } b(t) < 0, \end{cases} \tag{4.17}$$

where $b(t) = \lambda_2 + [(\psi_1(t)+\psi_2(t))e^{-qx_1^*(t)} - \psi_1(t)]$ *and* $\psi_1(t), \psi_2(t)$ *solves the adjoint system* (4.14).

Proof. Substituting $I(\tilde{\lambda}_2(t)) = (\lambda_2 - \tilde{\lambda}_2(t))\tilde{\lambda}_2(t)$ in to the Equation (4.13), then

$$\begin{aligned}
\tilde{\lambda}_2^*(t) &= \arg \max_{\tilde{\lambda}_2(t)\in\Lambda} ((\lambda_2 - \tilde{\lambda}_2(t))\tilde{\lambda}_2(t) + \tilde{\lambda}_2(t)((\psi_1(t) + \psi_2(t))e^{-qx_1^*(t)} - \psi_1(t))) \\
&= \arg \max_{\tilde{\lambda}_2(t)\in\Lambda} (-\tilde{\lambda}_2(t)^2 + \tilde{\lambda}_2(t)(\lambda_2 + ((\psi_1(t) + \psi_2(t))e^{-qx_1^*(t)} - \psi_1(t)))) \\
&= \arg \max_{\tilde{\lambda}_2(t)\in\Lambda} (-\tilde{\lambda}_2(t)^2 + \tilde{\lambda}_2(t)b(t))
\end{aligned}$$

It is trivial that the concave function $-\tilde{\lambda}_2(t)^2 + \tilde{\lambda}_2(t)b(t)$ achieve its maximum at either its extreme points 0 or λ_2 or $\frac{1}{2}b(t)$. Hence the result as desired follows. □

Corollary 4.3 presents an example of finding an optimal solution to the control problem for a special simply defined income rate function. For any other income rate function which satisfies the assumptions introduced in Section 4.2.1, we can use similar method to find the optimal solution. Next we discus the infinite horizon problem.

4.2.3 Deterministic Infinite Horizon Problem

In this section, we discuss the deterministic infinite horizon optimal control problem. Distinct from the finite horizon problem discussed in Section 4.2.2, in the infinite horizon problem the terminal cost function $h(X_2(T))$ is no longer present in the objective function. Rather, we have a discount factor $\rho > 0$. Specifically, we find an admissible control trajectory $\{\tilde{\lambda}_2(t)|t > 0\}$ which together with its corresponding state trajectory $\{x^{\tilde{\lambda}_2}(t)|t > 0\}$, solves the following problem.

$$\begin{aligned}
&\max_{\tilde{\lambda}_2} \quad J^D(\infty) = \int_0^\infty e^{-\rho t} g(x(t), \tilde{\lambda}_2(t))dt \\
&\text{s.t.} \quad \begin{cases} x(0) & = (0,0), \\ \frac{dx_i(t)}{dt} & = f_i(x(t), \tilde{\lambda}_2(t)), i = 1, 2. \end{cases}
\end{aligned} \tag{4.18}$$

In the infinite horizon problem, the Minimum Principle does not apply directly, due to the fact that there is no boundary condition on the adjoint variable ψ for the control problem as it is in the finite horizon problem. A detailed discussion on the infinite horizon problem in Aseev and Kryazhimskii [3] sheds light on a solution for the above question. First we discuss the existence of a solution to the problem.

Existence of the Optimal Solution

Before discussing the optimal solution of an infinite horizon control problem, we first need to check if there exists an optimal solution. Theorem 2.1 in [3] shows that if Condition (C1)-(C3) stated below are satisfied then an optimal solution exists.

(C1) There exists a $C_0 \geq 0$ such that for any $x \in \mathbb{R}^2, \tilde{\lambda}_2 \in \Lambda$,

$$\langle x, f(x, \tilde{\lambda}_2)\rangle \leq C_0(1 + \|x\|^2). \tag{4.19}$$

(C2) The control system in the optimization problem (4.18) is affine in control, i.e., it can be represented as

$$\frac{dx_i(t)}{dt} = \xi_i(x(t)) + \zeta_i(x(t))\tilde{\lambda}_2(t), i = 1, 2.$$

(C3) There exist positive functions μ and ω on $[0, \infty)$ such that $\mu(t) \to +0$ and $\omega(t) \to +0$ as $t \to \infty$ and, for every admissible pair $(x, \tilde{\lambda}_2)$, the following inequalities hold for any $t > 0$ and $T > 0$,

$$e^{-\rho t} \max_{\tilde{\lambda}_2 \in \Lambda} |g(x(t), \tilde{\lambda}_2)| \leq \mu(t), \tag{4.20}$$

$$\int_T^\infty e^{-\rho t} |g(x(t), \tilde{\lambda}_2(t))| dt \leq \omega(T), \tag{4.21}$$

We first state our result then check the conditions above in the proof.

Theorem 4.4. *There exists an optimal solution to Problem* (4.18)*.*

Proof. Now we check the above conditions (C1) -(C3). First we know that

$$\begin{aligned}\langle x, f(x, \tilde{\lambda}_2)\rangle &= \lambda_1 x_1 e^{-qx_2} + \tilde{\lambda}_2 x_1 e^{-qx_1} - \tilde{\lambda}_2 x_1 - \gamma x_1^2 \\ &\quad + \lambda_1 x_2 e^{-qx_2} + \tilde{\lambda}_2 x_2 e^{-qx_1} - \lambda_1 x_2 - \gamma x_2^2.\end{aligned}$$

Since $\lim_{x \to \infty} e^{-qx} = 0$ and $\lim_{x \to 0} e^{-qx} = 1$, then for any $x = (x_1, x_2) \in \mathbb{R}_+^2 \cup \{(0, 0)\}$, there exists C_1, such that

$$\lambda_1 x_1 e^{-qx_2} + \tilde{\lambda}_2 x_1 e^{-qx_1} \leq C_1 x_1^2,$$

$$\lambda_1 x_2 e^{-qx_2} + \tilde{\lambda}_2 x_2 e^{-qx_1} \leq C_1 x_2^2.$$

In addition, for any $x = (x_1, x_2) \in \mathbb{R}_+^2 \cup \{(0,0)\}$, there exists C_2, such that

$$-\tilde{\lambda}_2 x_1 - \gamma x_1^2 \leq C_2 x_1^2,$$

$$-\lambda_1 x_2 - \gamma x_2^2 \leq C_2 x_2^2.$$

Letting $C_0 = \max(C_1, C_2)$, Equation 4.19 in condition (C1) holds. Second, let

$$(\xi_1(x(t)), \xi_2(x(t))) = (\lambda_1 e^{-qx_2(t)} - \gamma x_1(t), \lambda_1 e^{-qx_2(t)} - \gamma x_2(t) - \lambda_1)$$

and

$$(\zeta_1(x(t)), \zeta_2(x(t))) = (e^{-qx_1(t)} - 1, e^{-qx_1(t)}).$$

Then we can write

$$\frac{dx_i(t)}{dt} = \xi_i(x(t)) + \zeta_i(x(t))\tilde{\lambda}_2(t), i = 1, 2,$$

and hence (C2) is satisfied. Third, using a similar method as in the proof of Lemma 3.5 in Chapter 3, we can show that $x(t)$ in the control problem (4.18) is bounded from both below and above. This together with the fact that $I(\cdot)$ satisfies the assumptions introduced in Section 4.2.1 implies that there exists a constant C such that for any $t \geq 0$,

$$\max|g(x(t), \tilde{\lambda}_2(t))| = \max|I(\tilde{\lambda}_2(t)) - a\gamma x_2(t)| \leq C.$$

Letting $\mu(t) = Ce^{-\rho t}$, Equation (4.20) in condition (C3) is satisfied. Further,

$$\begin{aligned}\int_T^\infty e^{-\rho t}|g(x(t), \tilde{\lambda}_2(t))|dt &\leq C\int_T^\infty e^{-\rho t}dt \\ &= -\frac{C}{\rho}(0 - e^{-\rho T}) \\ &= \frac{C}{\rho}e^{-\rho T}\end{aligned}$$

Letting $\omega(t) = \frac{C}{\rho}e^{-\rho T}$, the Equation (4.21) in Condition (C3) holds. Consequently, applying the result in Theorem 2.1. in Aseev and Kryazhimskii [3], there exists an optimal solution to the infinite horizon problem 4.18. □

Theorem 4.4 ensures that there exists an optimal solution for the infinite horizon control problem 4.18. Next we discuss the properties of an optimal solution.

Properties of an Optimal Solution

Now we discuss some important properties of an optimal solution. To facilitate our discussion, we give the definitions of the Hamiltonian function, the normalised fundamental matrix of the linear system of differential equations corresponding to the infinite horizon optimal control problem and its adjoint system as follows. For infinite optimal control problem (4.18) the Hamilton function $H_\infty : \mathbb{R}^2 \times [0,\infty) \times \Lambda \times \mathbb{R}^2$ and the Hamiltonian $H_\infty^* : \mathbb{R}^2 \times [0,\infty) \times \mathbb{R}^2$ are defined as:

$$H_\infty(x, t, \tilde{\lambda}_2, \psi) = e^{-\rho t}g(x, \tilde{\lambda}_2) + \psi^\intercal f(x, \tilde{\lambda}_2),$$

$$H_\infty^*(x, t, \psi) = \max_{\tilde{\lambda}_2 \in \Lambda} H_\infty(x, t, \tilde{\lambda}_2, \psi),$$

where ψ is the adjoint variable. For a given admission pair $(x, \tilde{\lambda}_2)$, *the normalised fundamental matrix of the linear system of differential equations*: $Y_{x,\tilde{\lambda}_2}(t) = \begin{pmatrix} y_{11}(t) & y_{12}(t) \\ y_{21}(t) & y_{22}(t) \end{pmatrix}$ for the whole infinite horizon $[0, \infty)$ is defined as

$$\begin{pmatrix} \frac{dy_{11}(t)}{dt} & \frac{dy_{12}(t)}{dt} \\ \frac{dy_{21}(t)}{dt} & \frac{dy_{22}(t)}{dt} \end{pmatrix} = \begin{pmatrix} \frac{df_1(x(t),\tilde{\lambda}_2(t))}{dx_1} & \frac{df_1(x(t),\tilde{\lambda}_2(t))}{dx_2} \\ \frac{df_2(x(t),\tilde{\lambda}_2(t))}{dx_1} & \frac{df_2(x(t),\tilde{\lambda}_2(t))}{dx_2} \end{pmatrix} \begin{pmatrix} y_{11}(t) & y_{12}(t) \\ y_{21}(t) & y_{22}(t) \end{pmatrix}.$$

The normalised fundamental matrix of the corresponding adjoint system

$$Z_{x,\lambda_2}(t) = \begin{pmatrix} z_{11}(t) & z_{12}(t) \\ z_{21}(t) & z_{22}(t) \end{pmatrix}$$

is defined as

$$\begin{pmatrix} \frac{dz_{11}(t)}{dt} & \frac{dz_{12}(t)}{dt} \\ \frac{dz_{21}(t)}{dt} & \frac{dz_{22}(t)}{dt} \end{pmatrix} = - \begin{pmatrix} \frac{df_1(x(t),\tilde{\lambda}_2(t))}{dx_1} & \frac{df_2(x(t),\tilde{\lambda}_2(t))}{dx_1} \\ \frac{df_1(x(t),\tilde{\lambda}_2(t))}{dx_2} & \frac{df_2(x(t),\tilde{\lambda}_2(t))}{dx_2} \end{pmatrix} \begin{pmatrix} z_{11}(t) & z_{12}(t) \\ z_{21}(t) & z_{22}(t) \end{pmatrix}.$$

Then for any $t \geq 0$, $Y^{\mathsf{T}}_{x,\tilde{\lambda}_2}(t) = Z^{-1}_{x,\tilde{\lambda}_2}(t)$.

Before we present the result of an optimal solution, we state three more conditions introduced in [3] in addition to Conditions (C1)-(C3) and show that all of them are satisfied in our system.

(C4) There exists constants κ_1 and κ_2 such that for any $x \in \mathbb{R}^2$ and $\tilde{\lambda}_2 \in \Lambda$,

$$\left\| \frac{\partial g(x, \tilde{\lambda}_2)}{\partial x} \right\| \leq \kappa_1 (1 + \|x\|^{\kappa_2}). \tag{4.22}$$

(C5) There exists numbers $k \in \mathbb{R}$, $k_1, k_2, k_3 \geq 0$ such that for any $t \geq 0$, any admissible pair $(x, \tilde{\lambda}_2)$ satisfies the conditions

$$\|x(t)\| \leq k_1 + k_2 e^{kt} \tag{4.23}$$

and

$$\|Y_{(x,\tilde{\lambda}_2)}(t)\| \leq k_3 e^{kt} \tag{4.24}$$

(C6) The following inequality holds

$$\rho > (\kappa_2 + 1)k. \tag{4.25}$$

Theorem 4.5. *Let $(x^*, \tilde{\lambda}_2^*)$ be an optimal admissible pair in problem 4.18. Then there exists an adjoint variable ψ such that the following conditions hold:*

(i) *the optimal admissible pair $(x^*, \tilde{\lambda}_2^*)$ together with the adjoint variable ψ satisfies the* *maximum principle in the normal form in the infinite interval* $[0, \infty)$*:*

$$\begin{cases} \frac{d\psi_1(t)}{dt} &= e^{-\rho t} a\gamma\psi_1(t) + (\psi_1(t) + \psi_2(t))\tilde{\lambda}_2^*(t) q e^{-qx_1^*(t)} \\ \frac{d\psi_2(t)}{dt} &= e^{-\rho t} a\gamma\psi_2(t) + (\psi_1(t) + \psi_2(t))\lambda_1 q e^{-qx_2^*(t)}, \end{cases}$$

$$H_\infty^*(x^*(t), t, \psi(t)) = H_\infty(x^*(t), t, \tilde{\lambda}_2^*(t), \psi(t));$$

(ii) *the optimal admissible pair* (x^*, λ_2^*) *together with the adjoint variable* ψ *satisfies the normal form stationarity condition for any* $t \geq 0$*:*

$$\begin{aligned} H_\infty^*(x^*(t), t, \psi(t)) &= \rho \int_t^\infty e^{-\rho s} g(x^*(s), \tilde{\lambda}_2^*(s)) ds \\ &= \rho \int_t^\infty e^{-\rho s} (I(\tilde{\lambda}_2^*(s)) - a\gamma x_2^*(s)) ds \end{aligned}$$

(iii) *for any* $t \geq 0$*, the integral* $\Upsilon^*(t) = \int_t^\infty e^{-\rho s} Y_{x^*, \tilde{\lambda}_2^*}^\intercal(s) \frac{\partial g(x^*(s), \tilde{\lambda}_2^*(s))}{\partial x} ds$ *converges absolutely (i.e.,* $\Upsilon^*(t) < \infty$*), where* $Y_{x^*, \tilde{\lambda}_2^*}(t)$ *satisfies that*

$$\begin{pmatrix} \frac{dy_{11}(t)}{dt} & \frac{dy_{12}(t)}{dt} \\ \frac{dy_{21}(t)}{dt} & \frac{dy_{22}(t)}{dt} \end{pmatrix} = - \begin{pmatrix} q\tilde{\lambda}_2^*(t)e^{-qx_1^*(t)} + \gamma & q\lambda_1 e^{-qx_2^*(t)} \\ q\tilde{\lambda}_2^*(t)e^{-qx_1^*(t)} & q\lambda_1 e^{-qx_2^*(t)} + \gamma \end{pmatrix} \begin{pmatrix} y_{11}(t) & y_{12}(t) \\ y_{21}(t) & y_{22}(t) \end{pmatrix}.$$

Moreover we have $\psi(t) = Z_{(x^*, \tilde{\lambda}_2^*)}(t)\Upsilon^*(t)$ *where* $Z_{(x^*, \tilde{\lambda}_2^*)}(t)$ *satisfies*

$$\begin{pmatrix} \frac{dz_{11}(t)}{dt} & \frac{dz_{12}(t)}{dt} \\ \frac{dz_{21}(t)}{dt} & \frac{dz_{22}(t)}{dt} \end{pmatrix} = \begin{pmatrix} q\tilde{\lambda}_2^*(t)e^{-qx_1^*(t)} + \gamma & q\tilde{\lambda}_2^*(t)e^{-qx_1^*(t)} \\ q\lambda_1 e^{-qx_2^*(t)} & q\lambda_1 e^{-qx_2^*(t)} + \gamma \end{pmatrix} \begin{pmatrix} z_{11}(t) & z_{12}(t) \\ z_{21}(t) & z_{22}(t) \end{pmatrix}.$$

Proof. If Conditions (C1)-(C6) hold, then the conclusion as desired follows from Theorem 12.1 in [3]. In the proof of Theorem 4.4, we have shown that Conditions (C1)-(C3) hold, here we only need to check (C4)-(C6). In our system,

$$\| \frac{\partial g(x, \tilde{\lambda}_2)}{\partial x} \| = \| (0, -a\gamma)^\intercal \|,$$

and hence for $\kappa_1 = |a\gamma|$ and $\kappa_2 = 0$, (C4) holds. Since $Y_{(x,\tilde{\lambda}_2)}(t)$ satisfies that

$$\begin{pmatrix} \frac{dy_{11}(t)}{dt} & \frac{dy_{12}(t)}{dt} \\ \frac{dy_{21}(t)}{dt} & \frac{dy_{22}(t)}{dt} \end{pmatrix} = \begin{pmatrix} -q\tilde{\lambda}_2(t)e^{-qx_1(t)} - \gamma & -q\lambda_1 e^{-qx_2(t)} \\ -q\tilde{\lambda}_2(t)e^{-qx_1(t)} & -q\lambda_1 e^{-qx_2(t)} - \gamma \end{pmatrix} \begin{pmatrix} y_{11}(t) & y_{12}(t) \\ y_{21}(t) & y_{22}(t) \end{pmatrix},$$

$\|Y_{(x,\tilde{\lambda}_2)}\|(t)$ is bounded, i.e., there exists a number k_3 such that $\|Y_{(x,\tilde{\lambda}_2)}\| \leq k_3$. This together with the fact that $x(t)$ is bounded implies that for $k = 0$, condition (C5) holds. Since $k = 0$ and $\kappa = 0$, $\rho > 0$ holds and so does condition (C6). Therefore the results as desired follow.

□

Although the infinite horizon optimization problem (4.5) is difficult to solve in general, Theorem 4.4 guarantees an optimal solution and Theorem 4.5 provides a way to compute the optimal solution for a specified income rate function. Next we move to the discussion for the observable model.

4.3 Observable Model

4.3.1 Model Description

This section addresses observable systems, where an employer observes the length of the queue of employers and employees in the system before making her decision on whether to join the system. Moreover, the information on the arrival rate of employees, the matching probability q are known to employers. Our goal is to develop a state dependent pricing policy for observable systems to maximize the long run average profit. As the system is observable, the strategy of employers plays an important role in the pricing. Hence we start our analysis by studying employers' strategies. For notational convenience we use the notation $r = 1 - q$ to denote the probability of not matching for a pair of employees and employers as defined in Chapter 2.

When an employer decides to join the system, she considers the chance of finding a match either by an immediate match upon arrival or a later match after spending sometime in the system with a later arriving employee. We assume that the cost of waiting in the system until finding a match cancels out the value of finding such a late match. Therefore employers decide on whether to join the system by evaluating the instant reward from joining the system. She looks at the number of employees in the system and evaluates the probability of finding an immediate match upon arrival. Suppose an immediate match upon arrival brings an employer a reward of $R \geq 0$, where R is a constant. Then when an arriving employer sees x_1 employees in the system, she computes her expected reward from joining in the system as

$$\mathbb{E}[\text{Reward}] = (1 - r^{x_1})R - p,$$

where p is the entry price for her to join in the system. This arriving employer joins the system and pays the system operator a fee p if and only if her expected reward is non-negative, i.e. $\mathbb{E}[\text{Reward}] \geq 0$. As a consequence, when there are x_1 employees in the system, the maximum entry price a system operator could gain is $p_{\max}(x_1) = (1-r^{x_1})R$. Therefore, at any time t when there are $X_1(t)$ employees in the system, if it is desirable to accept an arriving employer we charge her

$$p(t) = p_{\max}(X_1(t)) = (1 - r^{X_1(t)})R. \tag{4.26}$$

Since any price larger than $p_{\max}(X_1(t))$ will prevent an arriving employer from joining the system and hence result in a zero profit from her, for notational convenience, we denote $p(t) = \infty$ whenever it is better to reject an arriving employer. As a consequence, an optimal pricing strategy essentially tells when it is desirable for an arriving employer to be accepted or rejected.

We consider the long run average profit problem as defined in Equation (4.3). First we notice that the expected total profit generated in any time interval $[0, T]$ is the sum of the expected profit from each arriving employer during that interval. Defining $V(x_1, x_2)$ as the expected profit we gain from an arriving employer when there are x_1 employees and x_2 employers in the system, we have

$$V(x_1, x_2) = \begin{cases} (1 - r^{x_1})R - ar^{x_1}P(x_1, x_2)), & \text{if an arriving employer is accepted,} \\ 0, & \text{otherwise.} \end{cases} \tag{4.27}$$

where $P(x_1, x_2)$ is the probability that a newly joined employer will abandon the system when there are x_1 employees and x_2 employers in the system when she arrives, given that she cannot find a match upon arrival. Then the long run average profit we consider in an observable system is

$$J(\text{average}) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\int_0^T V(X_1(t), X_2(t))dA_2(t)\right] \tag{4.28}$$

Our aim is to find the optimal $V^*(X_1^*(t), X_2^*(t))$ with the corresponding queue length process $\{(X_1^*(t), X_2^*(t)\}$ to maximize Equation (4.28). As we see in Equation (4.27), to compute $V(x_1, x_2)$ we need to quantify the probability that an arriving employer abandons the system if she joins, which is very complicated due to the probabilistic matching structure. Therefore in Section 4.3.2 we analyse a myopic optimal pricing strategy which guarantees that the system operator receives the optimal amount of profit at each stage when an employer arrives.

4.3.2 Myopic Pricing

Due to the difficulty in completely quantifying the probability of abandonment $P(x_1, x_2)$ in Equation (4.27), a direct analysis on the optimal pricing with respect to the objective function (4.28) is very complicated. To provide a reasonable pricing scheme, we discuss the myopic pricing which assumes that the system operator wants to maximize the profit at each stage when an employer arrives, regardless of what will happen in the future. With this assumption, we should accept an employer to the system if the expected profit she brings to the system is positive and reject her otherwise. In particular, we propose bounds on the numbers of employers and employees in the system at which arriving employers should be accepted or rejected. The bounds proposed in this section are based on myopic pricing strategy adopted by system operators.

Assumption 4.6. *The system operator adopts a myopic pricing strategy when an employer arrives, regardless what will happen in the resulting future stages.*

Note that under Assumption 4.6, at any time t, the optimal control variable which is denoted as $V^{m*}(X_1^{m*}(t), X_2^{m*}(t))$ together with its corresponding queue length process

$\{(X_1^{m*}(t), X_2^{m*}(t))\}$ satisfies the following relationship:

$$V^{m*}(X_1^{m*}(t), X_2^{m*}(t)) = \max((1 - r^{X_1^{m*}(t)})R - ar^{X_1^{m*}(t)}P(X_1^{m*}(t), X_2^{m*}(t)), 0). \quad (4.29)$$

This essentially means that we accept an arriving employer if she brings a positive expected profit and reject her otherwise. Through determining pricing strategy, Theorem 4.7 provides bounds on the number of employees at which an arriving employer is accepted and rejected.

Theorem 4.7. *Under Assumption 4.6, the optimal myopic pricing $p^{m*}(t)$ at time t satisfies*

$$p^{m*}(t) = \begin{cases} p_{\max}(X_1^{m*}(t)), & \text{if } X_1^{m*}(t) \geq x_1^u = \log_r \frac{R}{R+a}, \\ \infty, & \text{if } X_1^{m*}(t) < x_1^l = \log_r \frac{R}{R+a\frac{\gamma}{\gamma+q\lambda_1}}, \end{cases}$$

where $p_{max}(X_1^{m*}(t))$ is defined as Equation(4.26).

Theorem 4.7 provides a pricing policy under the assumption of myopic pricing which tells us that regardless of the number of employers, when the number of employees reach above x_1^u, we should accept an arriving employer; when the number of employees drops below x_1^l, we should reject an arriving employer. The case when the number of employees falls in the interval of $[x_1^l, x_1^u]$ is discussed in Theorem 4.8.

Proof. By Equation (4.29) we know that if

$$(1 - r^{X_1^{m*}(t)})R - ar^{X_1^{m*}(t)}P(X_1^{m*}(t), X_2^{m*}(t)) \geq 0,$$

to accept the arriving employer brings a positive profit and hence we should set $p^{m*}(t) = p_{\max}(X_1^{m*}(t))$. Otherwise we should reject the arriving employer by setting the entry price larger than $p_{\max}(X_1^{m*}(t))$. Therefore we need to show the following two results: if $X_1^{m*}(t) \geq x_1^u$, then $(1 - r^{X_1^{m*}(t)})R - ar^{X_1^{m*}(t)}P(X_1^{m*}(t), X_2^{m*}(t)) > 0$; and if $X_1^{m*}(t) < x_1^l$, then$(1 - r^{X_1^{m*}(t)})R - ar^{X_1^{m*}(t)}P(X_1^{m*}(t), X_2^{m*}(t)) < 0$.

First if $X_1^{m*}(t) \geq x_1^u$, we have

$$(1 - r^{X_1^{m*}(t)})R - ar^{X_1^{m*}(t)} \geq 0.$$

Since $P(X_1(t), X_2(t)) \leq 1$ holds for any pricing strategy associated with its corresponding queue length process, we have, under the myopic optimal pricing strategy,

$$(1 - r^{X_1^{m*}(t)})R - ar^{X_1^{m*}(t)} P(X_1^{m*}(t), X_2^{m*}(t)) \geq (1 - r^{X_1^{m*}(t)})R - ar^{X_1^{m*}(t)} \geq 0.$$

Next we define E_1 to be the event that an employer arriving at time t finds a match after arrival if she joins the system when there are $X_1^{m*}(t)$ employees and $X_2^{m*}(t)$ employers in the system. Then $P(X_1^{m*}(t), X_2^{m*}(t)) = 1 - \mathbb{P}(E_1)$. If an employer who joins in the system does not get matched upon arrival, the only chance that she finds a match is to match with the later arriving employees who join the system before her patient time runs out. Conditioning on there are n employees arriving to the system before her patient time runs out, she need to be able to match at least one of them, which has a probability $1 - r^n$. Using the law of total probability, we have

$$\mathbb{P}(E_1) \leq \int_0^\infty \sum_{n=0}^{\infty} (1 - r^n) \frac{(\lambda_1 s)^n}{n!} e^{-\lambda_1 s} \gamma e^{-\gamma s} ds. \tag{4.30}$$

To simplify the right hand side of Inequality (4.30), we conduct the following computation:

$$\begin{aligned}\sum_{n=0}^{\infty} (1 - r^n) \frac{(\lambda_1 s)^n}{n!} e^{-\lambda_1 s} &= \sum_{n=0}^{\infty} \frac{(\lambda_1 s)^n}{n!} e^{-\lambda_1 s} - \sum_{n=0}^{\infty} \frac{(\lambda_1 s r)^n}{n!} e^{-\lambda_1 s} \\ &= 1 - \sum_{n=0}^{\infty} \frac{(\lambda_1 s r)^n}{n!} e^{-\lambda_1 r s} e^{\lambda_1 r s - \lambda_1 s} \\ &= 1 - e^{-\lambda_1 q s}.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\mathbb{P}(E_1) &\leq \int_0^\infty (1-e^{-\lambda_1 q s})\gamma e^{-\gamma s} ds \\ &= \int_0^\infty \gamma e^{-\gamma s} - \gamma e^{-\lambda_1 q s - \gamma s} ds \\ &= \frac{\lambda_1 q}{\lambda_1 q + \gamma},\end{aligned}$$

and hence $P(X_1^{m*}(t), X_2^{m*}(t)) \geq \frac{\gamma}{\lambda_1 q+\gamma}$. As a result, when $X_1^{m*}(t) < x_1^l$, the inequality $(1-r^{X_1^{m*}(t)})R - ar^{X_1^{m*}(t)} P(X_1^{m*}(t), X_2^{m*}(t)) = 1-(1-r^{X_1^{m*}(t)})R - ar^{X_1^{m*}(t)}(1-\mathbb{P}(E_1)) < 0$ holds and thus, the optimal strategy is to reject the arriving employer by setting $p^{m*}(t) = \infty$.

□

Next, to gain insight about what the optimal strategy should be if the number of employees lies in the interval between $[x_1^l, x_1^u]$, we indroduce Theorem 4.8. In particular, it takes the number of employers in to consideration.

Theorem 4.8. *Under Assumption 4.6, the optimal myopic pricing* $p^{m*}(t) = \infty$ *if* $X_1^{m*}(t) \in (x_1^l, x_1^u)$ *and* $X_2^{m*}(t) \geq x_2^u$*, where*

$$x_2^u = \begin{cases} \max(x \in \mathbb{N}^+ : 1 - r^{X_1^{m*}(t)} R - ar^{X_1^{m*}(t)} g(x) > 0), & \textit{if such } x \textit{ exists} \\ 0, & \textit{otherwise}, \end{cases}$$

$$g(x) = \left(1 + \frac{h(x)}{r}\right) \frac{\gamma}{\lambda_1 q + \gamma}, x \in \mathbb{N}^+$$

and

$$h(x) = \begin{cases} (1-r)^2 \frac{\lambda_1}{\lambda_1+\gamma}\frac{1}{2}, & x = 1 \\ (1-r)(\frac{\lambda_1}{\lambda_1+x\gamma}(1-\frac{1-r^{x+1}}{(x+1)q}) + \sum_{m=1}^{x-1}(\prod_{l=m+1}^{x-1}\frac{l}{\lambda_1+l\gamma})\frac{\lambda_1\gamma^{x-m}}{\lambda_1+m\gamma}(1-\frac{1-r^{m+1}}{(m+1)q})), & x \geq 2. \end{cases}$$

Proof. We prove the theorem by developing another bound on $P(X_1^{m*}(t), X_2^{m*}(t))$ based on $X_2^{m*}(t)$. First we define E_2 to be the event that the an employer matches with the first later arriving employee but does not get picked by her given she meets the the first later arriving employee before her patience time runs out and there are $x \geq 1$ other employers in the system when she joins in. Then if $x = 1$,

$$\mathbb{P}(E_2) = (1-r)\frac{\lambda_1}{\lambda_1+\gamma}(1-r)\frac{1}{2}.$$

If $x \geq 2$,

$$\begin{aligned}\mathbb{P}(E_2) = (1-r)&\frac{\lambda_1}{\lambda_1+x\gamma}\left(\sum_{k=1}^{x}\binom{x}{k}q^k r^{x-k}\frac{k}{k+1}\right)\\
&+(1-r)\left[\sum_{m=1}^{x-1}\frac{x\gamma}{\lambda_1+x\gamma}\cdots\frac{(m+1)\gamma}{\lambda_1+(m+1)\gamma}\frac{\lambda_1}{\lambda_1+m\gamma}\left(\sum_{k=1}^{m}\binom{m}{k}q^k r^{m-k}\frac{k}{k+1}\right)\right]\end{aligned}$$

To simplify the expression, we notice that,

$$\begin{aligned}\sum_{k=1}^{m}\binom{m}{k}q^k r^{m-k}\frac{k}{k+1} &= \sum_{k=1}^{m}\frac{m!}{k!\,(m-k)!}q^k(1-q)^{m-k}\frac{k}{k+1}\\
&= \frac{1}{m+1}\sum_{k=1}^{m}\binom{m+1}{k+1}q^k(1-q)^{m-k}k\\
\text{(letting } l=k+1\text{)} \qquad &= \frac{1}{(m+1)q}\sum_{l=2}^{m+1}\binom{m+1}{l}q^l(1-q)^{m+1-l}(l-1)\\
&= \frac{1}{(m+1)q}(\sum_{l=2}^{m+1}\binom{m+1}{l}q^l(1-q)^{m+1-l}l\\
&\quad -\sum_{l=2}^{m+1}\binom{m+1}{l}q^l(1-q)^{m+1-l})\\
&= 1-\frac{1-r^{m+1}}{(m+1)q}\end{aligned}$$

Similarly we have $\sum_{k=1}^{x}\binom{x}{k}q^k r^{x-k}\frac{k}{k+1} = 1-\frac{1-r^{x+1}}{(x+1)q}$, and hence $\mathbb{P}(E_2) = h(x)$ as defined in the theorem.

Suppose that an arriving employer joins the system when there are $X_2^{m*}(t)$ employers in the system and she does not get matched upon arrival, and we further assume that there are n employees arriving the system before she decides to abandon. Then the probability that this employer abandons the system is larger than the probability that she does not match with any of these n employees (which is r^n) plus the probability that she matches only with the first arriving employee but does not get picked (which is $r^{n-1}h(X_2^{m*}(t))$. That is, using the law of total probability,

$$P(X_1^{m*}, X_2^{m*}) \geq \int_0^\infty \sum_{n=0}^{\infty} (r^n + r^{n-1}h(X_2^{m*}(t)) \frac{(\lambda_1 s)^n}{n!} e^{-\lambda_1 s} \gamma e^{-\gamma s} ds.$$

To simplify this inequality, we first look at the total sum:

$$\begin{aligned}\sum_{n=0}^{\infty}(r^n + r^{n-1}h(x_2(t))\frac{(\lambda_1 s)^n}{n!}e^{-\lambda_1 s} &= \left(1 + \frac{h(X_2(t))}{1-q}\right)\sum_{n=0}^{\infty}\frac{(r\lambda_1 s)^n}{n!}e^{\lambda_1 r s}e^{\lambda_1 r s - \lambda_1 s} \\ &= \left(1 + \frac{h(X_2(t))}{1-q}\right)e^{-\lambda_1 q s}.\end{aligned}$$

Therefore we have

$$\begin{aligned} P(X_1^{m*}, X_2^{m*}) &\geq \int_0^\infty \left(1 + \frac{h(X_2^{m*})}{1-q}\right) e^{-\lambda_1 q s} \gamma e^{-\gamma s} ds \\ &= \left(1 + \frac{h(X_2^{m*})}{1-q}\right) \frac{\gamma}{\lambda_1 q + \gamma} \\ &= g(X_2^{m*}). \end{aligned}$$

As a consequence, the expected revenue of the arriving employer if she is accepted to the system can be bounded as

$$\begin{aligned} &(1 - r^{X_1^{m*}(t)})R - ar^{X_1^{m*}(t)} P(X_1^{m*}(t), X_2^{m*}(t)) \\ \leq\ &(1 - r^{X_1^{m*}(t)})R - ar^{X_1^{m*}(t)} g(X_2^{m*}(t)). \end{aligned}$$

Since $h(x)$ increases as x increases and so does $g(x)$, for a fixed $X_1^{m*}(t)$, if there exists an $\tilde{x}_2$, such that $\tilde{x}_2 = \max(x \in \mathbb{N}, (1-r^{X_1^{m*}(t)})R - ar^{X_1^{m*}(t)} g(X_2^{m*}(t)) \geq 0)$, then for any $X_2^{m*}(t) \geq \tilde{x}_2$, $(1-r^{X_1^{m*}(t)})R - ar^{X_1^{m*}(t)} P(X_1^{m*}(t), X_2^{m*}(t)) \leq 0$, the optimal strategy of the system controller is to reject the arriving employer and so the optimal entry price $p^{m*}(t) = \infty$; otherwise, if no such $\tilde{x}_2$ exists for some $X_1^{m*}(t)$, then for all $X_2^{m*}(t) \geq 0$, the optimal entry price $p^{m*}(t) = \infty$.

□

We learn that under the assumption of myopic pricing, when the number of employees reaches a constant upper bound, we should always accept an arriving employer by charging her the maximum price based on the number of employees in the system. On the other hand, when the number of employees drops below a constant lower bound, we should reject an arriving employer by setting the entry price to be higher than what she would like to afford. In addition, if the number of employees in the system is between the upper and lower bounds, then we turn to the number of employers to adjust the pricing strategy. In this case, if the number of employers is above a constant upper bound, then we should reject the arriving employer.

In this section we discuss myopic pricing mechanism which provides a way to maximize the profit at each stage. It guarantees that at each stage, we receive the best possible expected profit. However, to check if the optimal myopic pricing is the optimal over all pricing strategies, we have to verify Equation (4.29). In particular, we need to show Equation (4.29) by justifying the following two claims:

(a) $V^*(X_1^*(t), X_2^*(t)) \geq 0$.

(b) Whenever $(1 - r^{X_1^{m*}(t)})R - ar^{X_1^{m*}(t)}P(X_1^{m*}(t), X_2^{m*}(t)) > 0$, $V^*(X_1^*(t), X_2^*(t)) > 0$.

To see (a), first we can show that $V(x_1, x_2)$ increased in x_1 and decreases in x_2. Then if a pricing policy π accepts an employer who brings in a negative profit at some time t_0, then this employer results in either an increase in the number of employers or a decrease in the number of employees. We can define another policy $\tilde{\pi}$ which follows the same acceptance or rejection of employers as π at any time expect from t_0. Then before t_0, π and $\tilde{\pi}$ generate the same expected total profit but at t_0, $\tilde{\pi}$ receives a zero profit but π receives a negative profit. Since $\tilde{\pi}$ rejects the employer arriving at t_0, there will be at least (at most) as many employees (employers) as in the scenario under π for any later arriving employer. This guarantees that after t_0, $\tilde{\pi}$ generates at least as much profit as π. Therefore the total profit generate by π is worse then $\tilde{\pi}$. As a result, in the optimal pricing mechanism there is no stage when we accept an employer who brings in a negative profit.

However it is not so straightforward to check (b), which essentially says that whenever an arriving employer brings in a positive profit it is accepted by the optimal mechanism. We know that accepting an employer might lead to a decrease of the number of employees, which might further results in a decrease of price we can set for the next arriving employers and the expected profit we gain from her. In addition we need to know how the probability of abandonment $P(x_1, x_2)$ is affected by a decrease in x_1 and an increase in x_2, before we can draw our conclusion that the optimal myopic pricing is the optimal pricing. We leave this to be an open question for further research.

4.4 Conclusion

In this chapter we discuss the optimal pricing for probabilistic matching systems with the objective of profit maximization. We formulate the optimal control problems for both the unobservable and observable models. For the unobservable model, we consider the expected discounted total profit. We first introduce general economic assumptions on the income rate function and then suggest a deterministic control model based on fluid limits discussed in Chapter 3. In particular we provide an explicit solution for the finite time horizon problem for a general arrival rate function and present an example for a specific definition of arrival rate function. For the infinite time horizon problem, we show the existence of an optimal solution to determine some properties of an optimal solution.

For the observable model, we study the long run average profit maximization problem. We first analyse employers' strategies and derive the maximum price the system can gain from an arriving employer depending on the number of employees in the system. By charging exactly this maximum price we are accepting an arriving employer, while charging a price higher than the maximum price we are rejecting an arriving employer. Therefore we suggest that the pricing problem is equivalent to finding the optimal thresholds at which an employer is accepted or rejected. We develop the optimal pricing mechanism under the assumption of myopic pricing, which maximizes the profit at each stage regardless of future stages. In particular, we provide an upper bound and a lower bound on the number of employees at which an arriving employer should be rejected and accepted respectively. Furthermore, we show that when the number of employees stays between the lower and upper bounds and when the number of employers reaches an upper bound, any arriving employer should be rejected. We identify under which conditions the optimal myopic pricing is the optimal over all pricing strategies and we leave the justification of the conditions as an open question for a further research direction.

There are many interesting directions to carry on for further research on the topic of pricing for probabilistic matching systems. One might want to analyse the difference between the long run average profits generated by the unobservable and observable models. This will shed light on whether it is beneficial for the system operator to reveal system information to arriving employers. Another interesting direction is to study social welfare maximization for an observable model and compare the optimal pricing with the established results of profit maximization.

Chapter 5

Conclusion and Future Work

In this work, we propose a novel queueing model to analyse customer behaviours in probabilistic matching systems, where unlike the traditional queueing systems where users wait to access a resource, two classes of users arrive and wait to match with users from the other class. This model is motivated by internet portals, e.g. employment and rental portals, matrimonial and dating sites and multi-purpose portals. As an important feature, the users match probabilistically and this type of matching captures the nature of individuality in these portals, e.g. an employer hires an employee with some probability even when the employee has the necessary background.

We start our analysis by studying the transient behaviour of this new model and derive the basic probability distribution of the number of matches given the number of arrivals, e.g. the probability of exactly k people finding a job, when there are m employers hiring and k employees looking for jobs. After that, we prove that the system is unstable for any set of parameters if it is not controlled and characterize the nature of instability. We then suggest four different classes of admission control policies and prove that one type of the admission control stabilizes the system when matching probability $q = 0$ and other three controls stabilize the probabilistic matching systems for any system parameters. In addition we conduct performance analysis on the admission controls. Contrary to our intuition that the throughput should increase as the users get more likely to match, we show that under specific control policies throughput is insensitive to the matching probability. We also show similar insensitivity results relating to the average queue lengths and waiting times. Even more surprisingly, we show that under many control policies the throughput may decrease as the matching probability increases. By explaining the reasons behind this unexpected behaviour, we argue that these policies can be deemed to be more reasonable than the policies where the throughput is an increasing function of the matching probability.

Furthermore, we propose approximation methods based on fluid and diffusion limits using different scalings to characterize further features of probabilistic matching systems. We propose two different scalings to obtain fluid and diffusion approximations to the queue length processes of probabilistic matching systems. We suggest that the first approach of scaling space and time while the matching probability is kept fixed is suitable for those systems when the matching probability is considerably high. Then we introduce the second approach to study the systems with small matching probability and we incorporate the phenomenon of user abandonments in the model. Through the analysis of the fluid limits, we show that the difference between the average queue lengths of different classes of users is independent of the matching probability. This result is similar to that under some admission controls as shown by exact analysis. Based on the fluid limits obtained, we study the asymptotic behaviour and provide insights on the long run average numbers. First, when abandonment rate is zero, the two fluid limits, obtained with and without scaling the matching probability, converges to each other with time. Further, for non-zero abandonment rates, the fluid limits obtained in the second scaling converge to a unique fixed point, which represents the long run average number of users in the system. Carrying on analysis on the fixed point reveals that as the abandonment rate increases, we show that the number of users for the class with lower arrival rate first experiences an increase and then decrease while the number of users for the class with higher arrival rate decreases monotonically.

Finally we study optimal pricing and revenue management for probabilistic matching systems with the objective of profit maximization. We formulate mathematical models for both unobservable and observable models respectively. For the unobservable model, we impose general economic assumptions on the income rate function and suggest to change the control variable of pricing to the arrival rate. We suggest a deterministic optimal control for finite and infinite time horizontal problems respectively based on the fluid approximation of the queue length process. For the observable model, we analyse users' strategies and identify the maximum state dependent price the system can gain from an arriving user. Whenever it is profitable to accept an arriving employer we charge her exactly the maximum price, but more than the maximum price while otherwise. As a result, the optimal price essentially suggests that when an arriving user should be accepted/rejected. We develop an optimal pricing mechanism under the assumption of myopic pricing, which maximizes the profit at each stage. We show that under the assumption of myopic pricing, when the number of employees goes above (drops below) an upper (lower) bound, it is profitable to accept (reject) an arriving employer. Further when the number of employees lies between the upper and lower bounds, an arriving employer should be rejected if the number of employers reaches an upper bound. We identify the conditions under which the optimal myopic pricing policy is the overall optimal pricing policy and we leave the justification of these conditions to be an open question for further research.

Probabilistic matching systems exhibit many interesting properties and we provide a few promising future directions as follows. One interesting research direction is to consider probabilistic matching networks, where each class have several types of users and each pair of types has a different matching probability. In the network setting we can focus on matching strategies, in addition to devising admission control policies. Moreover we can consider fluid and diffusion limits for probabilistic matching networks and to develop a pricing mechanism with the objective of social farewell maximization.

Appendix A

Proof of Theorem 2.1

Proof of Theorem 1. When $q = 1$, the result is trivial. When $0 < q < 1$, the probability of no matchings, i.e., $k = 0$, when there are m and n class-1 and class-2 users, respectively is r^{mn}. Also, we know that $k \leq m \leq n$ and $P^q_{k,m,n} = 0$ for $k > m$. When $1 \leq k \leq m$, conditioning on whether a specific class-1 user matches with any of the n class-2 users or not, we get

$$P^q_{k,m,n} = r^n P^q_{k,m-1,n} + (1 - r^n) P^q_{k-1,m-1,n-1}. \tag{A.1}$$

It is clear that given $P^q_{0,m,n} = r^{mn}$ and $P^q_{k,m,n} = 0$ when $k > m$, the solution to (A.1) is unique. We now use induction to prove that for $m \geq 1$

$$P^q_{1,m,n} = (1 - r^n) r^{(m-1)n} \frac{1 - r^m}{r^{m-1}(1 - r)}. \tag{A.2}$$

For $m = 1$, $P^q_{1,1,n} = 1 - r^n$. Now, assume that (A.2) holds for $P^q_{k,m-1,n}$ where $2 \leq m \leq n$. Then,

$$\begin{aligned}
P^q_{1,m,n} &= r^n P^q_{k,m-1,n} + (1 - r^n) P^q_{0,m-1,n-1} \\
&= r^n (1 - r^n) r^{(m-2)n} \frac{1 - r^{m-1}}{r^{m-2}(1 - r)} + (1 - r^n) r^{(m-1)(n-1)} \\
&= (1 - r^n) r^{(m-1)n} \frac{1 - r^m}{r^{m-1}(1 - r)},
\end{aligned}$$

and hence (A.2) holds for any $1 \leq m \leq n$.

Now, we show that solving the three-dimensional recursion (A.1) can be reduced to solving a two-dimensional recursion. Suppose, $\{a_{k,m}, k \geq 0, m \geq 0\}$ solves

$$a_{k,m} = \begin{cases} 1 & k = 0, \\ r^{k-m} a_{k-1,m-1} + a_{k,m-1} & 1 \leq k \leq m, \\ 0 & k > m. \end{cases} \tag{A.3}$$

Then,

$$P^q_{k,m,n} = a_{k,m} r^{(m-k)n} \prod_{i=0}^{k-1} (1 - r^{n-i}) \tag{A.4}$$

solves (A.1), where if $k = 0$, $\prod_{i=0}^{-1}(1-r^{n-i})$ is assumed to be 1. To prove this statement, first observe that (A.4) implies, $P^q_{0,m,n} = r^{mn}$. When $k = 1$, $a_{1,0} = 0, a_{1,1} = 1$ and $a_{1,m} = r^{-m+1} + a_{1,m-1}$ when $m \geq 2$, which implies

$$a_{1,m} = \sum_{i=0}^{m-1} \left(\frac{1}{r}\right)^i = \frac{r^m - 1}{r^{m-1}(r-1)}.$$

Now, fix k, m and n such that $n \geq m \geq 2$ and $k \geq 1$. Suppose that for $0 \leq n' < n$, $0 \leq m' < n'$, $0 \leq k' \leq m'$, $0 \leq m'' < m$, $k'' \geq 0$ and $0 \leq k''' \leq k$, $P^q_{k',m',n'}$, $P^q_{k'',m'',n}$ and $P^q_{k''',m,n}$ given as (A.4) coincides with the solution of (A.1). Then, if $k + 1 \leq m$,

$$\begin{aligned} P^q_{k+1,m,n} &= r^n P^q_{k+1,m-1,n} + (1 - r^n) P^q_{k,m-1,n-1} \\ &= a_{k,m} r^{(m-k-1)n} \prod_{i=0}^{k} (1 - r^{n-i}) + a_{k,m-1} r^{(m-k-1)(n-1)} \prod_{i=0}^{k} (1 - r^{n-i}) \\ &= (a_{k+1,m-1} + r^{-m+k+1} a_{k,m-1}) r^{(m-k-1)n} \prod_{i=0}^{k} (1 - r^{n-i}) \\ &= a_{k+1,m} r^{(m-k-1)n} \prod_{i=0}^{k} (1 - r^{n-i}). \end{aligned}$$

This proves that if we can solve (A.3), (A.4) provides us with the solution of (A.1). Now, we provide a solution to the recursion (A.3). Using (A.3) $m - k + 1$ times, we get

$$\begin{aligned} a_{k,m} &= r^{-m+k} a_{k-1,m-1} + r^{-m+k+1} a_{k-1,m-2} + \ldots + \underbrace{a_{k-1,k-1}}_{1} + \underbrace{a_{k,k-1}}_{0} \\ &= \sum_{j=0}^{m-k} r^{-m+k+j} \cdot a_{k-1,m-1-j}. \end{aligned} \tag{A.5}$$

We now guess that for $0 \le k \le m$, $a_{k,m}$ has the following form

$$a_{k,m} = \sum_{i=0}^{k} r^{-mi} \alpha_{k,i}. \tag{A.6}$$

Then, $a_{0,m} = 1$ implies $\alpha_{0,0} = 1$. For $k \ge 1$, we plug (A.6) into both sides of (A.5) and we obtain:

$$\begin{aligned}
\sum_{i=0}^{k} r^{-mi} \alpha_{k,i} &= \sum_{j=0}^{m-k} r^{-m+k+j} \sum_{i=0}^{k-1} r^{-(m-1-j)i} \alpha_{k-1,i} \\
&= \sum_{i=0}^{k-1} \alpha_{k-1,i} r^{-m+k-mi+i} \sum_{j=0}^{m-k} \left(r^{(i+1)}\right)^{j} \\
&= \sum_{i=0}^{k-1} \alpha_{k-1,i} \frac{1 - r^{(i+1)(m-k+1)}}{1 - r^{(i+1)}} r^{-m+k-mi+i} \\
&= \sum_{i=0}^{k-1} \frac{r^{-m+k-mi+i} - r^{-ki+2i+1}}{1 - r^{i+1}} \alpha_{k-1,i} \\
&= \sum_{i=0}^{k-1} r^{-m(i+1)} \frac{r^{k+i}}{1 - r^{i+1}} \alpha_{k-1,i} + \sum_{i=0}^{k-1} \frac{-r^{-ki+2i+1}}{1 - r^{i+1}} \alpha_{k-1,i}.
\end{aligned}$$

We then shift the index of the first sum on the righthand side, and for $k \ge 1$, we get

$$\sum_{i=0}^{k} r^{-mi} \alpha_{k,i} = \sum_{i=1}^{k} r^{-mi} \frac{r^{k+i-1}}{1 - r^{i}} \alpha_{k-1,i-1} + \sum_{i=0}^{k-1} \frac{-r^{-ki+2i+1}}{1 - r^{i+1}} \alpha_{k-1,i}.$$

Now, comparing the coefficient of r^{-mi} for $0 \le i \le k$, we obtain,

$$\alpha_{k,0} = \sum_{i=0}^{k-1} \frac{-r^{-ki+2i+1}}{1 - r^{i+1}} \alpha_{k-1,i}, \tag{A.7}$$

and for every $1 \le i \le k$:

$$\alpha_{k,i} = \frac{r^{k+i-1}}{1 - r^{i}} \alpha_{k-1,i-1}. \tag{A.8}$$

Repeating (A.8) i times we have for $1 \le i \le k$,

$$\begin{aligned}\alpha_{k,i} &= \frac{r^{k+i-1}}{(1-r^i)}\frac{r^{k+i-3}}{(1-r^{i-1})}\frac{r^{k+i-5}}{(1-r^{i-2})}\cdots\frac{r^{k-i+5}}{(1-r^3)}\frac{r^{k-i+3}}{(1-r^2)}\frac{r^{k-i+1}}{(1-r^1)}\alpha_{k-i,0}\\ &= \frac{r^{ki}}{\prod_{j=1}^{i}(1-r^j)}\alpha_{k-i,0}. \qquad (A.9)\end{aligned}$$

Plugging (A.9) into (A.7) and substituting $\beta_k = \alpha_{k,0}$, we obtain a new recurrence:

$$\beta_k = \sum_{i=1}^{k-1}\frac{-r^{-ki+2i+1}}{(1-r^{i+1})}\frac{r^{(k-1)i}}{\prod_{j=1}^{i}(1-r^j)}\beta_{k-1-i} + \frac{-r\beta_{k-1}}{1-r} = \sum_{i=0}^{k-1}\frac{-r^{i+1}\beta_{k-1-i}}{\prod_{j=1}^{i+1}(1-r^j)},$$

and replacing $l = k-1-i$, we get

$$\beta_k = \sum_{l=0}^{k-1}\frac{-r^{k-l}\beta_l}{\prod_{j=1}^{k-l}(1-r^j)}.$$

Thus by now we have a recurrence of the form: $\beta_0 = 1$ and for every $k \ge 1$:

$$\beta_k = \sum_{l=0}^{k-1}\gamma_{k,l}\beta_l, \qquad (A.10)$$

where

$$\gamma_{i,j} = \frac{y_i}{y_j}z_{i-j}, \qquad (A.11)$$

with $y_i = r^i$ and $z_i = \frac{-1}{\prod_{j=1}^{i}(1-r^j)}$. The recursion (A.10) is a 1-dimensional recurrence, and for $k \ge 1$ has a general solution :

$$\begin{aligned}\beta_k &= \sum_{l>0}\left(\sum_{0=i_0<i_1<\ldots<i_{l-1}<i_l=k}\left(\prod_{j=1}^{l}\gamma_{i_j,i_{j-1}}\right)\right)\beta_0\\ &= \sum_{l>0}\left(\sum_{0=i_0<i_1<\ldots<i_{l-1}<i_l=k}\gamma_{i_l,i_{l-1}}\gamma_{i_{l-1},i_{l-2}}\cdots\gamma_{i_2,i_1}\gamma_{i_1,i_0}\right). \qquad (A.12)\end{aligned}$$

We prove this by induction.

(i) when $k = 1$, Equation (A.12) implies

$$\beta_1 = \sum_{l>0}\left(\sum_{0=i_0<i_1<\ldots<i_{l-1}<i_l=1}\left(\prod_{j=1}^{l}\gamma_{i_j,i_{j-1}}\right)\right)\beta_0 = \gamma_{1,0},$$

which satisfies Equation (A.10). Hence Equation (A.12) holds for β_1.

(ii) suppose Equation (A.12) holds for all β_n, $1 \leq n \leq k-1$, then according to Equation (A.10),

$$\begin{aligned}\beta_k &= \sum_{n=0}^{k-1} \gamma_{k,n}\beta_n \\ &= \sum_{n=0}^{k-1} \gamma_{k,n} \left(\sum_{l>0} \left(\sum_{0=i_0<i_1<\ldots<i_{l-1}<i_l=n} \left(\prod_{j=1}^{l} \gamma_{i_j,i_{j-1}} \right) \right) \right) \beta_0 \\ &= \gamma_{k,0} + \gamma_{k,1}\gamma_{1,0} + \cdots + \sum_{l>0} \left(\sum_{0=i_0<i_1<\ldots<i_{l-1}<i_l=n} \gamma_{k,n}\gamma_{n,i_{l-1}} \cdots \gamma_{i_1,0} \right) \\ &\quad + \cdots + \sum_{l>0} \left(\sum_{0=i_0<i_1<\ldots<i_{l-1}<i_l=k-1} \gamma_{k,k-1}\gamma_{k-1,i_{l-1}}\gamma_{i_{l-1},i_{l-2}} \cdots \gamma_{i_1,0} \right) \\ &= \sum_{l>0} \left(\sum_{0=i_0<i_1<\ldots<i_{l-1}<i_l=k} \left(\prod_{j=1}^{l} \gamma_{i_j,i_{j-1}} \right) \right).\end{aligned}$$

Therefore, if Equation (A.12) holds for all β_n, $1 \leq n \leq k-1$, then it also holds for $\beta_n, n = k$.

Now by adding equation (A.11) to (A.12), we have:

$$\beta_k = \sum_{l>0} \left(\sum_{0=i_0<i_1<\ldots<i_{l-1}<i_l=k} \frac{y_k}{y_0} \prod_{j=1}^{l} z_{i_j - i_{j-1}}. \right).$$

Further more, using substitution $d_j = i_j - i_{j-1} \geq 1$, where $d_1 + d_2 + \ldots + d_l = k$ we have:

$$\beta_k = \sum_{l>0} \left(\sum_{d_1+d_2+\ldots+d_l=k} \frac{y_k}{y_0} \prod_{i=1}^{l} z_{d_i} \right),$$

where indexes $d_1, d_2, \ldots$ are taken from $\mathbb{N} = \{1, 2, 3, \ldots\}$. Thus, for $k \geq 1$,

$$\beta_k = \sum_{l>0} \left(\sum_{d_1+d_2+\ldots+d_l=k} r^k \prod_{i=1}^{l} \frac{-1}{\prod_{j=1}^{d_i}(1-r^j)} \right). \tag{A.13}$$

Finally (A.9) implies for $1 \leq i \leq k$,

$$\alpha_{k,i} = \frac{r^{ki}}{\prod_{j=1}^{i}(1-r^j)} \beta_{k-i}.$$

Using (A.6), we have

$$\begin{aligned} a_{k,m} &= \sum_{i=0}^{k} r^{-mi}\alpha_{k,i} = (\sum_{i=0}^{k-1} r^{-mi}\alpha_{k,i}) + r^{-mk}\alpha_{k,k} \\ &= \alpha_{k,0} + (\sum_{i=1}^{k-1} r^{-mi}\alpha_{k,i}) + r^{-mk}\alpha_{k,k}. \end{aligned}$$

As a result,

$$\begin{aligned} a_{k,m} = &\sum_{l>0}\left(\sum_{d_1+d_2+...+d_l=k} r^k \frac{(-1)^l}{\prod_{i=1}^{l}\prod_{j=1}^{d_i}(1-r^j)}\right) + r^{-mk}\frac{r^{k^2}}{\prod_{j=1}^{k}(1-r^j)} \\ &+ \sum_{i=1}^{k-1} r^{-mi}\frac{r^{ki}}{\prod_{j=1}^{i}(1-r^j)}\sum_{l>0}\left(\sum_{d_1+d_2+...+d_l=k-i} r^{k-i}\frac{(-1)^l}{\prod_{w=1}^{l}\prod_{j=1}^{d_w}(1-r^j)}\right), \end{aligned}$$

with indexes $d_1, d_2, ...$ taken from $\mathbb{N}_+ = \{1, 2, 3, ...\}$. □

Appendix B

Stationary Probabilities under the ASQ Policy.

Proof of Theorem 2.11. Under the ASQ policy, the rate balance equations can be written as follows:

$$(\lambda_1 + \lambda_2)p_{0,0} = \lambda_1(1-r)p_{0,1} + \lambda_2(1-r)p_{1,0}, \tag{B.1}$$

$$(\lambda_1 + \lambda_2)p_{i,i} = \lambda_1 r^i p_{i-1,i} + \lambda_2 r^i p_{i,i-1} + \lambda_1(1-r^{i+1})p_{i,i+1} + \lambda_2(1-r^{i+1})p_{i+1,i},\ \ i \geq 1, \tag{B.2}$$

$$\lambda_2 p_{i+1,i} = \lambda_1 r^i p_{i,i} + \lambda_1(1-r^{i+1})p_{i+1,i+1},\ \ i \geq 0, \tag{B.3}$$

$$\lambda_1 p_{i,i+1} = \lambda_2 r^i p_{i,i} + \lambda_2(1-r^{i+1})p_{i+1,i+1},\ \ i \geq 0, \tag{B.4}$$

$$\sum_{i=0}^{\infty}\sum_{j=i-1}^{i+1} p_{i,j} = 1. \tag{B.5}$$

The state space has a very special structure where the removal of a state in the form $(i,i), i > 0$ disconnects the transition graph. This implies a rate balance for the transitions between states $\{(i,i-1),(i-1,i)\}$ and (i,i), which implies the following detailed balance type equations:

$$(\lambda_1 + \lambda_2)r^i p_{i,i} = \lambda_1(1-r^{i+1})p_{i,i+1} + \lambda_2(1-r^{i+1})p_{i+1,i},\ \ i \geq 0, \tag{B.6}$$

$$(\lambda_1 + \lambda_2)(1-r^i)p_{i,i} = \lambda_1 r^i p_{i-1,i} + \lambda_2 r^i p_{i,i-1},\ \ i \geq 1, \tag{B.7}$$

Equation (B.1) can obtained by setting $i = 0$ in (B.6) and further, summing up (B.6) and (B.7) for a given $i \geq 1$ we have (B.2). Hence, any solution to the set of equations (B.3)-(B.7) also solves (B.1)-(B.5) and hence should be unique. Further, Equations (B.3) and (B.4) imply $\frac{p_{i+1,i}}{p_{i,i+1}} = \frac{\lambda_1^2}{\lambda_2^2}$. Hence, substituting $p_{i+1,i} = \frac{\lambda_1^2}{\lambda_2^2} p_{i,i+1}$ into (B.6) we obtain

$$p_{i,i+1} = \frac{\lambda_2}{\lambda_1} \frac{1 - r^{i+1}}{r^i} p_{i,i} \text{ and } p_{i+1,i} = \frac{\lambda_1}{\lambda_2} \frac{1 - r^{i+1}}{r^i} p_{i,i}. \tag{B.8}$$

Then (B.8) and (B.7) together imply $p_{i+1,i+1} = \frac{r^i r^{i+1}}{(1-r^{i+1})^2} p_{i,i}$, $i \geq 0$, and hence, for $i \geq 1$,

$$p_{i,i} = \prod_{k=1}^{i} \frac{r^k r^{k-1}}{(1 - r^k)^2} p_{0,0} = \frac{r^{i^2} p_{0,0}}{\left[\prod_{k=1}^{i} (1 - r^k)\right]^2}. \tag{B.9}$$

Substituting (B.9) into (B.8) and defining $\prod_{k=1}^{0} (1 - r^k) = 1$, for $i \geq 0$

$$\begin{aligned} p_{i,i+1} &= \frac{\lambda_2 r^{i(i+1)} p_{0,0}}{\lambda_1 \prod_{k=1}^{i} (1 - r^k) \prod_{k=1}^{i+1} (1 - r^k)} \\ p_{i+1,i} &= \frac{\lambda_1 r^{i(i+1)} p_{0,0}}{\lambda_2 \prod_{k=1}^{i} (1 - r^k) \prod_{k=1}^{i+1} (1 - r^k)}. \end{aligned}$$

Finally the result follows from plugging all $p_{i,j}$ back in (B.5). □

Appendix C

Insensitivities of Functional Threshold Policies with $h(x) = x + d$

Lemma C.1. *Suppose that the functional threshold policy with a threshold function of the form $h(x) = x + d$, where $d \in \mathbb{N}$ is employed to stabilize a probabiltic matching system. For $(i,j) \in \mathbb{N}^2$, let $p_{i,j}$ be the stationary probability of being at state (i,j). Now, define $a_l = \sum_{i=0}^{\infty} p_{i,i+l}$ and $a_{-l} = \sum_{j=0}^{\infty} p_{j+l,j}$ Then, if $\lambda_1 \neq \lambda_2$,*

$$a_{d+1} = \frac{1 - \frac{\lambda_2}{\lambda_1}}{1 - (\frac{\lambda_2}{\lambda_1})^{2d+3}} \text{ and } a_l = (\frac{\lambda_2}{\lambda_1})^{d+1-l} a_{d+1}, \text{ for } \quad -d-1 \leq l \leq d,$$

and if $\lambda_1 = \lambda_2$, $a_l = \frac{1}{2d+3}$, $-d-1 \leq l \leq d+1$.

Proof. When $q = 1$, the process $\{X^{1,FT}(t) + d + 1, t \geq 0\}$ is stochastically equivalent to an $M/M/1/2d+2$ system and the result follows. When $0 < q < 1$, the state space can be written as $\mathbb{S} = \{(i, i+l) : i \in \mathbb{N}, -d-1 \leq l \leq d+1, i+l \in \mathbb{N}\}$, hence the global balance equations are

$$(\lambda_1+\lambda_2)p_{0,0} = \lambda_1(1-r)p_{0,1} + \lambda_2(1-r)p_{1,0}, \tag{C.1}$$

$$(\lambda_1+\lambda_2)p_{i,i} = \lambda_1 r^i p_{i-1,i} + \lambda_2 r^i p_{i,i-1} + \lambda_1(1-r^{i+1})p_{i,i+1} + \lambda_2(1-r^{i+1})p_{i+1,i},\ i \geq 1, \tag{C.2}$$

$$(\lambda_1+\lambda_2)p_{l,0} = \lambda_1 p_{l-1,0} + \lambda_1(1-r)p_{l+1,1} + \lambda_2(1-r^{l+1})p_{l+1,0}, \quad 1 \leq l \leq d \tag{C.3}$$

$$(\lambda_1+\lambda_2)p_{i+l,i} = \lambda_1 r^i p_{i+l-1,i} + \lambda_1(1-r^{i+1})p_{i+l,i+1} + \lambda_2 r^{i+l} p_{i+l,i-1} + \lambda_2(1-r^{i+l+1})p_{i+l+1,i},\ i \geq 1, 1 \leq l \leq d, \tag{C.4}$$

$$(\lambda_1+\lambda_2)p_{0,l} = \lambda_2 p_{0,l-1} + \lambda_2(1-r)p_{1,l+1} + \lambda_1(1-r^{l+1})p_{0,l+1}, \quad 1 \leq l \leq d \tag{C.5}$$

$$(\lambda_1+\lambda_2)p_{i,i+l} = \lambda_2 r^i p_{i,i+l-1} + \lambda_2(1-r^{i+1})p_{i+1,i+l} + \lambda_1 r^{i+l} p_{i-1,i+l} + \lambda_1(1-r^{i+l+1})p_{i,i+l+1},\ i \geq 1, 1 \leq l \leq d, \tag{C.6}$$

$$\lambda_2 p_{i+d+1,i} = \lambda_1 r^i p_{i+d,i} + \lambda_1(1-r^{i+1})p_{i+d+1,i+1},\ i \geq 0, \tag{C.7}$$

$$\lambda_1 p_{i,i+d+1} = \lambda_2 r^i p_{i,i+d} + \lambda_2(1-r^{i+1})p_{i+1,i+d+1},\ i \geq 0, \tag{C.8}$$

$$\sum_{i=0}^{\infty}\sum_{j=i-d-1}^{i+d+1} p_{i,j} = 1. \tag{C.9}$$

We sum (C.2) for $i = 1$ to ∞ and then add (C.1) to get

$$(\lambda_1+\lambda_2)a_0 = \lambda_1 a_{-1} + \lambda_2 a_1. \tag{C.10}$$

Repeating the same procedures for pairs (C.3) and (C.4), (C.5) and (C.6), (C.7) and (C.8),

$$(\lambda_1+\lambda_2)a_l = \lambda_2 a_{l+1} + \lambda_1 a_{l-1},\ 1 \leq l \leq d, \tag{C.11}$$

$$(\lambda_1+\lambda_2)a_{-l} = \lambda_1 a_{-l-1} + \lambda_2 a_{-l+1},\ 1 \leq l \leq d, \tag{C.12}$$

$$\lambda_2 a_{d+1} = \lambda_1 a_d, \tag{C.13}$$

$$\lambda_1 a_{-d-1} = \lambda_2 a_{-d}. \tag{C.14}$$

We notice that, similar to the case $q = 1$, if we replace $b_l = a_{l-d}$ in (C.10)-(C.14), we obtain the global balance equations of an $M/M/1/2d+2$ system. Hence, the result follows. □

Proof of Theorem 2.14. Using PASTA property, $c_1 = a_{d+1}$ and the result follows from Lemma C.1. □

Proof of Theorem 2.16. Without lost of generality, assume $d \geq 0$ is an integer. The difference of average queue lengths can be written as

$$\begin{aligned} L_1^q - L_2^q &= \sum_{l=-d-1}^{d+1} l a_l \\ &= \frac{1 - \frac{\lambda_2}{\lambda_1}}{1 - (\frac{\lambda_2}{\lambda_1})^{2d+3}} (-\frac{\lambda_2}{\lambda_1})^{d+2} \sum_{l=-d-1}^{d+1} (-l)(\frac{\lambda_2}{\lambda_1})^{-l-1}. \end{aligned}$$

Using $\rho = \frac{\lambda_2}{\lambda_1}$,

$$\begin{aligned} \sum_{l=-d-1}^{d+1} (-l)(\frac{\lambda_2}{\lambda_1})^{-l-1} &= \sum_{l=-d-1}^{d+1} \frac{\partial}{\partial \rho} \rho^{-l} \\ &= \frac{\partial}{\partial \rho} \frac{\rho^{d+1}(1 - \rho^{-2d-3})}{1 - \rho^{-1}} \\ &= \frac{((d+2)\rho^{d+1} + (d+1)\rho^{-d-2})(\rho - 1) - (\rho^{d+2} - \rho^{-d-1})}{(\rho - 1)^2} \end{aligned}$$

Hence,

$$L_1^q - L_2^q = \frac{(d+2)\rho^{2d+3} + d + 1}{1 - \rho^{2d+3}} + \frac{(1-\rho)\rho^{d+2}(\rho^{d+2} - \rho^{-d-1})}{(1 - \rho^{2d+3})(\rho - 1)^2}$$

□

Bibliography

[1] I. Adan and G. Weiss. Exact FCFS matching rates for two infinite muti-type sequences. *Operations Research*, 60(2):475–489, 2012.

[2] P. Afeche. Incentive-compatible revenue management in queueing systems: Optimal strategic delay. *Manufacturing and Service Operations Management*, 15(3):423–443, 2013.

[3] S. M. Aseev and A. V. Kryazhimskii. The pontryagin maximum principle and optimal economic growth problems. *Proceedings of the Steklov Institute of Mathematics*, 257:1–255, 2007.

[4] S. Asmussen. *Applied Probability and Queues.* Applications of Mathematics Series. Springer, New York, 2003.

[5] B. R. Atar and M. I. Reiman. Asymptotically optimal dynamic pricing for network revenue management. *Stochastic Systems*, 2(2):232–276, 2012.

[6] N. Bäuerle. Optimal control of queueing networks: An approach via fluid models. *Advances in Probability*, 34(2):313–328, 2002.

[7] D. P. Bertsekas. *Dynamic Programming and Optimal Control.* Athena Scientific, 1995.

[8] P. Billingsley. *Convergence of Probability Measures.* Wiley, New York, 1999.

[9] C. Borgs, J. T. Chayes, S. Doroudi, M. Harchol-Balter, and K. Xu. The optimal admission threshold in observable queues with state dependent pricing. *Probability in the Engineering and Informational Science*, 28:101–119, 2014.

[10] B. Büke and H. Chen. Diffusion approximations for probabilistic matching systems. *Submitted*.

[11] B. Büke and H. Chen. Optimal control and dynamic pricing for probabilistic matching systems. *Working Paper*.

[12] B. Büke and H. Chen. Stabilizing policies for probabilistic matching systems. *Queueing Systems*, 80:35–69, June 2015.

[13] A. Bušić, V. Gupta, and J. Mairesse. Stability of the bipartite matching model. *Advances in Applied Probability*, 45(2):351–378, 2013.

[14] R. Caldentey, E. Kaplan, and G. Weiss. FCFS infinite bipartite matching of servers and customers. *Advances in Applied Probability*, 41(3):695–730, 2009.

[15] E. Çınlar. *Probability and Stochastics*. Graduate Texts in Mathematics. Springer, New York, 2011.

[16] H. Chen and M. Z. Frank. State dependent pricing with a queue. *IIE Transactions*, pages 847–860, 2007.

[17] H. Chen and D. D. Yao. *Fundamentals of Queuing Networks, Performance, Asymptotics, and Optimization*. Springer, New York, 2001.

[18] J. G. Dai and S. He. Customer abandonment in many-server queues. *Mathematics of Operations Research*, 35(2):347–362, 2010.

[19] R. W. R. Darling and J. R. Norris. Structure of large random hypergraphs. *The Annals of Applied Probability*, 15(1A):125–152, 2005.

[20] M. El-Taha and S. Stidham. *Sample-path Analysis of Queueing Systems*. Kluwer Academic Publisher, Boston/Dordrecht/London, 1999.

[21] G. Fayolle, V. A. Malyshev, and M. V. Menshikov. *Topics in the Constructive Theory of Countable Markov Chains*. Cambridge Univ. Press, Cambridge, UK, 1995.

[22] G. Gallego and G. J. van Ryzin. Optimal dynamic pricing of inventory with stochastic demand over finite horizons. *Management Science*, 40(8):999–1020, 1994.

[23] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing & Operations Research*, 4(3):208–227, 2002.

[24] D. Gross and C. M. Harris. *Fundamantals of Queueing Theory*. Wiley Series in Probability and Statistics. John Wiley & Sons, New York, 1998.

[25] I. Gurvich and A. R. Ward. On the dynamic control of matching queues. *Stochastic Systems*, 4(2):475–523, 2014.

[26] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, 1981.

[27] J. M. Harrison. Assembly-like queues. *J. Appl. Probab.*, 1973.

[28] R. Hassin and M. Haviv. *To Queue Or Not To Queue*. Kluwer Academic Publishers, 2003.

[29] B. R. K. Kashyap. The double-ended queue with bulk service and limited waiting space. *Operations Research*, 14(5):822–834, 1966.

[30] P. E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, 1999.

[31] V. G. Kulkarni. *Modeling and Analysis of Stochastic System*. Chapman and Hall/CRC, London, UK, 1996.

[32] G. Latouche. Queues with paired customers. *J. Appl. Probab.*, 18(3):684–696, 1981.

[33] X. Liu, Q. Gong, and V. G. Kulkarni. Diffusion models for double-ended queues with renewal arrival processes. 2014. Working Paper, arXiv:1401.5146.

[34] C. Maglaras. Revenue management for a multiclass single-server queue via a fluid model analysis. *Operations Research*, 54(5):914–932, 2006.

[35] A. Mandelbaumi and P. Momčilović. Queues with many servers and impatient customers. *Mathematics of Operations Research*, 37(1):41–65, 2012.

[36] P. Naor. The regulation of queue size by levying tolls. *Econometrica*, 37(1):15–24, 1969.

[37] S. H. Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering (Studies in Nonlinearity)*. Westview Press, 1994.

[38] A. Ward and S. Kumar. Asymptotically optimal admission control of a queue with impatient customers. *Mathematics of Operations Research*, 33(1):167–202, 2008.

[39] A. R. Ward and P. W. Glynn. A diffusion approximation for a markovian queue with reneging. *Queueing Systems*, 43(1/2):103–128, 2003.

[40] A. R. Ward and P. W. Glynn. A diffusion approximation for a $GI/GI/1$ queue with balking or reneging. *Queueing Systems*, 50(4):371–400, 2005.

[41] W. Whitt. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and their Application to Queues*. Springer-Verlag, Florham Park, NJ, USA, 2001.

[42] U. Yildirim and J. J. Hasenbein. Admission control and pricing in a queue with batch arrivals. *Operations Research Letters*, 38:427–431, 2010.