THE UNIVERSITY OF CHICAGO

PERFORMANCE ANALYSIS AND CONTROL OF QUEUING SYSTEMS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE UNIVERSITY OF CHICAGO
BOOTH SCHOOL OF BUSINESS
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
XIAOSHAN PENG

CHICAGO, ILLINOIS
JUNE 2017

To my family

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor Baris Ata for his generous and continuous support, guidance, and feedback throughout my PhD. He is the most influential person on my intellectual development. His support helped me complete the transition from consuming knowledge to creating knowledge. He not only spent tremendous time discussing the research problems but also trained me the essential skills of being a qualified researcher. Through the rough road to pursuing my PhD, he has provided endless support for me academically and emotionally. He has always been the source of encouragement and inspiration during the most difficult times when finishing this thesis. He not only gave me feedback on the research but also provided hearty advices on building my academic career. I am especially thankful for all the support and help I obtained from him during the job market year. I feel deeply grateful to have him as my advisor. In addition, his sincerity and dedication towards research was contagious and motivational for me. I appreciate the excellent example he has provided as a successful researcher and professor.

I would like to express sincere appreciation to my committee members, Prof. Don Eisenstein, Prof. Peter Glynn, Prof. Sunil Kumar and Prof. Varun Gupta for their interests in my research, their insightful feedback and advice, and their generous support and help in my job market year. They generously provided me the guidance to find the research direction, develop research ideas and progress the research.

I am also thankful to all other faculty members in the Ops group at Booth and other institutions who have helped me in various way: Prof. John Birge, Prof. René Caldentey, Prof. Ozan Candogan, Prof. Seyed Emadi, Prof. Itai Gurvich, Prof. Ozge Islegen, Prof. Chris Ryan and Prof. Yuan Zhong.

I also want to thank my friends for their endless support and encouragement emotionally during the doctoral jorney. They also made this process more enjoyable and colorful!

Words cannot express my gratitude towards my family for their unconditional support and love. I received from my parents all the help and love one could ask for during these

years, especially during the most difficult times. Finally, seeing my husband's passion and motivation on pursuing a demanding career also always filled me with energy. I dedicate this work to them.

ABSTRACT

This dissertation studies the performance analysis and control of a queueing system when customers make their own routing decision (which queue to join) and abandonment decision (when to abandon the queue). In particular, there are two problems studied.

The first part of the dissertation studies a multiclass queueing system with endogenous abandonments where the congestion affects customers' abandonment behavior and vice versa. This model captures this interaction by developing two closely related models: an abandonment model and a queueing model. In the abandonment model, customers take the virtual waiting time distribution as given. Class k customers receive a reward r_k from service and incur a cost c_k per period of waiting. Customers are forward looking and make wait or abandon decisions dynamically to maximize their expected discounted utilities. The queueing model takes the customers' abandonment time distribution as an input and studies the resulting virtual waiting time distribution. Because the multiclass queueing system is not amenable to exact analysis, we resort to an approximate analysis in the conventional heavy traffic limit (under the hazard rate scaling). Leveraging the so-called state-space collapse property, we provide a characterization of the system performance. Combining the results for the two models, we show that there exists a unique equilibrium in which the customers' abandonment time and the virtual waiting time for the various classes are consistent in the two models. Moreover, building on the closed-form results of Baccelli and Hebuterne [22] for the $M/M/1+G$ queue, we prove the existence and uniqueness of the equilibrium in that special case via an exact analysis. Lastly, we provide computational schemes to calculate the equilibrium numerically for both the single- and multi-class queueing systems.

The second part of the dissertation studies how to manage the callback option effectively to mitigate congestion due to temporary surges in the arrivals to a call center. The call arrival process can be an arbitrary point process, allowing uncertainty and temporary surges in the arrival rate, provided that the system is stable. However, particular attention will be paid to the Poisson process with the Cox-Ingresoll-Ross (CIR) process as its stochastic intensity

both in our model development and numerical results because of its practical importance, although our theoretical results hold for any arbitrary point process. When a customer arrives, the call center manager reviews the system state and decides whether to keep him in the online queue or to offer the callback option. For each customer in the online queue, she incurs a waiting cost of h per time unit. Similarly, whenever she routes a customer to an offline queue (for a callback later), she incurs a one-time penalty of p. Initially, we allow complete foresight policies that look into the entire future. We first study the case where all customers are willing to accept a callback offer. A simple lookahead policy that looks into the future for the next p/h time units is pathwise optimal among the complete foresight policies. Next, we consider the setting where some customers may reject the callback offer. We show that a modified lookahead policy that looks into the future arrivals and service completion times for the next p/h time units and uses the current number of customers in the system who previously rejected a callback offer (but does not look into the accept/reject decisions of future customers) is pathwise optimal among the complete foresight policies. Building on the insights gleaned from the optimal lookahead policies, we also propose a non-anticipating policy, referred to as the line policy, to decide when to offer the callback option. Lastly, we conduct a simulation study using a dataset from a US bank call center which shows that the line policy has excellent performance.

CHAPTER 1

INTRODUCTION

Starting with Naor [89], a large and growing literature studying rational customers in queueing systems emerged. The decisions of arriving customers can include whether to join the queue or balk (joining decision) and which queue to join (routing decision). After joining the queue, customers may decide whether to keep waiting in the queue or leave (abandoning or reneging decision) and whether to switch to a different queue (jockeying decision). These models use the equilibrium approach to study the queueing system when customers make rational decisions on their own by imposing consistency conditions on customers' rational decisions. Hassin and Haviv [60] provide a comprehensive survey of this literature.

This dissertation studies the performance analysis and control of a queueing system when customers make their routing decisions (which queue to join) and abandonment decisions (when to abandon the queue). In particular, we study two specific problems. Chapters 2 - 4 study the equilibrium of a multiclass queueing system in which customers waiting in the queues make rational abandonment decisions. Chapter 5 studies the optimal control of a single class queue with the callback option in which customers decide to take or reject the offered callback option.

Customers' abandonment decision and its resulting system characterization depend on whether they can observe the system state or not; see Chapters 2 and 3 of Hassin and Haviv [60] for a discussion of the observable and unobservable queues, respectively. In the observable case, customers join a physical queue and observe the system state, especially the queue length, and other customers' abandonment decisions. There are several papers studying customers' rational abandonment behavior in the observable case. For example, Afèche and Sarhangian [1] study the rational abandonment decisions of customers in an observable two-class priority system, and pricing as a tool to control this behavior. Other studies include Assaf and Haviv [11], Hassin [57] and Maglaras et al. [88]. In the unobservable setting, customers cannot observe the queue length and other customers abandonment

behavior. A number of papers study how information about the system impacts customers' abandonment behavior. For example, Armony et al. [10] studies the system equilibrium using a fluid model under delay announcements. Jennings and Pender [67] study a ticket queue in which a newly arriving customer cannot observe the queue length but knows her ticket number. Moreover, the abandonment decisions are unobservable. They compare this ticket queue with the observable queue and prove that they are indistinguishable in heavy traffic. Kuzu et al. [77] examine customers' abandonment behavior in a ticket queue using a dataset of a Turkish bank. Their empirical study reveals that customers update their forecast of the waiting time over time and adjust their abandonment decisions accordingly. Aksin et al. [4] conducts an empirical study of the impact of different delay announcements on system performance. Examples of unobservable queues include virtual queues in call centers and waiting lists of patients for organ transplant; see for example Ata et al. [20]. To the best of our knowledge, the first papers in this literature are Hassin and Haviv [59], Haviv and Ritov [61], Zohar et al. [121], Mandelbaum and Shimkin [86] and Shimkin and Mandelbaum [100]. These papers assume that the customers determine whether to join the queue and their abandonment time upon arrival. Hassin and Haviv [59] consider a queue in which the reward from service reduces to zero after a certain time and study the Nash equilibrium of customers' strategies. Haviv and Ritov [61] study the symmetric Nash equilibrium of customers' abandonment strategies of a queue in which customers have an increasing and convex waiting cost and a fixed reward from service. In Zohar et al. [121]'s model, the customers' abandonment time follows a parametric distribution. The parametric distribution depends only on a single statistic measure of the system performance, for instance, the average anticipated waiting time. Mandelbaum and Shimkin [86] and Shimkin and Mandelbaum [100] analyze rational models of a Markovian queue with a general abandonment time distribution (an M/M/s+GI queue). In both papers, the authors assume that the customers make a rational decision on when to abandon upon arrival with the waiting cost and the service utility of the callers given. Callers abandon the system if their actual waiting time exceeds

their optimal abandonment time.

Chapters 2 - 4 study the endogenous abandonment decisions of customers in the unobservable setting and incorporates such endogenous behavior for a multiclass queue. In contrast to the static abandonment models in the literature, we adopt the dynamic abandonment model developed by Aksin et al. [3]; also see Aksin et al. [4]. In this model, customers decide whether or not to abandon dynamically while they wait in the queue to maximize their utilities. Customers' utilities depend on the waiting costs, the reward from the service and their idiosyncratic random shocks. Customers maximize their utilities by solving an optimal stopping problem. Aksin et al. [3] apply this dynamic abandonment model to the dataset of an Israeli bank call center and study the customers' abandonment behavior in this call center. In equilibrium, customers' belief of the waiting time distribution should be consistent with the actual waiting time distribution. Rather than characterizing the system equilibrium analytically, Aksin et al. [3] focus on estimating preference parameters, i.e. the waiting costs and rewards of the customers. The main contribution of our paper is to propose a theoretical framework to study the system equilibrium and the corresponding system performance under the aforementioned endogenous abandonment model. We show that there exists a unique equilibrium and provide an algorithm to compute it and illustrate its performance.

We propose a modeling framework to characterize the equilibrium of the multiclass system in Chapter 2. The framework includes two parts: an abandonment model and a queueing model. The abandonment model characterizes how customers make abandonment decisions to maximize their utilities given their beliefs of the system performance characteristics. In particular, customers decide whether or not to abandon dynamically as they wait to maximize their expected discounted utility. The queueing model studies the system dynamics and performance given the customers' abandonment behavior. In an equilibrium, customers' abandonment behavior (as a function of the system behavior) and the implied system behavior must be consistent with each other. The characterization of the system dynamics and performance in a multiclass system is difficult even when the abandonment behavior is given

exogenously. Therefore, we study two special case: the exact analysis of a $Geo/Geo/1+G$ queue in Chapter 3 and the approximate analysis of a multiclass system in the conventional heavy traffic regime in Chapter 4.

Chapter 5 studies a demand-side intervention by offering the callback option to manage the arrival uncertainty when customers are allowed make their decisions of taking or rejecting the offered callback option to them. To be more specific, the callback option works as follows: when the system is congested, the call center manager notifies the arriving customer and presents him with the option to hang up to be called back later within a reasonable time window. When the system congestion decreases, outbound calls are initiated for such customers. Consequently, the callback option allows the call center manager to shift some of the calls arriving when the system is undergoing temporary arrival surges to a period when the system is less busy. A survey conducted by Software Advise shows that more than half of the customers are willing to wait for more than 30 minutes offline for the call center to get back to them. Thanks to the customers who are willing to accept the callback option, the call center manager can improve the performance metrics of the online queue significantly using the callback option.

This problem is studied using a canonical queueing model. The call center consists of two queues: An online queue and an offline queue. When a customer arrives, the call center manager examines the system state and decides whether to offer the incoming customer the callback option or not. If the call center offers the callback option to the customer, he may decide to accept or reject the offer. A customer is routed to the offline queue (to be called back later) only if he is offered the callback option and he accepts it. Otherwise, he is routed to the online queue. The call center incurs a waiting cost of h per unit time for each customer waiting in the online queue, whereas it incurs a one-time penalty of p if the customer is routed to the offline queue. We refer the call center manager's decision as the callback decision. In what follows, we assume that the time commitment for the callback option is sufficiently large so that we can relax the delay commitment in our theoretical model. Under this assumption,

the call center manager makes the callback decision by solving an admission control problem which allows customers to reject or follow the call center manager's admission decisions. This admission control problem is interesting in its own right. However, what makes it relevant to the callback option and the call center operations is the mean-reverting nature of call arrival rate processes on the mesoscopic time scale as observed by Glynn et al. [50]. To be more specific, although the results proved below hold for general point process, we pay particular attention to the Poisson process with stochastic intensity as a call arrival model because of its practical importance. In particular, we model the stochastic intensity as a Cox-Ingersoll-Ross (CIR) process, a mean reverting diffusion process, following Glynn et al. [50]. The mean-reversion feature of the CIR process makes it a good candidate for modeling temporary surges in arrivals in the mesoscopic time scale. When a temporary surge occurs, the call center offers the callback option. After the surge ends, which may last from a few minutes to an hour, the call center uses its excess service capacity to serve customers waiting for the callback. Given the mean-reverting nature of the CIR process (and that the system is stable), the delay commitment made to the offline queue will be met with high probability. Indeed, our simulation study shows that the average delays for the offline queue are around 30 minutes (unless the abandonment rate is high)[1]. The simulation study thus provides a guidance as to how the callback time commitment should be given.

1. The simulation study also allows abandonments. It shows that average delays as well as the 90^{th} percentile of the offline waiting time is less than one hour for most cases unless the abandonment rate is high.

CHAPTER 2
EQUILIBRIUM ANALYSIS OF A MULTICLASS QUEUE WITH ENDOGENOUS ABANDONMENTS

2.1 Introduction

Queueing models with abandonments have been studied extensively in the operations research literature both from the performance analysis and optimization perspectives; see Ward [107] for a survey. In these models, customers are typically endowed with exogenously given abandonment times. The distribution of the abandonment time is not affected by the system dynamics or any policy change. This assumption not only enables stochastic simulation studies to determine the impact of major operational changes, e.g. a change in the service discipline, on system performance, but also helps efforts for analytical characterizations.

The exogenous abandonments assumption may not always be suitable. On the contrary, in many service systems customers behave strategically. This gives rise to the endogenous customer (abandonment) behavior. As such systems go through important changes, e.g. policy changes or other important design/structural changes, it is important to take into account customers' endogenous abandonment behavior to circumvent potential unintended consequences. Not doing so may lead to significant miscalculations as we illustrate below (see for example Section 4.2.2). However, incorporating endogenous abandonments in service systems brings about additional challenges. For example, simulating the system performance under a new policy is no longer straightforward and requires an equilibrium approach.

In this paper, we study a multiclass queueing system with endogenous abandonments using an equilibrium approach. We propose a modeling framework to characterize the equilibrium of the multiclass system. The framework includes two parts: an abandonment model and a queueing model. The abandonment model characterizes how customers make abandonment decisions to maximize their utilities given their beliefs of the system performance characteristics. In particular, customers decide whether or not to abandon dynamically

as they wait to maximize their expected discounted utility. The queueing model studies the system dynamics and performance given the customers' abandonment behavior. In an equilibrium, customers' abandonment behavior (as a function of the system behavior) and the implied system behavior must be consistent with each other. while the performance characterization of the queueing model relies on the system infrastructure and the service discipline.

The characterization of the queueing systems' behavior depends on the system infrastructure and the service discipline. We study two specific cases: an single-class Markovian queue, i.e. a Geo/Geo/$1 + GI$ queue, and a multiclass system in heavy traffic. For the Geo/Geo/$1 + GI$ queue, we derive a discrete time analog of the characterization in Baccelli and Hebuterne [22]. To be more specific, we obtain a closed form characterization of the virtual offered waiting time (VOWT) distribution. This enables us to characterize the exact equilibrium of a single class queue. The characterization of the system dynamics and performance in a multiclass system is difficult even when the abandonment behavior is given exogenously. Therefore, we pursue an approximate analysis in the conventional heavy traffic regime. The heavy traffic approximation allows us to express the virtual offered waiting time in terms of queue lengths. Moreover, by the virtue of the state-space collapse result [e.g. 30], we characterize the one-dimensional workload process for the entire system as well as the individual queue length processes. Consequently, the system equilibrium (in the heavy traffic limit) is fully characterized by the workload process and a suitably defined aggregate abandonment rate function. We characterize the system equilibrium in heavy traffic by a fixed point equation defined in a suitable function space.

Our main result shows that there exists a unique equilibrium for both cases studied. The exact analysis of the single class system relies on the results in Baccelli and Hebuterne [22] and Baccelli et al. [23]. They provide necessary and sufficient conditions for existence of the steady-state virtual offered waiting time in the $G/G/s + GI$ systems. In addition, Baccelli and Hebuterne [22] obtain a closed form characterization of the (steady-state) distribution of

the virtual offered waiting time for the $M/M/s+GI$ system. We prove the main result for the multiclass system under the heavy traffic approximation by analyzing the workload process. We also provide a computational scheme and illustrate the equilibrium computation. To compute the equilibrium, we approximate it by a truncated one, where the abandonment probabilities of customers waiting beyond a threshold are replaced by a certain (exogenously given) value. We then provide an iterative algorithm to compute the truncated equilibrium. Lastly, we illustrate the effectiveness of the algorithm using data from an Israeli bank call center.

2.1.1 Main Contributions

The contribution of this paper lies in the theoretical framework it provides for studying the system equilibrium and the proof of existence and uniqueness of the equilibrium. This framework allows studying the performance under the endogenous (dynamic) abandonment model. An important antecedent of this paper is Aksin et al. [3]; see also Aksin et al. [4]. Aksin et al. [3] propose a dynamic abandonment model and estimate preference parameters, i.e. the waiting costs and rewards of the customers. Rather than estimating the parameters, our paper focuses on characterizing the system equilibrium analytically and providing a theoretical foundation for studying the system performance. Therefore, the equilibrium analysis in this paper provides a theoretical framework for the counterfactual analysis conducted in Aksin et al. [3] and other empirical work that builds on it. Moreover, we apply a novel asymptotic analysis to study the equilibrium of the multiclass system where the exact analysis is intractable. This approach is motivated by Armony and Maglaras [8][1]; see also Armony and Maglaras [9]. Lastly, we provide an algorithm to compute the system equilib-

1. Armony and Maglaras [8] use this approach to study the call center with a callback option and characterize the equilibrium asymptotically. Another closely related paper is Maglaras et al. [88], which also uses the asymptotic analysis to study a system with endogenous abandonment decisions. Our paper studies the system where customers know the service rate but cannot observe the system state; whereas the system state in Maglaras et al. [88] is observable but the customers learn the service rate by observing the evolution of the system.

rium and illustrate its performance using several numerical examples and the the data from an Israeli bank call center.

Our proof of uniqueness also makes a methodological contribution[2]. To the best of our knowledge, there is no theory that readily addresses or can be extended to address the question of uniqueness. Therefore, we establish the uniqueness proof from first principals. We express the equilibrium distributions of the waiting and abandonment times for the various customer classes as the solution to a system of non-linear ODEs (or dynamical system for the exact analysis of the single class queue). The existence and uniqueness of the equilibrium correspond to the existence and uniqueness of a solution to this system of ODEs. We reformulate the uniqueness question as one of the stability of a related auxiliary system of ODEs[3]. Ultimately, the analysis boils down to studying the eigenvalues of the state-transition matrix of the dynamical system[4]. A difficulty arises in our setting because one eigenvalue of the underlying state-transition matrix fails to satisfy the usual condition of stability. We resolve this by expressing the corresponding state variable in terms of others. This eliminates the "nonconforming" eigenvalue but contributes a perturbation error to the new state-transition matrix, which vanishes asymptotically as the waiting time, i.e. the time index of the auxiliary dynamical system, gets large. Ultimately, these steps translate the system into a well-behaved system of ODEs (with a perturbation).

The aforementioned approach may be applicable to establish the uniqueness of the equilibrium in other settings where forward-looking agents facing shared-resource constraints make decisions dynamically. For example, one important application area is deceased-donor organ transplants where patients make accept-reject decisions for the organ offers strate-

2. The proof of existence is relatively straightforward and follows from the Schauder-Tychonoff fixed point theorem in a suitably defined space of functions.

3. We proceed by contradiction, thereby postulating two distinct solutions, and study the system of ODEs governing evolution of their difference.

4. See for example Gurvits [54], Czornik [36] and Protasov et al. [92] for conditions for stability of discrete-time dynamical systems and Chapter 7 of Verhulst [106] for their continuous-time analogs, i.e. a system of nonlinear ODEs.

gically. Similar to abandonments in our setting, these (endogenous) decisions impact the evolution of the transplant queues. The deceased-donor allocation policies go under frequent changes; see for example Ata et al. [13] and Ata and Friedewald [14]. It is of great importance to predict the future (equilibrium) outcomes under such policy changes, which calls for an analysis that is similar to the analysis in this paper. The rest of this chapter is organized as follows.

2.1.2 Outline of Chapters 2 - 4

Chapters 2 - 4 are organized as follows. The rest of Chapter 2 introduces the general model frame work of a multiclass queueing system with endogenous abandonments. Section 2.2 reviews the related literature. Section 2.3 introduces a discrete-time multiclass queueing system with endogenous abandonments. It also provides the definition of the equilibrium. Chapter 3 provides the exact analysis of a single class Markovian queue. To be specific, we characterize the equilibrium for this special case and state the main result, i.e. the existence and uniqueness of the equilibrium in Section 3.1. We then prove the existence and uniqueness results in Sections 3.2. Section 3.3 provides a computational scheme to calculate the equilibrium numerically. We then conduct a numerical study to analyze the impact of modeling the abandonment decisions endogenously in Section 3.4. The asymptotic analysis of the multiclass system is provided in Chapter 4. Section 4.1 analyzes the equilibrium of a multiclass system in heavy traffic. Section 4.2 provides a computational scheme to calculate the equilibrium. It also provides a numerical example to illustrate the importance of modeling the abandonment decisions endogenously in various counterfactual scenarios using a data set from a bank call center. The proofs omitted in the main text are provided in the appendices.

The appendices are organized as follows. The proofs of the results in Section 2.3 are given in A.1.1. Appendicies A.1.2 - A.1.3 present the proofs of lemmas, corollaries and the main theorem in Sections 3.1.1 - 3.2. Appendix A.1.4 contains the proofs of propositions in Section

3.3. Appendix A.2 provides the proof of a technical lemma that facilitates the proof of the uniqueness of the equilibrium. We also provide a roadmap for the proof of the uniqueness of the equilibrium in Appendix A.3. Appendix B.1.1 provides the derivations of the results in Section 4.1.1. Proofs of results in Sections 4.1.2 and 4.1.3 are provided in Appendix B.1.2. Appendix B.1.3 contains the proofs of the results in Section 4.2. Appendix B.2 provides the proof of Lemma 13, a technical lemma that facilitates the proof of the uniqueness. Appendix B.3 provides a detailed roadmap for the proofs of the uniqueness and Lemma 13. The reader may benefit from reviewing Appendix B.3 before reading B.2. Appendix B.4.1 provides the maximum likelihood formulation used to compute the parameter estimates in Section 4.2.2. Appendix B.4.2 provides the Kolmogorov-Smirnov test conducted to compare the predicted (steady-state) distributions of VOWT from the endogenous and exogenous models.

2.2 Literature review

In the traditional approach, one endows customers with exogenous abandonment time (patience time) distributions. A customer abandons when his waiting time exceeds his patience time. The exact results are rare and require distributional assumptions on the arrival, service and abandonment processes; see Gans et al. [44] and Aksin et al. [2] for overviews and other examples. Because most queueing models with abandonments are not amenable to exact analysis, the heavy traffic asymptotic regime is often used for approximate analysis of such systems. For example, Ward and Glynn [108] study the heavy traffic limit of the $G/G/1+G$ queue in the conventional heavy traffic limit regime. Rubino and Ata [95] study a multiclass parallel-server queueing system with abandonments in the conventional heavy traffic regime. Ata and Tongarlak [21] study a multiclass queue with abandonments under nonlinear costs. We refer the readers to Ward [107] for a comprehensive survey of the approximate analysis of queueing systems with abandonments. In this paper, we use the hazard rate scaling in the conventional heavy traffic regime introduced by Reed and Ward [93]. Under the hazard rate scaling, the diffusion approximations of the virtual offered waiting time and queue length

processes use the entire abandonment distribution given in the $G/G/1+G$ system. Kim and Ward [69] study the optimal control problem in a multiclass $G/G/1+G$ queue under the hazard rate scaling.

Another stream of literature models customer abandonments endogenously whereby customers make their abandonment decisions by maximizing their utilities. The literature studying rational customers in queueing systems starts with Naor [89]. Arriving customers' decisions includes whether to join the queue or balk (joining decision) and which queue to join (routing decision). Customers waiting in the queues may decide whether to keep waiting in the queue or leave (abandoning or reneging decision) and whether to switch to a different queue (jockeying decision). Hassin and Haviv [60] provide a comprehensive survey of this literature. This paper falls into the literature studying customers' abandonment decisions. These decisions and the resulting system behavior (and its characterization) depend on whether the customers can observe the system state or not; see Chapters 2 and 3 of Hassin and Haviv [60] for a discussion of the observable and unobservable queues, respectively. If the queue is unobservable, the only information of a customer is her own waiting time. To the best of our knowledge, the first papers in the unobservable setting are Hassin and Haviv [59], Haviv and Ritov [61], Zohar et al. [121], Mandelbaum and Shimkin [86] and Shimkin and Mandelbaum [100]. These papers assume that the customers determine their abandonment time upon arrival to maximize their utilities. Haviv and Ritov [61] consider the symmetric Nash equilibrium of customers' abandonment strategies with an increasing and convex waiting cost and a fixed reward from service. Zohar et al. [121] present a model in which the customers' patience time follow a parametric distribution, which depends only on a single performance measure, e.g. average anticipated waiting time. Mandelbaum and Shimkin [86] and Shimkin and Mandelbaum [100] analyze rational abandonment behavior of impatient customers in a Markovian queue with a general abandonment time distribution (an M/M/m+G queue). In both papers, the authors assume that the waiting cost and the service utility of the callers are given, and customers, depending on these parameters, act

rationally and decide upon arrival when to abandon if they do not receive service. Callers abandon the system if their actual waiting time exceeds their optimal abandonment time.

In the observable case, customers join a physical queue and observe the system state, especially the queue length, and other customers' abandonment decisions. There are several papers studying customers' rational abandonment behavior in the observable case. Assaf and Haviv [11] study an $M/M/1$ queue in which customers observe the system state and make abandonment decisions to maximize their utilities. The authors characterize the stationary Markovian strategies in an $\epsilon-$Nash equilibrium of the system. Afèche and Sarhangian [1] study the rational abandonment decisions of customers in an observable two-class priority system, with pricing as a tool to control this behavior. They characterize the equilibrium abandonment strategy of low-priority customers and show how it depends on the high-priority queue length and on the fee structure. Welfare maximization requires charging customers only for service, whereas revenue maximization requires charging for joining the queue and offering a partial refund for order cancellation. A ticket queue falls in-between the observable and unobservable cases. A customer in a ticket queue cannot observe the queue length but can observe the abandonments and service completions in the system. The empirical study of Gao and Kuzu [47] finds that customers react dynamically and reconsider their abandonment decisions when they see abandonments and service completions.

The most closely related papers to ours are Aksin et al. [3, 4] and Maglaras et al. [88]. Aksin et al. [3, 4] propose a dynamic abandonment model. They assume that the customers make their abandonment decisions dynamically as they wait by solving an optimal stopping problem. Customers maximize their utilities which depend on the waiting costs, the reward from the service and their idiosyncratic random shocks. Aksin et al. [3] focus on estimating the waiting costs and rewards of the customers and use an iterative simulation-based approach to calculate the steady state distribution of the virtual offered waiting time and the abandonment time in the equilibrium. Aksin et al. [4] incorporate delay announcements and develop a Markovian approximation for a two-class queueing system, which gives rise

to a set of system equations to characterize the equilibrium. Whereas the focus of Aksin et al. [3, 4] is the empirical study of abandonments, this paper develops a framework for studying the equilibrium theoretically. In particular, we establish existence and uniqueness of the equilibrium and provide an algorithm to compute it.

Maglaras et al. [88] analyze the equilibrium of a queueing system in which customers observe the evolution of the queue but do not know the service rate. Customers first join the queue and observe their progress to estimate their wait times and subsequently decide whether to stay in the queue or abandon. They study systems where the arrival rate and service capacity gets large, but the individual behavior of customers remains fixed. They characterize the equilibrium of this asymptotic system and use it to approximate the behavior of the pre-limit system. They show that the system dynamics are significantly impacted by long service times, characterized by the tail of the service time distribution. Both Maglaras et al. [88] and this paper use an asymptotic approach to study the equilibrium of the queue where customers in the queue make endogenous abandonment decisions. In contrast, this paper assumes the service rate is known whereas the queue is unobservable.

From the style of analysis perspective, important antecedents of this paper are Armony and Maglaras [8, 9] that study a call center in which customers choose to either stay in the queue or receive a call-back service to maximize their utilities. When customers choose the call-back service, which has a maximum-delay guarantee, they leave the real-time queue and wait in an offline queue. Characterizing the equilibrium analytically using the standard queueing models is not tractable. Therefore, the authors take a novel approach and study instead an approximate limiting system (in the so-called heavy traffic limit) which is amenable to analysis. We follow Armony and Maglaras' approach and characterize the equilibrium for the limiting system (in the heavy traffic limit).

From the methodological perspective, our model relates to self-interacting Markov chains and the nonlinear Perron-Frobenius theory. The defining feature of a self-interacting Markov chain is that its evolution depends on the occupancy measure of the past values; see Del Moral

and Miclo [39]. In our model, the evolution of the system dynamics depends on the past random abandonment decisions. Thus, our problem can be modeled as a self-interacting Markov chain. Moreover, finding the steady state distribution of self-interacting Markov chains can be viewed as finding the eigenvalues and eigenvectors of a nonlinear mapping, which is the focus of nonlinear Perron-Frobenius Theory; see Lemmens and Nussbaum [79]. In both streams of literature, establishing uniqueness of the equilibrium essentially relies on contraction mapping arguments. The abstract fixed point equation characterizing the equilibrium is not a contraction in our setting. Therefore, our proof of uniqueness proceeds from first principles.

2.3 The Model

We consider a multiclass queueing system with endogenous abandonments. That is, customers' abandonment decisions depend on the congestion in the system. Characterizing the (aggregate) behavior of such a system requires studying two closely-related questions:

1. How the congestion affects customers' abandonment decisions;

2. How customers' abandonment decisions impact system performance.

To address the first question, we introduce a model of forward-looking customers who make abandonment decisions to maximize their utility given the underlying system dynamics. We then incorporate the resulting abandonment behavior into the queueing model to characterize the equilibrium. To this end, the next section describes the primitives of the queueing model.

2.3.1 The queueing model

We consider a discrete time model of a single-server queueing system with K customers classes, indexed by $k = 1, \ldots, K$. In each period, at most one customer arrives to the system. A key primitive of the arrival process is the function $a(t)$ (for $t \geq 0$) which denotes

the probability that there is an arrival in the current period given that no customers arrived in the previous $t-1$ periods. Assuming the interarrival times of the customers are i.i.d., let A denote an interarrival time. Its cumulative distribution is given as follows:

$$\mathbb{P}(A \leq t) = 1 - \prod_{i=1}^{t}(1 - a(i)) \quad \text{for} \quad t = 1, 2, \ldots. \tag{2.1}$$

If the interarrival time is t, then a newly arrived customer is a class k customer with probability $p_k(t)$, where $\sum_{k=1}^{K} p_k(t) = 1$ for $t \geq 1$. The arrival rate of class k customers, denoted by λ_k, is given by the following:

$$\lambda_k = \left[\sum_{i=1}^{\infty} p_k(i)a(i) \prod_{j=1}^{i-1}(1 - a(j))\right] / \left[\sum_{i=1}^{\infty} ia(i) \prod_{j=1}^{i-1}(1 - a(j))\right]. \tag{2.2}$$

Let $\sigma_{A,k}$ denote the standard deviation of the interarrival time of class k customers, which can be calculated from the primitives (a, p_k).

We assume a non-preemptive, work-conserving scheduling policy which dictates the next customer class to work on every time the server is done with the customer currently in service. Within a class, customers are served in a First-Come-First-Served (FCFS) fashion. The service times of class k customers are i.i.d. and have general distributions. Letting S_k denote a class k service time distribution, its hazard rate is given by b_k. That is, $b_k(s)$ denotes the probability that the service is completed in the next period given that it has lasted longer than $s-1$ periods. Therefore, the cumulative distribution of the service time S_k is given as follows:

$$\mathbb{P}(S_k \leq s) = 1 - \prod_{i=1}^{s}(1 - b_k(i)) \quad \text{for} \quad s = 1, 2, \ldots \quad \text{and} \quad k = 1, \ldots, K. \tag{2.3}$$

Letting μ_k denote the service rate for class k customers, it is given as follows:

$$\mu_k = \left[\sum_{i=1}^{\infty} i b_k(i) \prod_{j=1}^{i-1} (1 - b_k(j)) \right]^{-1}. \tag{2.4}$$

The standard deviation of the service time for class k customers, denoted by $\sigma_{S,k}$, can be calculated from equation (2.3). Let $\rho_k = \lambda_k/\mu_k$ denote the offered load for class k for $k = 1, \ldots, K$. We only consider the underloaded case, i.e. $\sum_{k=1}^{K} \rho_k \leq 1$.

A crucial feature of our model is that customers make real time abandonment decisions while waiting; a detailed description of that will be provided in the next section. Until then, we assume the abandonment time follows a general distribution, where a class k customer who has waited for w periods abandons with probability $q_k(w)$, $k = 1, \ldots, K$ and $w \geq 0$.

The system dynamics are fully characterized by the primitives of the arrival process (a, p_k), the service process b_k, the scheduling policy and the abandonment probabilities q_k. Therefore, we can characterize how the system evolves in steady state. A quantity of primary interest for us is $\beta_k(w)$, the probability of entering service in the next period for a class k customer given that she has been in the queue for w periods. That is, $\beta_k(w)$ is the hazard rate of the virtual offered waiting time distribution for class k. The probability $\beta_k(w)$ will be the key input into our customer abandonment model, summarizing the queueing dynamics. Closed-form expressions for $\beta_k(\cdot)$ are not available in general. We provide a closed-form formula for $\beta(\cdot)$ in Section 3.2 for the special case of a (single-class) $M/M/1 + G$ system; Section 4.1 provides an approximation for the multiclass case.

A detailed description of our abandonment model is given in the next section, which builds on Aksin et al. [3].

2.3.2 An abandonment model with forward-looking customers

In our abandonment model, customers make their abandonment decisions based on their beliefs of the probability of entering service, which also characterizes their waiting time.

We assume that[5] the customers have "true" beliefs, i.e. their beliefs coincide with $\beta_k(\cdot)$. Consider a class k customer who has been waiting in the queue for w periods. Figure 2.1 shows the sequence of events that she experiences during the next period. We denote her

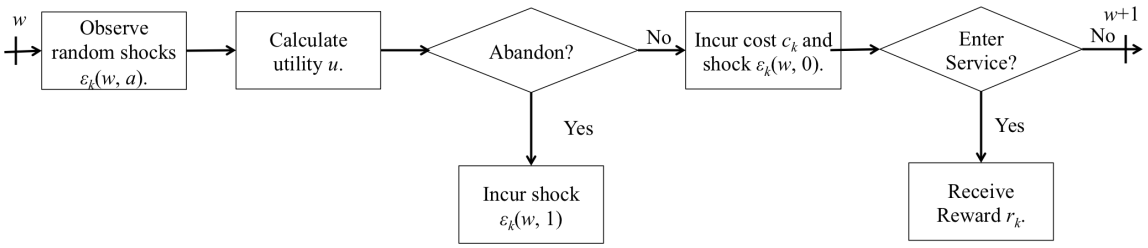


Figure 2.1: Timeline of the Events

abandonment decision by $a \in \{0, 1\}$; $a = 0$ corresponds to staying in the queue whereas $a = 1$ corresponds to abandoning. There is a random shock, denoted by $\epsilon_k(w, a)$ for a class k customer, associated with each action a. First, she learns the realizations of the random shocks and calculates her utility under each action $a \in \{0, 1\}$, which we denote by $u_k(w, a, \varepsilon_k(w, a))$. Then she makes the abandonment decision so as to maximize her utility. If she chooses to abandon ($a = 1$), she only incurs the random shock $\varepsilon_k(w, 1)$. Her expected discounted future utility from abandoning is normalized to zero. If she chooses to stay in the queue ($a = 0$), then she incurs a waiting cost of c_k for that period as well as the shock $\varepsilon_k(w, 0)$. Moreover, she enters service with probability $\beta_k(w)$ in which case she receives a reward r_k from service. In making her decision of whether to abandon or wait, she takes into account her expected discounted future utility (or "value-to-go") as well. In summary, the (expected) utility of the customer as a function of her action (at the time of making her abandonment decision) is given as follows:

$$u_k(w, a, \varepsilon_k(w, a)) = \begin{cases} -c_k + \alpha[\beta_k(w) r_k + (1 - \beta_k(w)) J_k(w + 1)] + \varepsilon_k(w, 0), & \text{if } a = 0, \\ \varepsilon_k(w, 1), & \text{if } a = 1, \end{cases} \tag{2.5}$$

5. We assume customers learn this through previous experience.

where α is the discount factor and $J_k(w+1)$ is the value-to-go function having waited for $w+1$ periods (or, the expected discounted future utility of waiting). The expected discounted future utility of waiting $J_k(w)$ is given as follows:

$$J_k(w) = \mathbb{E}_{\varepsilon_k}\left[\max_{a\in\{0,1\}} u_k(w,a,\varepsilon_k(w,a))\right], \tag{2.6}$$

where $\varepsilon_k(w) = (\varepsilon_k(w,0), \varepsilon_k(w,1))$. The customer's optimal action $a_k^*(w,\varepsilon_k(w))$ is given by

$$a_k^*(w,\varepsilon_k(w)) = \arg\max_{a\in\{0,1\}} u_k(w,a,\varepsilon_k(w,a)). \tag{2.7}$$

We make the following two assumptions on the idiosyncratic shocks.

Assumption 1. *The idiosyncratic shocks satisfy the following properties:*

1. *The idiosyncratic shocks are independent across different classes for $k = 1,\ldots,K$ and customers.*

2. *For each class k they are i.i.d. with zero mean for all $w \geq 1$ and $a \in \{0,1\}$, i.e. $\mathbb{E}[\epsilon_k(w,a)] = 0$.*

3. *The cumulative distribution function of $\epsilon_k(1,1) - \epsilon_k(1,0)$, denoted by $F_k(\cdot)$, has a positive support on $[-c_k, r_k]$. In addition, $F_k(\cdot)$ admits a continous probability density function $f_k(\cdot)$ on $[-c_k, r_k]$.*

For notational brevity, we suppress the dependence of idiosyncratic shocks on the waiting time and write $\varepsilon_k(a)$, $a \in \{0,1\}$. Assumption 2 involves the idiosyncratic shocks, the service reward r_k and the per-period waiting cost c_k.

Assumption 2. *Customers prefer receiving the service immediately to waiting for one period before entering service, i.e. $r_k > \mathbb{E}_\epsilon[\max(\varepsilon_k(1), -c_k + \alpha r_k + \varepsilon_k(0))]$ for all k.*

The expected discounted utility of waiting $J_k(w)$ solves the following Bellman equation:

$$J_k(w) = \mathbb{E}_{\varepsilon_k}\left[\max\{\varepsilon_k(1), -c_k + \alpha[\beta_k(w)r_k + (1-\beta_k(w))J_k(w+1)] + \varepsilon_k(0)\}\right]. \tag{2.8}$$

In light of (2.7), the probability that a class k customer abandons the queue after waiting for w periods, $q_k(w)$, is given as follows:

$$q_k(w) = \mathbb{E}_{\varepsilon_k}[a_k^*(w, \varepsilon_k)] = \mathbb{P}(u_k(w, 1, \varepsilon_k(1)) \geq u_k(w, 0, \varepsilon_k(0))). \tag{2.9}$$

Substituting equation (2.5) into equation (2.9) yields the following equation that derives the abandonment probability $q_k(\cdot)$ from the expected value of waiting $J_k(\cdot)$:

$$\begin{aligned} q_k(w) &= \mathbb{P}(\varepsilon_k(1) - \varepsilon_k(0) \geq -c_k + \alpha\left\{\beta_k(w)r_k + (1-\beta_k(w))J_k(w+1)\right\}) \\ &= \bar{F}_k\left(-c_k + \alpha\left\{\beta_k(w)r_k + (1-\beta_k(w))J_k(w+1)\right\}\right), \end{aligned} \tag{2.10}$$

where $\bar{F}_k(\cdot) = 1 - F_k(\cdot)$. The following proposition (see Section A.1.1 for its proof) shows that the expected value of waiting $J_k(\cdot)$, and thus the abandonment probability $q_k(\cdot)$, is unique given the belief $\beta_k(\cdot)$.

Proposition 1. *For $k = 1, \ldots, K$, given $\beta_k(\cdot)$, the expected value of waiting $J_k(\cdot)$ for a class k customer is the unique solution to the Bellman equation* (2.8)*. Moreover, the corresponding abandonment probability $q_k(\cdot)$ is uniquely characterized by equation* (2.10)*.*

As mentioned earlier, under Assumption 1, the expected discounted future utility of abandoning is zero, i.e.

$$\mathbb{E}[u_k(w, 1, \varepsilon_k(w, 1)] = 0 \quad \text{for} \quad w \geq 1.$$

In addition, by substituting equation (2.10) into equation (2.5), the utility of staying in the queue can be written in terms of the abandonment probability, i.e.

$$u_k(w, 0, \varepsilon_k(w, 0)) = \bar{F}_k^{-1}(q_k(w)) + \varepsilon_k(0).$$

Thus, we obtain the following corollary that expresses the expected value of waiting $J_k(\cdot)$ in terms of the abandonment probability $q_k(\cdot)$.

Corollary 1. *The following holds:*

$$J_k(w) = \mathbb{E}_{\varepsilon_k}\left[\bar{F}_k^{-1}(q_k(w)) + \varepsilon_k(0) - \varepsilon_k(1)\right]^+, \ k = 1, \dots K. \tag{2.11}$$

By Corollary 1 and Assumptions 1 and 2, the following corollary shows that for any given β_k, the abandonment probability is bounded away from zero.

Corollary 2. *For any given $\beta_k(\cdot)$, we have that $J_k(w) \leq r_k$ for all w and k. Moreover, there exists a constant $\underline{q}_k > 0$ such that $q_k(w) \geq \underline{q}_k$ for all w and $k = 1, \dots, K$.*

2.3.3 Definition of a system equilibrium

Note that the belief of entering service $\beta_k(w)$ and the abandonment probability $q_k(w)$ (for $w \geq 1$) can be viewed as infinite dimensional vectors. Let $\Omega = [0,1]^\infty$ so that both $\beta_k, q_k \in \Omega$. Also let $\beta = (\beta_1, \dots, \beta_K)$ and $q = (q_1, \dots, q_K)$ and note that $\beta, q \in \Omega^K$. Section 2.3.1 describes the queueing model which takes the abandonment probability q as a primitive along with the arrival and service processes and the scheduling policy. We refer the reader to Peng (2015) for the details of the description of the evolution of the queueing process and the resulting characterization of the probability β of entering service. To the best of our knowledge, there is no closed form representation of β in general. Nonetheless, we let $\Phi : \Omega^K \to \Omega^K$ denote the mapping from the abandonment probability q to the beliefs β. Note that the mapping Φ depends on other system primitives and, in particular on the scheduling policy. Equations (2.8) and (2.10) of Section 2.3.2 provides a characterization of the mapping from the belief β_k to the abandonment probability q_k for class k customers. We denote that mapping by $\Gamma_k : \Omega \to \Omega$.

In an equilibrium, the beliefs of entering service β and the abandonment probabilities q must be consistent with each other. The following definition of system equilibrium imposes

the consistency requirement on customers' abandonment probabilities q and their beliefs β.

Definition 1. *We say that* $e^* = (\beta^*, q^*)$*, where* $\beta^* = (\beta_1^*, \ldots, \beta_K^*)$ *and* $q^* = (q_1^*, \ldots, q_K^*)$*, is a system equilibrium if the following conditions are both satisfied:*

1. *The class* k *customers make abandonment decisions with beliefs* β_k^**, i.e.* $q_k^* = \Gamma_k(\beta_k^*)$*,* $k = 1, \ldots, K$*.*

2. *The beliefs* β^* *are consistent with the actual probability of entering service with abandonment probabilities* q^**, i.e.* $\beta^* = \Phi(q^*)$*.*

In the following two chapters, we characterize the equilibrium for two specific cases: the $Geo/Geo/1 + G$ queue and the multiclass system in the conventional heavy traffic limit.

CHAPTER 3

EXACT ANALYSIS OF A SINGLE-CLASS MARKOVIAN QUEUE

3.1 Characterization of the Equilibrium of a single-class Markovian Queue

The section introduces the framework to study a single-class queueing with endogenous abandonments. We first analyze the queueing model for this special case in Section 3.1.1. The abandonment model follows Section 2.3.2 with $k = 1$. We then state the main result at the end of this section.

3.1.1 The queueing model

The queueing model characterizes the system performance by taking the distribution of customers' abandonment time as given. We consider a Geo/Geo/1+GI queue in which customers' abandonment times follow a general distribution. In particular, we consider a single class queue with a single server[1]. In each period, a new customer arrives at the system with probability a. We assume a non-idling service policy. Customers are served in a First-Come-First-Served (FCFS) fashion. The service times follow a geometric distribution with probability b. We consider only the underloaded case[2], i.e. $a < b$.

1. The single-server assumption eliminates the possibility of having multiple service completions in one period, simplifying the characterization of the waiting time distribution significantly in our discrete-time model. Under this assumption, the characterization of the hazard rate of the waiting time distribution is a straightforward analogue of that of Baccelli and Hebuterne [22] in their continuous-time model. Their characterization of the waiting time distribution is valid for the multi-server queueing systems as well. This observation leads us to conjecture that our existence and uniqueness results can be extended to the multiserver setting as well.

2. Although we restrict attention to the underloaded case, i.e. $a < b$, we conjecture that the existence and uniqueness results continue to hold in the general case when the stability condition $a < bG(\infty)$ holds; see Equation (3.1) for the definition of G. The intuition for this stems from the fact that the proofs in Section 3.2 rely merely on the properties of the abandonment decisions and performance metrics of the queue when the waiting time is large. Focusing on the underloaded case, i.e. $a < b$, relieves us from the burden of

A crucial feature of our model is that customers make real time abandonment decisions while waiting; a detailed description of that will be provided in the next section. Until then, we assume the abandonment time follows a general distribution. To be specific, a customer who has waited in the queue for w periods abandons with probability $q(w)$ for $w \geq 1$.

A quantity of primary interest for us is $\beta(w)$, the steady-state probability of entering service in the next period for a customer who has been in the queue for w periods. This is because the (steady-state) probability β of entering service is a key input of customers' abandonment decisions described in Section 2.3.2. We first study the virtual offered waiting time (VOWT) process, denoted by $\{V(t) : t \geq 1\}$, where $V(t)$ is the number of periods a customer arriving at time t has to wait if she does not abandon. Then the (steady-state) probability β of entering service is the hazard rate of the steady state distribution of the VOWT.

Baccelli and Hebuterne [22] calculate the VOWT distribution of an $M/M/s+GI$ queue in closed form. The authors observe that the steady state distributions of the VOWT of the following two systems are equivalent, which is key to their analysis:

- (System 1) A customer calculates her VOWT upon arrival and balk if it exceeds her patience time.
- (System 2) A customer joins the queue regardlessly and abandons when her patience time expires.

The system of interest for us falls into the second case (System 2). Following this observation, we can study the VOWT of System 2 by analyzing the VOWT of System 1. Baccelli and Hebuterne [22] derive the generalized Lindley's equation [22, Equation 2.1] to characterize the steady state distribution of the VOWT (of System 1). We adapt their approach to our discrete-time setting. To this end, let $G(w)$ denote the probability that a customer abandons

working with different characterizations of the system as a function of the waiting time, and thus, simplifies the proof of the main result.

the queue within w periods. Note that

$$G(w) = 1 - \prod_{i=1}^{w}(1 - q(i)), \tag{3.1}$$

and let $\bar{G}(w) = 1 - G(w)$. Conditioning on the service time of a (potential) customer who arrives in period t, we write a recursive equation to characterize the dynamics of the VOWT $V(t)$ (for $t \geq 1$). This equation is the discrete time analogue of the generalized Lindley's equation in Baccelli and Hebuterne [22], which is given as follows: For $m = 1, 2, \ldots$,

$$V(t+1) = \begin{cases} (V(t)-1)^+, & \text{w.p.} \quad (1-a) + aG(V(t)), \\ V(t) - 1 + m & \text{w.p.} \quad a(1 - G(V(t)))(1-b)^{m-1}b. \end{cases} \tag{3.2}$$

Note that the VOWT $V(t)$ is the sum of the service time of all customers (in System 1) at the beginning of period t. Thus, if no customers enters the system in period t, the VOWT decreases by one in a non-empty system. If the system is empty, the VOWT remains zero. This happens either when no customer arrives (with probability $1 - a$) or when a customer arrives, finds that the VOWT exceeds her patience time (with probability $aG(V(t))$), and balks. This gives the first case in equation (3.2). The second case represents the scenario when a customer arrives the system and does not balk (with probability $a(1 - G(V(t)))$) in period t. Conditioning on the service time of this arriving customer, we obtain that the VOWT $V(t+1)$ becomes $V(t) - 1 + m$ (with probability $a(1 - G(V(t)))(1-b)^{m-1}b$).

The system dynamics are fully characterized by the arrival probability a, the probability of service completion b and the abandonment probabilities $q(\cdot)$. Therefore, we can characterize how the system evolves in steady state. By analyzing equation (3.2), we obtain Proposition 2 that is the discrete-time analogue of Baccelli and Hebuterne's result. It characterizes the (steady-state) probability of entering service $\beta(\cdot)$ given their abandonment probabilities $q(\cdot)$; see Appendix A.1.2 for its proof.

Proposition 2. *The probability of entering service after waiting for w periods in steady*

state is given as follows:

$$\beta(w) = \left(1 + \sum_{t=w+1}^{\infty} \prod_{i=w+1}^{t} \frac{1-b}{1-a\bar{G}(i)}\right)^{-1}, \ w \geq 1. \tag{3.3}$$

The following recursive equation is immediate from equation (3.3):

$$\frac{1}{\beta(w)} = 1 + \frac{1-b}{1-a\bar{G}(w+1)} \frac{1}{\beta(w+1)}. \tag{3.4}$$

The queueing model characterizes the system performance as if the abandonment probabilities $q(\cdot)$ were known. We complete the characterization of the system by introducing a rational abandonment model in the next subsection. The abandonment model takes the system performance (in steady state) as an input and gives the abandonment probabilities $q(\cdot)$ as the output.

3.1.2 Existence and uniqueness of the system equilibrium

In an equilibrium, the probability of entering service β and the abandonment probability q must be consistent with each other. To facilitate the formal definition of the equilibrium, let $\Omega = [0,1]^{\infty}$. Note that the probability of entering service $\beta(w)$ and the abandonment probability $q(w)$ (for $w \geq 1$) can be viewed as infinite dimensional vectors. Thus, $\beta, q \in \Omega$. We let $\Phi : \Omega \to \Omega$ denote the mapping from the vector q of the abandonment probability to the vector β of the probability of entering service, which is characterized by Proposition 2. Equations (2.8) and (2.10) provide the characterization of the mapping from the vector β of the probability of entering service to the vector q of the abandonment probability. We denote that mapping by $\Gamma : \Omega \to \Omega$.

The following definition of system equilibrium imposes the consistency requirement on customers' abandonment probability q and their probability of entering service β.

Definition 2. *We say that $e^* = (\beta^*, q^*)$ is a system equilibrium (in steady state) if the*

following conditions are both satisfied:

1. *The customers make abandonment decisions with the (steady state) probability of entering service β^*, i.e. $q^* = \Gamma(\beta^*)$.*

2. *The (steady state) probability of entering service β^* is consistent with the actual probability of entering service with abandonment probability q^*, i.e. $\beta^* = \Phi(q^*)$.*

The definition of the system equilibrium requires that the customers' beliefs on the system performance, which is characterized by β, are consistent with the actual system performance in steady state[3]. The following corollary, which combines the analysis from the abandonment and queueing models, provides the characterization of the system in equilibrium. It is immediate from Propositions 1 and 2.

Corollary 3. *The equilibrium is characterized by equations* (2.8), (2.10)(3.1) *and* (3.3), .

We end this section by stating our main result which establishes the existence and uniqueness of the equilibrium. Next section proves this theorem.

Theorem 1. *There exists a unique system equilibrium e^*.*

3.2 Proof of Theorem 1

The proof of Theorem 1 includes two parts. Section 3.2.1 establishes the existence of the equilibrium building on several auxiliary lemmas. Section 3.2.2 proves the uniqueness.

3.2.1 The Existence of the Equilibrium

Let $e^* = (\beta^*, q^*)$ be an equilibrium. By definition of an equilibrium, the probability $\beta^*(\cdot)$ of entering service is the solution to the fixed point problem $\beta^* = \Phi(\Gamma(\beta^*))$. We use Brouwer-Schauder fixed point theorem to prove the existence the equilibrium. We first state two

3. The definition of the system equilibrium is a symmetric Nash equilibrium with indistinguishable infinitely many players; see 1.1 of Chapter 1 in Hassin and Haviv [60] for a detailed discussion.

lemmas in preparation for applying the fixed point theorem. We then prove the existence result at the end of this subsection.

The following lemma provides an upper-bound and a lower-bound for the belief $\beta^*(\cdot)$; see Appendix A.1.3 for its proof.

Lemma 1. *Given $\beta(w) \in [0, b]$ for $w \geq 1$, let $\tilde{\beta} = (\tilde{\beta}(w) : w \geq 1)$ be defined as $\tilde{\beta} = \Phi(\Gamma(\beta))$. Then $\tilde{\beta}(w)$ is increasing in w and satisfies the following inequality:*

$$\frac{b - a(1-\underline{q})^{w+1}}{1 - a(1-\underline{q})^{w+1}} \leq \tilde{\beta}(w) \leq b \quad \textit{for} \quad w \geq 1, \tag{3.5}$$

where $\underline{q} \in (0, 1)$ is the constant given in Corollary 2.

Since the probability $\beta^*(\cdot)$ of entering service in equilibrium satisfies $\beta^* = \Phi(\Gamma(\beta^*))$, it also satisfies equation (3.5). To state the next lemma, define the set of infinite sequences $\mathcal{B}$ as follows:

$$\mathcal{B} = \left\{ \beta \in l^\infty : \frac{b - a(1-\underline{q})^{w+1}}{1 - a(1-\underline{q})^{w+1}} \leq \beta(w) \leq b, w \geq 1 \right\}, \tag{3.6}$$

where l^∞ is the space of infinite sequences endowed with the topology induced by the sup norm. Lemma 2 shows that the fixed point problem $\beta^* = \Phi(\Gamma(\beta^*))$ satisfies the conditions of the Brouwer-Schauder fixed point theorem; see Appendix A.1.3 for its proof.

Lemma 2. *The set $\mathcal{B}$ is compact in l^∞. In addition, the mapping $\Phi(\Gamma(\cdot))$ is continuous.*

Thus, we immediately obtain the existence result stated in the following proposition.

Proposition 3. *There exists a system equilibrium e^*.*

Proof. By Lemma 1, we can restrict our attention to $\beta \in \mathcal{B}$. By Lemma 2, the set $\mathcal{B}$ is compact in l^∞ and the mapping $\Phi(\Gamma(\cdot))$ is continuous. By Brouwer-Schauder fixed point theorem [118], the fixed point problem $\beta = \Phi(\Gamma(\beta))$ has a solution $\beta^* \in \mathcal{B}$. Therefore, $(\beta^*, \Gamma(\beta^*))$ is a system equilibrium. □

3.2.2 The Uniqueness of the Equilibrium

The uniqueness is proved by contradiction. We first state some properties of the probability $\beta^*(\cdot)$ of entering service and the abandonment probability $q^*(\cdot)$ in equilibrium. We then assume that there exist multiple equilibria. We explore the properties of the difference of two different equilibria. The properties of the difference contradict the other properties of $\beta^*(\cdot)$ and $q^*(\cdot)$ in equilibrium alluded to immediately above.

Intuitively, as a customer waits in the queue, there remains fewer customers ahead of her both due to service completions and abandonments. Thus, her probability of entering service in the next period increases with waiting time. Also, note that for a customer at the head of the queue, $\beta = b$, which is formalized in the next corollary.

Corollary 4. *In equilibrium, the probability $\beta^*(w)$ of entering service (after waiting for w periods) is increasing in w and satisfies inequality* (3.5)*. Moreover, we have that*

$$\lim_{w\to\infty} \beta^*(w) = b.$$

The next lemma shows the monotonicity of the expected value of waiting and abandonment probability in equilibrium.

Lemma 3. *In equilibrium, the expected discounted utility of waiting $J^*(w)$ is increasing whereas the abandonment probability $q^*(w)$ is decreasing in the waiting time w.*

Since the expected value of waiting J^* is increasing and bounded above by r, it converges as w tends to infinity. Hence, we have the following corollary.

Corollary 5. *There exist constants $J_\infty \leq r$ and $q_\infty \geq \underline{q}$ such that*

$$\lim_{w\to\infty} J^*(w) = J_\infty \quad \textit{and} \quad \lim_{w\to\infty} q^*(w) = q_\infty,$$

where J_∞ is the unique fixed point of following equation:

$$x = \mathbb{E}_\varepsilon[-c + \alpha[br + (1-b)x] - (\varepsilon(1) - \varepsilon(0))]^+, \tag{3.7}$$

and $q_\infty = \bar{F}(-c + \alpha[br + (1-b)J_\infty])$.

To construct the contradiction, we assume that there exist at least two different equilibria. Suppose e_1^* and e_2^* are two different equilibria where $e_i^*(w) = (\beta_i^*(w), q_i^*(w))$ for $w \geq 1$ and $i = 1, 2$. Define

$$\delta_\beta(w) = \beta_1^*(w) - \beta_2^*(w),\ \delta_q(w) = q_1^*(w) - q_2^*(w) \ \text{ and } \ \delta_{\bar{G}}(w) = \bar{G}_1^*(w) - \bar{G}_2^*(w).$$

Recall from (3.1) that $\bar{G}_i^*(w) = \prod_{k=1}^w (1 - q_i^*(k))$ for $w \geq 1$ and $i = 1, 2$. It is immediate from Corollaries 4 and 5 that

$$\lim_{w\to\infty} \delta_\beta(w) = 0 \ \text{ and } \ \lim_{w\to\infty} \delta_q(w) = 0. \tag{3.8}$$

To reach a contradiction, we show that (3.8) cannot hold. Lemmas 4-7 facilitate this contradiction argument. The following lemma proves an auxliary relationship between $\delta_{\bar{G}}(w)$, $\delta_\beta(w)$ and $\delta_q(w)$, which plays a key role in proving Lemmas 5-6.

Lemma 4. *There exist two positive sequences $g_1(w)$ and $g_2(w)$ for $w \geq 1$ with $\lim_{w\to\infty} g_i(w) = 0$ for $i = 1, 2$ such that*

$$\delta_{\bar{G}}(w) = g_1(w)\delta_\beta(w) + g_2(w)\delta_q(w). \tag{3.9}$$

The proof of Lemma 4 is provided in Appendix A.2[4].

The following lemma shows that e_1^* and e_2^* must be everywhere different.

4. Here is a brief outline of the proof in Appendix A.2. To prove this result, we define an auxiliary function $f_w(\cdot)$ in Appendix A.2.1 implicitly and study its properties (especially the monotonicity and convergence of its partial derivatives as w gets large). This function helps characterize $\bar{G}$ in terms of β and q. We then apply the mean value theorem to $f_w(\cdot)$ to establish the result in Lemma 4.

Lemma 5. *If $e_1^* \neq e_2^*$ are two different equilibria, then $e_1^*(w) \neq e_2^*(w)$ for all $w \geq 1$.*

Proof. We prove this lemma by contradiction. Suppose this is not true. Thus, there exists n such that $e_1^*(n) = e_2^*(n)$. We first show that $e_1^*(w) = e_2^*(w)$ for $w < n$. We then use the assumption $e_1^*(n) = e_2^*(n)$ to show that $e_1^*(w) = e_2^*(w)$ for $w > n$. Combining these two results, we conclude that this contradicts to the assumption that $e_1^* \neq e_2^*$.

It follows from equation (3.9) that $\delta_{\bar{G}}(n) = 0$. In other words, $\bar{G}_1^*(n) = \bar{G}_2^*(n)$. Note that $e_i^*(w)$ and $\bar{G}_i^*(w)$ for $i = 1, 2$ and $w \leq n$ can be computed recursively by equations (3.4), (2.10)-(2.11) and the following equation:

$$\bar{G}_i^*(w-1) = \frac{\bar{G}_i^*(w)}{1 - q_i^*(w)}, w \leq n.$$

Thus, we have that

$$e_1^*(w) = e_2^*(w) \text{ and } \bar{G}_1^*(w) = \bar{G}_2^*(w), \ w = 1, \ldots, n.$$

In addition, by inverting equations (3.1), (3.4) and (2.10)-(2.11), we can calculate $\beta(w), q(w)$ and $\bar{G}(w)$ for $w > n$ by the following equations recursively: For $w \geq n$,

$$q(w+1) = \bar{F}\left[H^{-1}\left(\frac{\bar{F}^{-1}(q(w)) + c - \alpha\beta(w)r}{\alpha(1-\beta(w))}\right)\right], \tag{3.10}$$

$$\bar{G}(w+1) = \bar{G}(w)(1 - q(w+1)), \tag{3.11}$$

$$\beta(w+1) = \frac{1-b}{1 - a\bar{G}(w+1)} \frac{\beta(w)}{1-\beta(w)}, \tag{3.12}$$

where $H(x) = \int_{-\infty}^{x} F(y)\,\mathrm{d}y$ for $x \geq 0$. Thus, we have that $e_1^*(w) = e_2^*(w)$ and $\bar{G}_1^*(w) = \bar{G}_2^*(w)$ for all $w > n$ as well. This gives $e_1^* = e_2^*$, which contradicts to the assumption that $e_1^* \neq e_2^*$. □

To facilitate the analysis to follow, let $\mathcal{M}_n$ denote the set of all $n \times n$ matrices, $x = (x_1, \ldots, x_n)'$ be a n-dimensional vector, and $M = [m_{ij}] \in \mathcal{M}_n$. Also let $||\cdot||_\infty$ and $|||\cdot|||_\infty$

denote the vector and matrix norms, respectively, given by

$$||x||_\infty = \max_{1\le i\le n} |x_i| \text{ and } |||M|||_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |m_{ij}|.$$

Building on Lemma 4, the next lemma characterizes the evolution of $(\delta_q(w), \delta_\beta(w))'$ as w evolves.

Lemma 6. *There exist a matrix A and a sequence of matrices $B(w)$, $w \ge 1$ such that the following holds:*

1. *The sequence of vectors $(\delta_q(w), \delta_\beta(w))'$ for $w \ge 1$ satisfy the following recursive equation:*
$$\begin{bmatrix} \delta_q(w+1) \\ \delta_\beta(w+1) \end{bmatrix} = (A + B(w)) \begin{bmatrix} \delta_q(w) \\ \delta_\beta(w) \end{bmatrix}. \tag{3.13}$$
2. *The two eigenvalues of the matrix A, denoted by λ_1 and λ_2, satisfy $\lambda_1 > \lambda_2 > 1$.*
3. $\lim_{w\to\infty} |||B(w)|||_\infty = 0$.

Proof. To facilitate the proof, define the constants a_1, a_2, a_3 as follows:

$$a_1 = \frac{1}{\alpha(1-q_\infty)(1-b)},\ a_2 = \frac{f(\bar{F}^{-1}(q_\infty))(\bar{F}^{-1}(q_\infty) + c - \alpha\beta r)}{\alpha(1-q_\infty)(1-\beta)^2} \text{ and } a_3 = \frac{1}{1-b}.$$

In addition, define the matrix A as follows:

$$A = \begin{bmatrix} a_1 & a_2 \\ 0 & a_3 \end{bmatrix}.$$

Let $[\varepsilon_q(w+1), \varepsilon_\beta(w+1)]^T$ denote the difference of two vectors given as follows:

$$\begin{bmatrix} \varepsilon_q(w+1) \\ \varepsilon_\beta(w+1) \end{bmatrix} = \begin{bmatrix} \delta_q(w+1) \\ \delta_\beta(w+1) \end{bmatrix} - A \begin{bmatrix} \delta_q(w) \\ \delta_\beta(w) \end{bmatrix},\ w \ge 1. \tag{3.14}$$

In addition, define the matrices $B(w)$ as follows: For $w \geq 1$,

$$B(w) = \begin{bmatrix} \dfrac{\varepsilon_q(w+1)}{|\delta_q(w)| + |\delta_\beta(w)|}\text{sign}(\delta_q(w)) & \dfrac{\varepsilon_q(w+1)}{|\delta_q(w)| + |\delta_\beta(w)|}\text{sign}(\delta_\beta(w)) \\ \dfrac{\varepsilon_\beta(w+1)}{|\delta_q(w)| + |\delta_\beta(w)|}\text{sign}(\delta_q(w)) & \dfrac{\varepsilon_\beta(w+1)}{|\delta_q(w)| + |\delta_\beta(w)|}\text{sign}(\delta_\beta(w)) \end{bmatrix},$$

where $\text{sign}(x)$ is the sign of x. It is easy to check that the matrices A and $B(w)$ satisfy equation (3.13).

We complete the proof by showing that $\lim_{w\to\infty} |||B(w)|||_\infty = 0$. That is, for any $\epsilon > 0$, there exists w_0 such that for all $w \geq w_0$, $|||B(w)|||_\infty \leq \epsilon$. Note that this is equivalent to show that for any $\epsilon > 0$, there exists w_0 such that for all $w \geq w_0$,

$$|\varepsilon_q(w+1)| \leq \epsilon(|\delta_q(w)| + |\delta_\beta(w)|) \text{ and } |\varepsilon_\beta(w+1)| \leq \epsilon(|\delta_q(w)| + |\delta_\beta(w)|). \tag{3.15}$$

The rest of the proof shows (3.15). Recall that $H(x) = \int_{-\infty}^{x} F(y)\,\mathrm{d}y$ for $x \geq 0$. To facilitate the proof, define functions $\psi_1(\cdot,\cdot)$ and $\psi_2(\cdot,\cdot)$ as follows:

$$\begin{aligned} \psi_1(\beta, q) &= \bar{F}\left[H^{-1}\left(\frac{\bar{F}^{-1}(q) + c - \alpha\beta r}{\alpha(1-\beta)}\right)\right], \\ \psi_2(\beta, \bar{G}) &= \frac{1-b}{1-a\bar{G}}\frac{\beta}{1-\beta}, \end{aligned}$$

The partial derivatives of the functions $\psi_1(\cdot,\cdot)$ and $\psi_2(\cdot,\cdot)$ are given as follows:

$$\begin{aligned} \frac{\partial\psi_1(\beta,q)}{\partial\beta} &= -\frac{f(\bar{F}^{-1}(\psi_1(\beta,q)))}{1-\psi_1(\beta,q)}\frac{\bar{F}^{-1}(q) + c - \alpha\beta r}{\alpha(1-\beta)^2}, \\ \frac{\partial\psi_1(\beta,q)}{\partial q} &= \frac{f(\bar{F}^{-1}(\psi_1(\beta,q)))}{f(\bar{F}^{-1}(q))\alpha(1-\beta)(1-\psi_1(q,\beta))}, \\ \frac{\partial\psi_2(\beta,\bar{G})}{\partial\beta} &= \frac{1-b}{1-a\bar{G}}\frac{1}{(1-\beta)^2}, \\ \frac{\partial\psi_2(\beta,\bar{G})}{\partial\bar{G}} &= \frac{(1-b)\beta}{1-\beta}\frac{a}{(1-a\bar{G})^2}. \end{aligned}$$

It follows from equations (3.10) - (3.12) that for $w \geq 1$ and $i = 1, 2$,

$$q_i^*(w+1) = \psi_1(q_i^*(w), \beta_i^*(w)) \text{ and } \beta_i^*(w+1) = \psi_2(\beta_i^*(w), \bar{G}_i^*(w+1)).$$

By the mean value theorem [Theorem 8.4, 7], we have that (for $w \geq 1$)

$$\delta_q(w+1) = \frac{\partial \psi_1(\beta_1(w), q_1(w))}{\partial q}\delta_q(w) + \frac{\partial \psi_1(\beta_1(w), q_1(w))}{\partial \beta}\delta_\beta(w), \tag{3.16}$$

$$\delta_\beta(w+1) = \frac{\partial \psi_2(\beta_2(w), \bar{G}_2(w+1))}{\partial \bar{G}}\delta_{\bar{G}}(w+1) + \frac{\partial \psi_2(\beta_2(w), \bar{G}_2(w+1))}{\partial \beta}\delta_\beta(w), \tag{3.17}$$

where

$$(\beta_1(w), q_1(w)) = c_1(w)(\beta_1^*(w), q_1^*(w)) + (1 - c_1(w))(\beta_2^*(w), q_2^*(w)),$$
$$(\beta_2(w), \bar{G}_2(w+1)) = c_2(w)(\beta_1^*(w), \bar{G}_1^*(w+1)) + (1 - c_2(w))(\beta_2^*(w), \bar{G}_2^*(w+1))$$

for some $c_1(w), c_2(w) \in (0, 1)$. It follows from Corollaries 4 and 5 that for $i = 1, 2$,

$$\beta_i^*(w) \to b,\ q_i^*(w) \to q_\infty \text{ and } \bar{G}_i^*(w) \to 0 \text{ as } w \to \infty.$$

Since $(\beta_1(w), q_1(w))$ and $(\beta_2(w), \bar{G}_2(w+1))$ are convex combinations of the equilibrium quantities, the following holds:

$$(\beta_1(w), q_1(w)) \to (b, q_\infty) \text{ and } (\beta_2(w), \bar{G}_2(w+1)) \to (b, 0) \text{ as } w \to \infty.$$

Then it follows from the continuity of the partial derivatives of $\psi_1(\cdot)$ and $\psi_2(\cdot)$ that

$$\begin{aligned}
\lim_{w\to\infty} \frac{\partial\psi_1(\beta_1(w), q_1(w))}{\partial q} &= \frac{\partial\psi_1(b, q_\infty)}{\partial q} = a_1,\\
\lim_{w\to\infty} \frac{\partial\psi_1(\beta_1(w), q_1(w))}{\partial \beta} &= \frac{\partial\psi_1(b, q_\infty)}{\partial \beta} = a_2,\\
\lim_{w\to\infty} \frac{\partial\psi_2(\beta_2(w), \bar{G}_2(w+1))}{\partial \bar{G}} &= \frac{\partial\psi_2(b, 0)}{\partial \bar{G}} = ab,\\
\lim_{w\to\infty} \frac{\partial\psi_2(\beta_2(w), \bar{G}_2(w+1))}{\partial \beta} &= \frac{\partial\psi_2(b, 0)}{\partial \beta} = a_3.
\end{aligned}$$

Combining these with (3.16)-(3.17), we conclude that for any $\epsilon > 0$, there exits w_1 such that for $w \geq w_1$,

$$\left|\delta_q(w+1) - a_1\delta_q(w) - a_2\delta_\beta(w)\right| \leq \epsilon(|\delta_q(w)| + |\delta_\beta(w)|), \tag{3.18}$$

$$\left|\delta_\beta(w+1) - ab\delta_{\bar{G}}(w+1) - a_3\delta_\beta(w)\right| \leq \epsilon(|\delta_q(w)| + |\delta_\beta(w)|). \tag{3.19}$$

In particular, combining (3.18) with (3.14) yields that

$$|\varepsilon_q(w+1)| = |\delta_q(w+1) - a_1\delta_q(w) - a_2\delta_\beta(w)| \leq \epsilon(|\delta_q(w)| + |\delta_\beta(w)|),$$

which gives the first inequality in (3.15).

To complete the proof, we now focus on the second inequality in (3.15). By Lemma 4, there exists $w_2 \geq w_1$ such that

$$|\delta_{\bar{G}}(w)| \leq \frac{\epsilon}{2}\left(|\delta_q(w)| + |\delta_\beta(w)|\right) \quad \text{for} \quad w \geq w_2. \tag{3.20}$$

In addition, note that

$$\begin{aligned}
|\delta_q(w+1)| &= \left|\delta_q(w+1) - a_1\delta_q(w) - a_2\delta_\beta(w) + a_1\delta_q(w) + a_2\delta_\beta(w)\right| \\
&\leq \left|\delta_q(w+1) - a_1\delta_q(w) - a_2\delta_\beta(w)\right| + \left|a_1\delta_q(w) + a_2\delta_\beta(w)\right| \\
&\leq \epsilon(|\delta_q(w)| + |\delta_\beta(w)|) + a_1|\delta_q(w)| + a_2|\delta_\beta(w)| \\
&\leq M_1(|\delta_q(w)| + |\delta_\beta(w)|),
\end{aligned} \tag{3.21}$$

where $M_1 = \max\{a_1, a_2\}+\epsilon$ and the second inequality follows from (3.18). Since $\bar{G}_1^*(w+1) \to 0$ as $w \to \infty$, there exists $w_3 \geq w_2$ such that

$$\bar{G}_1^*(w+1) \leq \frac{\epsilon}{2M_1} \quad \text{for} \quad w \geq w_3. \tag{3.22}$$

Combining equations (3.21)-(3.22), we have the following inequality:

$$|\bar{G}_1^*(w+1)\delta_q(w+1)| \leq \frac{\epsilon}{2}(|\delta_q(w)| + |\delta_\beta(w)|) \quad \text{for} \quad w \geq w_3. \tag{3.23}$$

Thus, by substituting (3.11) into the definition of $\delta_{\bar{G}}$, we have that for $w \geq w_3$,

$$\begin{aligned}
|\delta_{\bar{G}}(w+1)| =& |\bar{G}_1^*(w+1) - \bar{G}_2^*(w+1)| \\
=& |\bar{G}_1^*(w)(1 - q_1^*(w+1)) - \bar{G}_2^*(w)(1 - q_2^*(w+1))| \\
=& |\bar{G}_1^*(w)(1 - q_1^*(w+1)) - \bar{G}_1^*(w)(1 - q_2^*(w+1)) \\
&+ \bar{G}_1^*(w)(1 - q_2^*(w+1)) - \bar{G}_2^*(w)(1 - q_2^*(w+1))| \\
=& |-\bar{G}_1^*(w)\delta_q(w+1) + (1 - q_2^*(w+1))\delta_{\bar{G}}(w)| \\
\leq& |\bar{G}_1^*(w)\delta_q(w+1)| + |(1 - q_2^*(w+1))\delta_{\bar{G}}(w)| \\
\leq& |\bar{G}_1^*(w)\delta_q(w+1)| + |\delta_{\bar{G}}(w)| \\
\leq& \frac{\epsilon}{2}(|\delta_q(w)| + |\delta_\beta(w)|) + \frac{\epsilon}{2}(|\delta_q(w)| + |\delta_\beta(w)|) \\
=& \epsilon(|\delta_q(w)| + |\delta_\beta(w)|).
\end{aligned} \tag{3.24}$$

The first inequality simply follows from $|-a+b| \leq |a|+|b|$ for any values a, b. The second inequality holds because $|1-q_2^*(w+1)| \leq 1$. The last inequality follows from equations (3.20) and (3.23). In summary, we can rewrite (3.24) as follows: For $w \geq w_3$,

$$|\delta_{\bar{G}}(w+1)| \leq \epsilon(|\delta_q(w)|+|\delta_\beta(w)|).$$

Then we observe the following for $w \geq w_3$,

$$\begin{aligned}\left|\delta_\beta(w+1)-a_3\delta_\beta(w)\right| &= \left|\delta_\beta(w+1)-a_3\delta_\beta(w)-ab\delta_{\bar{G}}(w+1)+ab\delta_{\bar{G}}(w+1)\right| \\ &\leq \epsilon(|\delta_q(w)|+|\delta_\beta(w)|)+ab|\delta_{\bar{G}}(w+1)| \\ &\leq 2\epsilon(|\delta_q(w)|+|\delta_\beta(w)|),\end{aligned} \tag{3.25}$$

where the first inequality follows from (3.19) while the last inequality follows because $ab \leq 1$. Thus, it follows from equations (3.14) and (3.25) that

$$|\varepsilon_\beta(w+1)| = |\delta_\beta(w+1)-a_3\delta_\beta(w)| \leq 2\epsilon(|\delta_q(w)|+|\delta_\beta(w)|),$$

which gives the second inequality in (3.15). □

The following technical lemma facilitates the proof of uniqueness.

Lemma 7. *Let $x(w)$ for $w \geq 1$ be a sequence of vectors in $\mathbb{R}^n$ with $x(w) \neq 0$ for all $w \geq 1$ such that*

$$x(w+1) = (A+B(w))x(w), \tag{3.26}$$

where $A \in \mathcal{M}_n$ with eigenvalues $\lambda_1 > \cdots > \lambda_n > 1$ and $B(w) \in \mathcal{M}_n$ with $\lim_{w\to\infty} |||B(w)|||_\infty = 0$. Then $x(w)$ cannot converge to zero, i.e. $x(w) \not\to 0$ as $w \to \infty$.

Proof. We can write (3.26) equivalently as

$$x(w) = (A+B(w))^{-1}x(w+1)$$

for w large enough because $A+B(w)$ is invertible for large w. This is because the eigenvalues of A are larger than one, and $B(w)$ is negligible for large w. Defining

$$C(w) = (A + B(w))^{-1} - A^{-1}, \tag{3.27}$$

we first show that

$$|||C(w)|||_\infty \to 0 \ \text{ as } \ w \to \infty.$$

Note that for $w \geq 1$,

$$\begin{aligned} C(w) =& - A^{-1} + (A + B(w))^{-1} \\ =& - A^{-1} \left[I - A(A + B(w))^{-1} \right] \\ =& - A^{-1} \left[(A + B(w))(A + B(w))^{-1} - A(A + B(w))^{-1} \right] \\ =& - A^{-1}(A + B(w) - A)(A + B(w))^{-1} \\ =& - A^{-1}B(w)(A + B(w))^{-1}. \end{aligned} \tag{3.28}$$

Therefore, the following holds: For $w \geq 1$,

$$\begin{aligned} |||C(w)|||_\infty =& |||A^{-1}B(w)(A + B(w))^{-1}|||_\infty \\ \leq& |||A^{-1}|||_\infty |||B(w)|||_\infty |||(A + B(w))^{-1}|||_\infty. \end{aligned} \tag{3.29}$$

It follows from (3.28) that $(A + B(w))^{-1} = A^{-1} - A^{-1}B(w)(A + B(w))^{-1}$. Therefore, we have that (for $w \geq 1$)

$$\begin{aligned} |||(A + B(w))^{-1}|||_\infty \leq& |||A^{-1}|||_\infty + |||A^{-1}B(w)(A + B(w))^{-1}|||_\infty \\ \leq& ||A^{-1}|||_\infty + |||A^{-1}|||_\infty |||B(w)|||_\infty |||(A + B(w))^{-1}|||_\infty. \end{aligned}$$

By rearranging the terms, we obtain that (for $w \geq 1$)

$$|||(A + B(w))^{-1}|||_\infty \leq \frac{|||A^{-1}|||_\infty}{1 - |||A^{-1}|||_\infty |||B(w)|||_\infty}.$$

Substituting this inequality into (3.29), we have that (for $w \geq 1$)

$$|||C(w)|||_\infty \leq |||A^{-1}|||_\infty \frac{|||A^{-1}|||_\infty}{1 - |||A^{-1}|||_\infty |||B(w)|||_\infty} |||B(w)|||_\infty.$$

Since $|||B(w)|||_\infty \to 0$, we conclude that $|||C(w)|||_\infty \to 0$ as $w \to \infty$.

Next, we show that $x(w) \not\to 0$ by contradiction. Suppose that $x(w) \to 0$. Note that the eigenvalues of A^{-1} are $\lambda_1^{-1} < \cdots < \lambda_n^{-1} < 1$. Thus, A^{-1} is diagonalizable and there exists a matrix S (of the eigenvectors) such that $A^{-1} = S^{-1}\Lambda S$ where Λ is a diagonal matrix with diagonal entries $\lambda_1^{-1}, \ldots, \lambda_n^{-1}$. Or equivalently, we can write that $\Lambda = SA^{-1}S^{-1}$. Define $||\cdot||_S$ and $|||\cdot|||_S$ as follows:

$$||x||_S = ||Sx||_\infty \quad \text{and} \quad |||M|||_S = |||SMS^{-1}|||_\infty.$$

Therefore, the following holds: For $w \geq 1$,

$$\begin{aligned}
|||A^{-1} + C(w)|||_S =& |||S(A^{-1} + C(w))S^{-1}|||_\infty \\
=& |||\Lambda + SC(w)S^{-1}|||_\infty \\
\leq& |||\Lambda|||_\infty + |||SC(w)S^{-1}|||_\infty \\
\leq& \lambda_n^{-1} + |||S|||_\infty |||S^{-1}|||_\infty |||C(w)|||_\infty,
\end{aligned}$$

where the last inequality follows from $|||\Lambda|||_\infty = \lambda_n^{-1} < 1$. Since $|||C(w)|||_\infty \to 0$, the second term on the right-hand side tends to zero as $w \to \infty$. Thus, there exists w_0 and $\epsilon > 0$ such that for $w \geq w_0$,

$$||||A^{-1} + C(w)|||_S \leq \lambda_n^{-1} + \epsilon < 1.$$

Therefore, it follows from (3.27) that for $w \geq w_0$,

$$\begin{aligned}
||x(w)||_S =& ||(A + B(w))^{-1} x(w+1)||_S \\
=& ||(A^{-1} + C(w))x(w+1)||_S \\
=& ||S(A^{-1} + C(w))S^{-1} S x(w+1)||_\infty \\
\leq& |||S(A^{-1} + C(w))S^{-1}|||_\infty ||Sx(w+1)||_\infty \\
=& |||A^{-1} + C(w)|||_S ||x(w+1)||_S \\
\leq& (\lambda_n^{-1} + \epsilon)||x(w+1)||_S.
\end{aligned}$$

We proceed by contradiction. Suppose $x(w) \to 0$ as $w \to \infty$. Denote $d = ||x(w_0)||_S$. Since $x(w_0) \neq 0$ by assumption, it holds that $d > 0$. Moreover, since $x(w_0 + n) \to 0$ as $n \to \infty$, there exists n_0 such that $||x(w_0 + n)||_S < d$ for all $n \geq n_0$. Therefore, the following holds:

$$d = ||x(w_0)||_S \leq (\lambda_n^{-1} + \epsilon)^n ||x(w_0 + n)||_S \leq (\lambda_n^{-1} + \epsilon)^n d < d.$$

This contradiction shows that $x(w)$ cannot converge to zero as $w \to \infty$. □

It is immediate from Lemma 6 that our problem satisfies the conditions of Lemma 7 (for $n = 2$). This observation facilitates the uniqueness proof; see proof of Proposition 9. Next we state the uniqueness result whose proof follows from Corollaries 4 and 5 and Lemmas 4-7.

Proposition 4. *The system equilibrium e^* is unique.*

Proof. We prove the uniqueness by contradiction. Suppose e_1^* and e_2^* are two different equilibria. On the one hand, by Corollaries 4 and 5, equation (3.8) holds. On the other hand, the two eigenvalues of matrix A in Lemma 6 are $(1-b)^{-1}$ and $[\alpha(1-q_\infty)(1-b)]^{-1}$, which are different and strictly greater than 1. Thus, it follows from Lemmas 5 and 6 that the sequence $(\delta_q(w), \delta_\beta(w))$ for $w \geq 1$ satisfies the two conditions in Lemma 7. By Lemma 7, either $\lim_{w\to\infty} \delta_\beta(w) \neq 0$ or $\lim_{w\to\infty} \delta_q(w) \neq 0$, which contradicts equation (3.8). □

3.3 An Algorithm to Compute the Equilibrium Numerically

This section provides an algorithm to compute the equilibrium. To calculate the equilibrium numerically, we introduce the notion of a truncated equilibrium in which the abandonment decisions are only partially endogenous. The abandonment probability of customers who have waited for more than N periods are given exogenously. They make abandonment decisions as if they had waited in the system indefinitely. Customers who have waited for less than N periods make their abandonment decisions endogenously. Formally, the truncated equilibrium is defined as follows:

Definition 3. *For $N \geq 1$, we call $e_N = (\beta_N, q_N)$ a truncated equilibrium if it satisfies the following conditions:*

1. *$\beta_N(w) = b$ and $q_N(w) = q_\infty$ for all $w \geq N$.*

2. *$\beta_N(w) = \Phi(q_N)(w)$ and $q_N(w) = \Gamma(\beta_N)(w)$ for all $w < N$,*

where $\Phi(\cdot)$ and $\Gamma(\cdot)$ are the mappings given in Definition 2.

Given N, the truncated equilibrium e_N is fully characterized if the values of $\beta_N(N)$, $q_N(N)$ and $\bar{G}_N(N)$ are known, where $\bar{G}_N(\cdot) = 1 - G_N(\cdot)$ and $G_N(\cdot)$ is the cdf induced by the abandonment probability $q_N(\cdot)$; see equation (3.1). To be specific, recall from equations (3.4) and (2.10)-(2.11) that the probability of entering service $\beta_N(w)$ and the abandonment probability $q_N(w)$ for all $w = 1, \ldots, N-1$, can be characterized by the following equations recursively:

$$\beta_N(w) = \left(1 + \frac{1-b}{1 - a\bar{G}_N(w+1)} \frac{1}{\beta_N(w+1)}\right)^{-1}, \tag{3.30}$$

$$q_N(w) = \bar{F}\left(-c + \alpha\left\{\beta_N(w)r + (1 - \beta_N(w))J_N(w+1)\right\}\right), \tag{3.31}$$

where

$$J_N(w+1) = \mathbb{E}_\varepsilon\left[\bar{F}^{-1}(q_N(w+1)) + \varepsilon(0) - \varepsilon(1)\right]^+, \tag{3.32}$$

and

$$\bar{G}_N(w) = \frac{\bar{G}_N(w+1)}{1 - q_N(w+1)}. \tag{3.33}$$

By the definition of the truncated equilibrium, the values of $\beta_N(N)$ and $q_N(N)$ are given exogenously. Thus, characterizing the truncated equilibrium is equivalent to determining the value of $\bar{G}_N(N)$. The following lemma shows that the truncated equilibrium is unique; see Appendix A.1.4 for its proof.

Lemma 8. *There exists a unique truncated equilibrium e_N for $N \geq 1$.*

Corollaries 4 and 5 suggest that the exogenous abandonments in the truncated equilibrium approximate the endogenous abandonment decisions in the (untruncated) equilibrium well for large N. The following proposition verifies this intuition and shows that the equilibrium can be approximated by the truncated one closely; see Appendix A.1.4 for its proof.

Proposition 5. *The truncated equilibrium e_N converges to the equilibrium e^* uniformly as $N \to \infty$.*

Thus, we use the truncated equilibrium to approximate the equilibrium e^*. Fixing the truncation period N, we next provide an algorithm to compute the truncated equilibrium e_N. As mentioned earlier, the term $\bar{G}_N(N)$ determines the truncated equilibrium through equations (3.30)-(3.33). Also note that we must have $\bar{G}_N(0) = 1$ by definition. The idea behind the algorithm is to start with a guess of $\bar{G}_N(N)$ and to recursively calculate $q_N(w)$, $\beta_N(w)$ and $\bar{G}_N(w)$ for $w < N$. If the guess of $\bar{G}(N)$ is correct, then the $\bar{G}(0)$ value calculated recursively must equal 1. Lemma 17 shows that $\bar{G}_N(0)$ is a monotone function of $\bar{G}_N(N)$ (see Appendix A.1.4 for its proof); and this observation leads to a simple algorithm.

Lemma 9. *If $\bar{G}^1_N(N) > \bar{G}^2_N(N)$, then $\bar{G}^1_N(0) > \bar{G}^2_N(0)$, where $\bar{G}^1_N(0)$ and $\bar{G}^2_N(0)$ are the values obtained from equations* (3.30)-(3.33) *recursively by substituting $\bar{G}_N(N) = \bar{G}^1_N(N)$ and $\bar{G}_N(N) = \bar{G}^2_N(N)$, respectively.*

By Lemma 17, if $\bar{G}_N(0) < 1$, the true value of $\bar{G}_N(N)$ is greater than the guessed value. So the initial guess must be increased. Otherwise, i.e. $\bar{G}_N(0) > 1$, we should lower the initial guess. This observation is key to the algorithm provided in Table 3.1.

Table 3.1: The algorithm for calculating the truncated equilibrium.

Algorithm 1: The truncated equilibrium in the single-class case.

Initialize: $G_N(N) \leftarrow g^0 \in (0,1)$ and $\bar{g} \leftarrow 1$ and $\underline{g} \leftarrow 0$.
Update the value of $\bar{G}_N(N)$**:**
while $\bar{g} - \underline{g} > \varepsilon$
    Calculate $\beta_N(\cdot)$, $q_N(\cdot)$ and $\bar{G}_N(\cdot)$ via equations (3.30)-(3.33).
    if $\bar{G}_N(0) = 1$
        stop
    else
        if $\bar{G}_N(0) > 1$
            $\bar{g} \leftarrow \bar{G}_N(N)$
        else
            $\underline{g} \leftarrow \bar{G}_N(N)$
        end if
    end if
    Pick $g \in (\underline{g}, \bar{g})$ and $\bar{G}_N(N) \leftarrow g$
end while

3.4 A Numerical Example

This section presents a numerical example to illustrate the effectiveness of the algorithm proposed in Section 3.3. We first compare the result from the numerical computation with the output of a simulation. In addition, we show the importance of modeling abandonments endogenously by comparing the predictions of the model with endogenous abandonments to those of the model with exogenous abandonments. We end this section by studying how the parameter changes impact the predictions of the system performance.

3.4.1 The Setup of the Numerical Example

Consider the Geo/Geo/1 queue in which customers make abandonment decisions to maximize their utility. The probability of arrival a equals to 0.5. In addition, the service rate b equals to 0.8. The per period waiting cost of customers is $c = 2$ and the reward from service is $r = 6$.

The idiosyncratic shocks $\varepsilon(0)$ and $\varepsilon(1)$ both follow the type I extreme value distribution, whose cumulative distribution function is given as follows[5]:

$$F_{\varepsilon(0)}(x) = F_{\varepsilon(1)}(x) = e^{e^{(-x)}},\ x \geq 0.$$

We refer this setting as the original system.

Suppose that the system then undergoes a change and the service rate is reduced to $b = 0.51$. To predict the new system performance, and in particular, the abandonment behavior for the new system, we approximate the equilibrium by the truncated one using the truncation period $N = 30$. In addition, in the model with exogenous abandonments, we assume that the distribution of the abandonment times remain the same as the one in the earlier system (with $b = 0.8$).

To compare the two models, we first simulate the system equilibrium corresponding to $b = 0.51$ iteratively. In the simulation, we start with using the equilibrium probability of entering service $\beta_N(\cdot)$ and the probability of abandoning $q_N(\cdot)$ computed via the algorithm given in Table 3.1. The simulation gives an empirical distribution of the VOWT. We use this empirical distribution as input and update the abandonment time distribution using the model of Section 2.3.2. We then simulate the system again with the updated distributions of the abandonment time and keep updating the distribution of the abandonment times and the VOWT until the simulation converges numerically.

Table 4.4 compares the means of the VOWT and abandonment time as well as the fraction of customers that abandon in equilibrium calculated from the simulation, the numerical computation and the exogenous model. In addition, Figure 3.1 shows the cumulative distribution function of the VOWT obtained from the simulation, the equilibrium computation and the exogenous model.

Both Table 4.4 and Figure 3.1 show that the exogenous model mistakenly predicts a longer

5. This distributional assumption is commonly made in models studying discrete consumer choice, cf. Anderson et al. [6]. Aksin et al. [3] also makes this assumption.

Table 3.2: The mean of the VOWT and abandonment time and the fractions of customers that abandon under the simulation, the equilibrium computation and the exogenous model ($a = 0.5$, $b = 0.51$, $c = 2$, $r = 6$). The numbers in the parentheses are the standard deviation of the statistics.

	Simulation	Eq. Computation	Error	Exogenous Model	Error
VOWT	2.537 (.055)	2.559	0.87%	4.232	66.8%
Mean abandonment time	5.68 (.063)	5.79	1.9%	13.79	142.7%
Percentage abandoning	42.50% (1.6%)	42.59%	0.2%	18.18%	57.22%

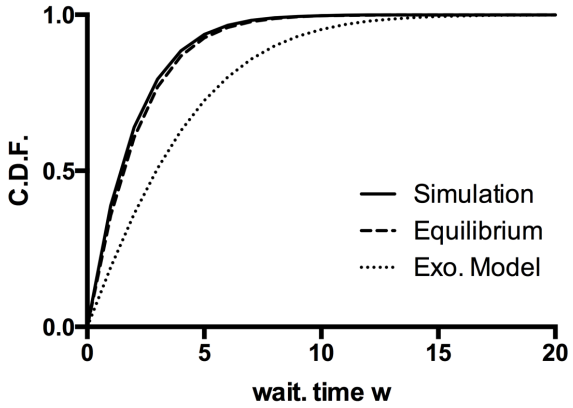


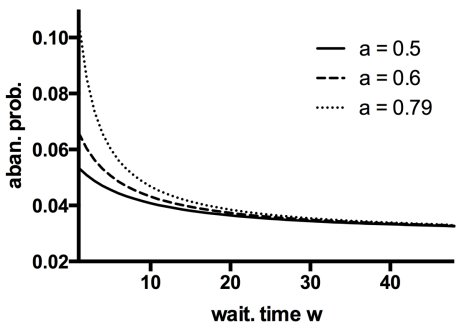
Figure 3.1: The cumulative distribution function of the VOWT with new service rate computed via the simulation, the equilibrium computation and the exogenous model ($a = 0.5$, $b = 0.51$, $c = 2$, $r = 6$).

waiting time and a lower probability of abandoning the queue. This is mainly because the model with exogenous abandonments ignores the impact of the service rate change on the abandonment behavior. Under the original service rate (which is higher), the customers are more patient because the probability of entering service is higher. When the service rate drops, the customers are more likely to abandon the system. However, the exogenous abandonment model does not capture this change in customers' behavior. In addition, the comparison of the simulation and the equilibrium computation shows that the proposed truncated equilibrium approximates the equilibrium well in all examples we tried. Thus, we only compare the predictions from the numerical computation of the equilibrium and the exogenous model in the rest of this section.

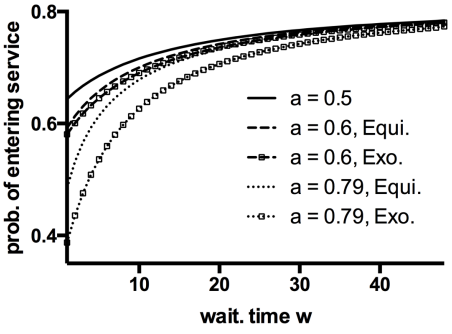
3.4.2 A Comparative Statics Analysis

We end the numerical study by a comparative statics analysis. To be specific, we study the impact of the changes of the arrival rate a and the service rate b on the predictions of the system performance.

Impact of the arrival rate. We first study the impact of the arrival rate a on the predictions. We keep the service rate b the same as the original system, i.e. $b = 0.8$, and increase the arrival rate a from 0.5 to 0.79 gradually. Figure 3.2 shows the numerical characterization of the system equilibrium for three different arrival rates. Figure 3.2a shows that the abandonment probability (as a function the waiting time) increases as the arrival rate increases. In other words, customers become more impatient when the system becomes more congested. Since the exogenous model assumes that the abandonment probability keeps unchanged, it fails to capture the change in customers' abandonment behavior. Figure 3.2b shows the probability of entering service in systems with different arrival rates. It shows that the probability of entering service $\beta(\cdot)$ decreases as the arrival rate increases, though customers are more likely to abandon. This is because the system becomes more congested when the arrival rate increases. Figure 3.2b also shows that the exogenous model underestimates the probability of entering service. A more comprehensive comparison between the equilibrium model and the exogenous model is given in Figure 5.1.

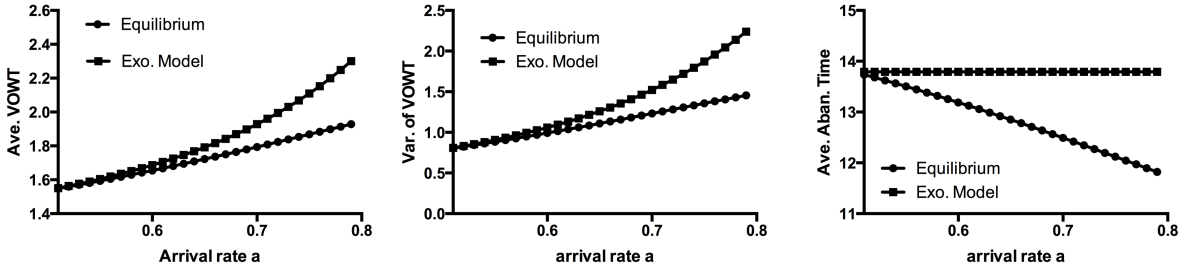


(a) The abandonment probability $q(w)$ as a function of the waiting time w

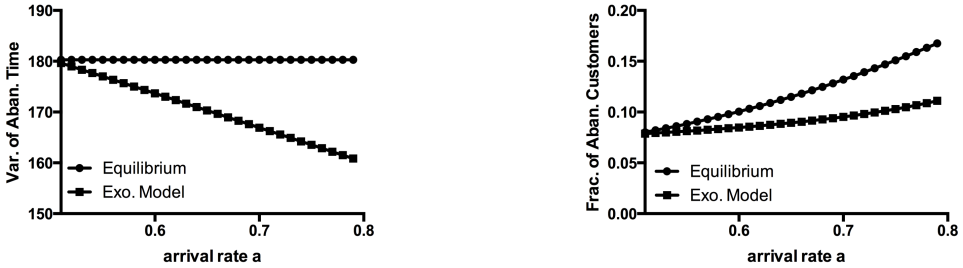


(b) The probability of entering service $\beta(w)$ as a function of the waiting time w

Figure 3.2: The system equilibrium under different arrival rates ($b = 0.8$, $c = 2$ and $r = 6$).



(a) The average VOWT (b) The variance of the VOWT (c) The average abandonment time



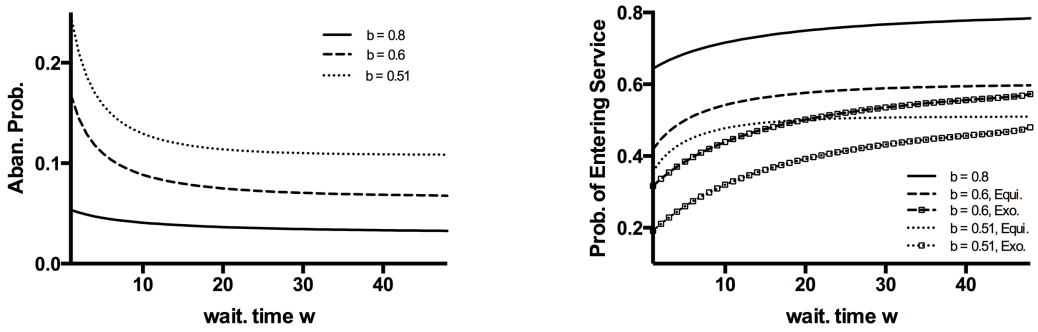
(d) The variance of the abandonment time (e) The fraction of customers abandoning

Figure 3.3: The predictions of the system performance under different arrival rates ($b = 0.8$, $c = 2$ and $r = 6$).

Figure 5.1 compares the predictions of the system performance from the equilibrium computation and the exogenous model. It shows that the prediction of the average VOWT from the exogenous model is mistakenly higher when the arrival rate is higher; see Figure 3.3a. In addition, Figure 3.3b shows that the variance of the VOWT predicted using the exogenous model is higher as well. This is because when the arrival rate is higher, the system becomes more congested. Thus, the customers are more likely to abandon in the more congested system; see Figures 3.3c-3.3e. The exogenous model ignores the change in customer's abandonment behavior. Therefore, it underestimates the abandonments and thus results in a higher prediction of the VOWT.

Impact of the service rate. Next, we study the impact of the service rate b on the predictions of the system performance using a similar method. To be specific, we keep the arrival rate a unchanged, i.e. $a = 0.5$, and reduce the service rate b from 0.8 to 0.51 gradually. Figure 3.4 shows the system equilibrium for different service rates b. Figure 3.5 compares the predictions of the system performance for the equilibrium computation and the exogenous

model. This study also shows that the predictions from the exogenous model is less accurate when the system is more congested, i.e. the service rate b is smaller.

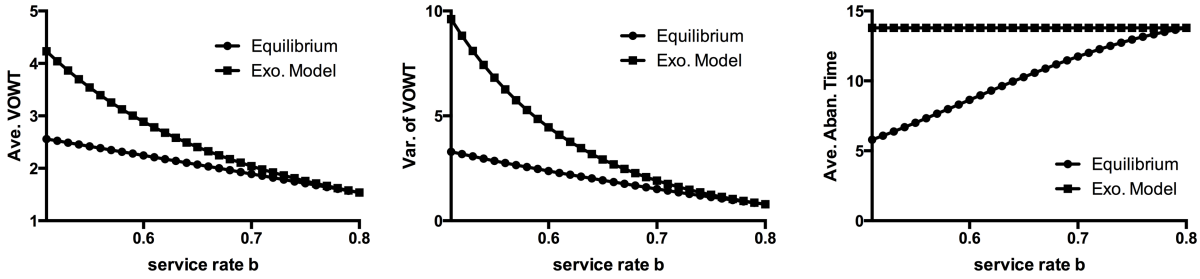


(a) The abandonment probability $q(w)$ as a function of the waiting time w

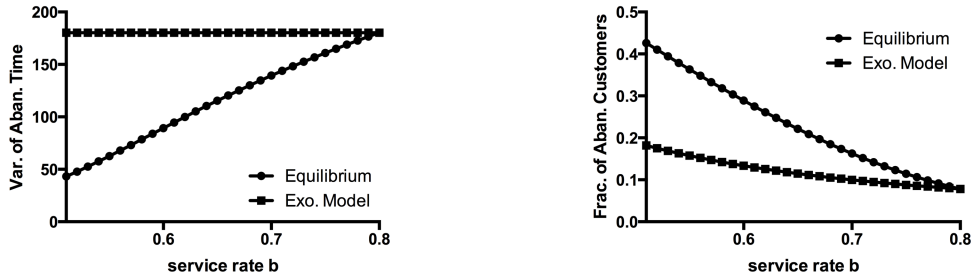
(b) The probability of entering service $\beta(w)$ as a function of the waiting time w

Figure 3.4: The system equilibrium under different service rates ($a = 0.5$, $c = 2$ and $r = 6$).

Impact of both arrival and service rates. The last comparative statics analysis compares the predictions from both the equilibrium computation and the exogenous model under different combinations of the arrival rates and the service rates. Figure 3.6 shows that the difference of the predictions from the two models is most significant when the arrival rate $a = 0.5$ and the service rate $b = 0.51$. This coincides with the scenario when the system is most congested; see Figure 3.6a. This study again emphasizes the importance on the abandonment assumption in congested systems.

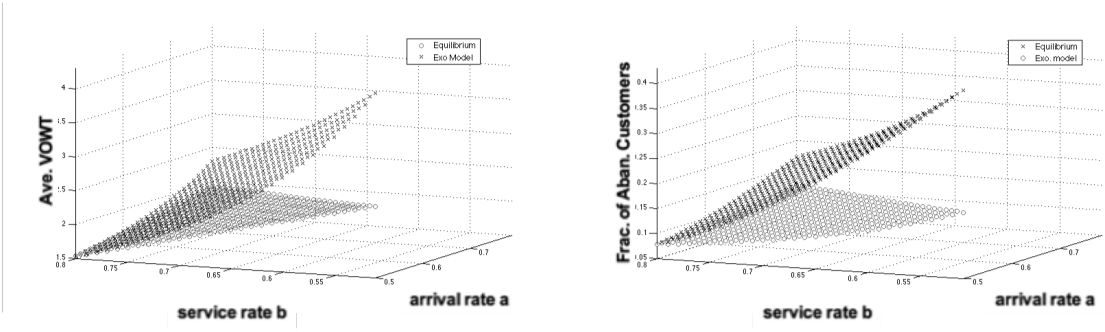


(a) The average VOWT (b) The variance of the VOWT (c) The average abandonment time



(d) The variance of the abandonment time (e) The fraction of customers abandoning

Figure 3.5: The predictions of the system performance under different service rates ($a = 0.5$, $c = 2$ and $r = 6$).



(a) The average VOWT (b) The fraction of customers abandoning

Figure 3.6: The predictions of the system performance under different arrival rates and service rates ($c = 2$ and $r = 6$).

CHAPTER 4

ASYMPTOTIC ANALYSIS OF THE MULTICLASS SYSTEM

4.1 Asymptotic Analysis of the Multiclass Queueing System

This section approximates the multiclass system by a diffusion model in the heavy traffic limit, and characterizes the equilibrium for the diffusion model. Section 4.1.1 introduces the heavy traffic regime and characterizes the system dynamics in the heavy traffic limit. By virtue of state-space collapse, Section 4.1.2 characterizes the equilibrium in the heavy traffic limit. Section 4.1.3 establishes the existence and uniqueness of the equilibrium.

4.1.1 An Asymptotic Analysis of the System Dynamics, the Virtual Offered Waiting Time and the Abandonment Probabilities in the Heavy Traffic Limit

The heavy traffic asymptotic regime involves a closely-related sequence of systems indexed by n, whose formal limit gives rise to a diffusion model of the underlying queueing system. The resulting diffusion model is analytically tractable and enables an approximate characterization of the original system. A superscript n is attached to quantities of interest associated with the n^{th} system. The arrival process of the n^{th} system is described through the sequence of i.i.d. vectors $\{(v^n(i), \kappa^n(i)) : i \geq 1\}$ where $v^n(i)$ and $\kappa^n(i)$ denote the interarrival time and the class, respectively, of the i^{th} arriving customer. For all n, the hazard rate function of $v^n(i)$ is $a(\cdot)$, and $\kappa^n(i)$ has a multinomial distribution with probabilities $(p_1(v^n(i)), \ldots, p_K(v^n(i)))$. In particular, we retain the same primitives $(a(t), p_k(t), k = 1, \ldots, K)$ for the arrival process (introduced in Section 2.3.1) across all members of the sequence of systems under consideration.

The service process for class k customers in the n^{th} system is defined through a sequence of i.i.d. random variables $\{s_k^n(i) : i \geq 1\}$. The service processes for the various classes are

mutually independent. The hazard rate function for $s_k^n(i)$ is denoted by $b_k^n(\cdot)$. We assume for $k = 1, \ldots, K$ that $b_k^n(\cdot) \to b_k(\cdot)$ as $n \to \infty$. Time is scaled in the n^{th} system such that periods are of length[1] $1/2^n$. In particular, the arrival rate (per unit of time) of class k in the n^{th} system is $\lambda_k^n = 2^n \lambda_k$. The traffic intensity of the n^{th} system is $\rho^n = \sum_{k=1}^K \rho_k^n$, where $\rho_k^n = \lambda_k^n/\mu_k^n$ is the offered load for class k and μ_k^n is the service rate for class k in the n^{th} system. By the assumption that $b_k^n(\cdot) \to b_k(\cdot)$ and the assumed time scaling, we expect that the service rate (per time unit) for class k in the n^{th} system is close to $2^n \mu_k$, which is made precise by the following heavy traffic assumption.

Assumption 3. *(The heavy traffic assumption). We assume that the following holds:*

1. *There exists θ_k such that $\sqrt{2^n}\left(\rho_k^n - \frac{\lambda_k}{\mu_k}\right) \to \theta_k$ as $n \to \infty$ for $k = 1, \ldots, K$.*
2. $\sum_{k=1}^K \frac{\lambda_k}{\mu_k} = 1$.

It is immediate from the heavy traffic assumption that $\sqrt{2^n}(\rho^n - 1) \to \sum_{k=1}^K \theta_k$ as $n \to \infty$. Under the heavy traffic assumption, we expect the queue lengths to be of order $\sqrt{2^n}$, whereas the delays to be of order $1/\sqrt{2^n}$ in the n^{th} system. Therefore, the reward from service r^n and the per-period delay cost c^n are scaled in the n^{th} system as follows[2]:

$$r_k^n = r_k \quad \text{and} \quad c_k^n = \frac{c_k}{\sqrt{2^n}} \quad \text{for} \quad k = 1, \ldots, K \quad \text{and} \quad n \geq 1. \tag{4.1}$$

Let $A^n(t)$ be the total number of customers arriving by time t across all classes and $A_k^n(t)$ be the total number of class k customers arriving by time t for $k = 1, \ldots, K$. In addition, denote $S_k^n(t)$ as the total number of class k customers served by time t if the server is continuously busy working on class k jobs during $[0, t]$. We use the sequences $\{(v^n(i), \kappa^n(i)) : i \geq 1\}$ and $\{s_k^n(i) : i \geq 1\}$ for $k = 1, \ldots, K$ to define the continuous-time

1. This dyadic partition of time leads to a nested period structure across the sequence of models considered.

2. Without scaling the per period delay cost in the n^{th} system would be $c_k/2^n$. Scaling this with $\sqrt{2^n}$ to account for the fact that the delays will be of order $1/\sqrt{2^n}$ gives (4.1).

arrival and service completion processes: For $t \geq 0$, $k = 1, \ldots, K$ and $n \geq 1$, define

$$A^n(t) = \sup\left\{ i \geq 0 : \frac{1}{2^n} \sum_{j=1}^{i} v^n(j) \leq t \right\}, \tag{4.2}$$

$$A_k^n(t) = \sum_{i=1}^{A^n(t)} \mathbb{1}_{\{\kappa^n(i)=k\}}, \tag{4.3}$$

$$S_k^n(t) = \sup\left\{ i \geq 0 : \frac{1}{2^n} \sum_{j=1}^{i} s_k^n(j) \leq t \right\}. \tag{4.4}$$

To describe the evolution of the queue length processes, let $R_k^n(t)$ denote the cumulative number of class k customers that have abandoned by time t for $k = 1, \ldots, K$. Also let $T^n(t) = (T_1^n(t), \ldots, T_K^n(t))$ denote the server's scheduling policy, where $T_k^n(t)$ is the amount of time the server dedicates to class k during $[0, t]$ for $k = 1, \ldots, K$. Let $I^n(t)$ denote the cumulative idleness incurred by the server during $[0, t]$. It is given by the following:

$$I^n(t) = t - \sum_{k=1}^{K} T_k^n(t), \ t \geq 0, \tag{4.5}$$

We assume that the system is empty initially. Then letting $Q_k^n(t)$ denote the number of class k jobs in the system at time t, it follows for each class $k = 1, \ldots, K$ that

$$Q_k^n(t) = A_k^n(t) - S_k^n(T_k^n(t)) - R_k^n(t), \ t \geq 0. \tag{4.6}$$

Define $Q^n = (Q_k^n)$ to be the vector-valued queue-length process. We restrict our attention

to non-preemptive and work-conserving policies that satisfy the following conditions:

$$T^n \text{ is non-anticipating with respect to } Q^n, \tag{4.7}$$

$$T^n \text{ is continuous and increasing with } T^n(0) = 0, \tag{4.8}$$

$$I^n \text{ is increasing with } I^n(0) = 0, \tag{4.9}$$

$$Q^n(t) \geq 0, \ t \geq 0. \tag{4.10}$$

The original problem of interest can be viewed as a specific element of this sequence of problems, determined by the particular choice of the parameter n. The underlying assumption of the heavy traffic approximation is that the system parameter corresponding to the original problem of interest is large enough that various (scaled) processes of the original system can be approximated by the corresponding processes of the diffusion model.

One arrives at the diffusion model by centering various processes around their means, rescaling them and passing to the formal limit as the scaling parameter gets large. In this vein, define the scaled arrival and service processes as follows: For $t \geq 0$ and $n \geq 1$,

$$\hat{A}_k^n(t) = \frac{A_k^n(t) - \lambda_k^n t}{\sqrt{2^n}} \ \text{ and } \ \hat{S}_k^n(t) = \frac{S_k^n(t) - \mu_k^n t}{\sqrt{2^n}}, \ k = 1, \ldots, K. \tag{4.11}$$

Similarly, for $t \geq 0$ and $n \geq 1$, define the scaled queue-length, abandonment and cumulative idleness processes as follows:

$$\hat{Q}_k^n(t) = \frac{Q_k^n(t)}{\sqrt{2^n}}, \ \hat{R}_k^n(t) = \frac{R_k^n(t)}{\sqrt{2^n}}, \ k = 1, \ldots, K, \tag{4.12}$$

$$\hat{I}^n(t) = \sqrt{2^n} I^n(t). \tag{4.13}$$

Then the dynamics of the scaled queue-length process are expressed as follows: For $k = 1, \ldots, K$ and $n \geq 1$,

$$\hat{Q}_k^n(t) = \hat{X}_k^n(t) - \hat{R}_k^n(t) + \frac{\mu_k^n}{2^n} \hat{I}_k^n(t), \tag{4.14}$$

where $\hat{X}_k^n(t) = \hat{A}_k^n(t) - \hat{S}_k^n(T_k^n(t)) + \sqrt{2^n}(\rho_k^n t - \rho_k t)$ and $\hat{I}_k^n(t) = \sqrt{2^n}(\rho_k t - T_k^n(t))$. By a straightforward application of the functional central limit theorem [e.g. 112], we conclude that as $n \to \infty$

$$(\hat{A}_1^n, \ldots, \hat{A}_K^n, \hat{S}_1^n, \ldots, \hat{S}_K^n) \Rightarrow (\lambda_1^{\frac{3}{2}}\sigma_1^A B_1^A, \ldots, \lambda_K^{\frac{3}{2}}\sigma_K^A B_K^A, \mu_1^{\frac{3}{2}}\sigma_1^S B_1^S, \ldots, \mu_K^{\frac{3}{2}}\sigma_K^S B_K^S), \quad (4.15)$$

where $\Rightarrow$ denotes weak convergence [28], and B_k^A and B_k^S for $k = 1, \ldots, K$, are independent standard Brownian motions. Recall that we restrict attention to work-conserving, non-preemptive scheduling policies. In addition, we assume that

$$(T_1^n(\cdot), \ldots, T_K^n(\cdot)) \Rightarrow (\rho_1(\cdot), \ldots, \rho_K(\cdot)) \quad \text{as} \quad n \to \infty, \quad (4.16)$$

where $\rho_k(t) = \rho_k t$ for $t \geq 0$ and $k = 1, \ldots, K$. We also assume for $k = 1, \ldots, K$ that

$$\hat{I}_k^n \Rightarrow \hat{I}_k \quad \text{as} \quad n \to \infty \quad (4.17)$$

for some process $\hat{I}_k$. Note that $\hat{I}^n(t) = \sum_{k=1}^K \hat{I}_k^n(t)$ for $t \geq 0$. Thus, equation (4.17) implies that $\hat{I}^n \Rightarrow \hat{I}$ as $n \to \infty$ where $\hat{I}(\cdot)$ is a non-decreasing process with $\hat{I}(0) = 0$. We also conclude from equations (4.15)-(4.16) and the heavy traffic assumption that

$$(\hat{X}_1^n, \ldots, \hat{X}_K^n) \Rightarrow (\hat{X}_1, \ldots, \hat{X}_K), \quad (4.18)$$

where $\hat{X}_k$ (for $k = 1, \ldots K$) are independent Brownian motions with drift parameters θ_k and variance parameters $\lambda_k^3(\sigma_k^A)^2 + \rho_k \mu_k^3 (\sigma_k^S)^2$.

The stochastic process $R_k^n(\cdot)$, the cumulative number of class k abandonments, is characterized by individual customers' abandonment decisions in system n. Letting $q_k^n(w)$ denote the probability of abandoning in the next period after waiting for w time units (where w takes values in $\{j/2^n : j \geq 1\}$), it is characterized by the framework advanced in Section

2.3.2; see Lemma **??** and Proposition 1. We focus our attention on the case when the abandonment process $R_n^k(\cdot)$ is of the same order of magnitude as the queue length $Q_k^n(\cdot)$, which is of order $1/\sqrt{2^n}$. To be more specific, we study the abandonment process $R_k^n(\cdot)$ under the hazard-rate scaling, introduced by Reed and Ward [93]. Under the hazard rate scaling, the entire abandonment time distribution is used crucially to formulate the limiting problem. Note that $q_k^n(\cdot)$ is also the per period hazard rate of the abandonment time. Thus, the hazard rate per unit of time is given by $2^n q_k^n(\cdot)$ in the n^{th} system. Applying the hazard rate scaling[3] used in Reed and Ward [93], we define the scaled hazard rate, denoted by $\hat{q}_k^n(\cdot)$, as follows[4]:

$$\hat{q}_k^n(w) = 2^n q_k^n \left(\frac{\lfloor \sqrt{2^n} w \rfloor}{2^n} \right), \ w \geq 0. \tag{4.19}$$

We assume that the (scaled) hazard rate function $\hat{q}_k^n(\cdot)$ converges as $n \to \infty$, i.e. there exists a function $\hat{q}_k(\cdot)$ such that $\hat{q}_k^n \to \hat{q}_k$ uniformly as $n \to \infty$.

Following Kim and Ward [69], the scaled process $\hat{R}_k^n(\cdot)$ is characterized by the scaled abandonment probability $\hat{q}_k^n(\cdot)$ approximately as follows:

$$\hat{R}_k^n(t) \approx \lambda_k \int_0^t \int_0^{\hat{Q}_k^n(s)/\lambda_k} \hat{q}_k^n(x) \, \mathrm{d}x \, \mathrm{d}s, \tag{4.20}$$

see Appendix B.1.1 for the derivation. Combining equations (4.14)-(4.18) and (4.20), we approximate $\hat{Q}_k^n$ by $\hat{Q}_k$, i.e. $(\hat{Q}_1^n, \ldots, \hat{Q}_K^n) \overset{D}{\approx} (\hat{Q}_1, \ldots, \hat{Q}_K)$ for n large, where (for $k = 1, \ldots, K$)

$$\hat{Q}_k(t) = \hat{X}_k(t) - \lambda_k \int_0^t \int_0^{\hat{Q}_k(s)/\lambda_k} \hat{q}_k(y) \, \mathrm{d}y \, \mathrm{d}s + \mu_k \hat{I}_k(t), \ t \geq 0. \tag{4.21}$$

We let $V_k^n(t)$ denote the virtual offered waiting time (VOWT) for class k in the n^{th}

3. Recall that the time is speeded up by a factor of 2^n in the n^{th} system.

4. The scaling of Reed and Ward [93] is as follows: $h^n(x) = h(\sqrt{n}x)$, where h is the limiting hazard rate function and h^n is the hazard rate function of the n^{th} system. This can equivalently be written as $h(y) = h^n(y/\sqrt{n})$. Using this would give $\hat{q}_k^n(w) = 2^n q_k^n(w/\sqrt{2^n})$. The only remaining step is to make sure that the argument of $q_k^n(\cdot)$ is a multiple of $1/2^n$, i.e. of the form $j/2^n$, which is the end of period j in the n^{th} system. Thus, we multiply $w/\sqrt{2^n}$ by 2^n, truncate and divide the value by 2^n. This gives (4.19).

system and define its scaled version for $k = 1, \ldots, K$ and $n \geq 1$ as follows:

$$\hat{V}_k^n(t) = \sqrt{2^n} V_k^n(t) \text{ for } t \geq 0.$$

By the snapshot principle [e.g. 94], the (scaled) virtual offered waiting time can be approximated as follows:

$$\hat{V}_k^n(t) \approx \frac{\hat{Q}_k^n(t)}{\lambda_k}, \ t \geq 0 \tag{4.22}$$

for large n. Assuming $\hat{V}_k^n \Rightarrow \hat{V}_k$ as $n \to \infty$ $(k = 1, \ldots, K)$, where $\hat{V}_k = \hat{Q}_k/\lambda_k$, we characterize the dynamics of the limiting virtual waiting time process $\hat{V}_k$ for class k as follows:

$$\hat{V}_k(t) = \frac{1}{\lambda_k}\hat{X}_k(t) - \int_0^t \int_0^{\hat{V}_k(s)} \hat{q}_k(y)\, \mathrm{d}y\, \mathrm{d}s + \frac{1}{\rho_k}\hat{I}_k(t), \ t \geq 0. \tag{4.23}$$

Let $\beta_k^n(w)$ denote the probability of entering service in the next period when a class k customer waits for w time units in steady state. In other words, $\beta_k^n(\cdot)$ is the per period hazard rate of (the steady state distribution of) the VOWT $V_k^n(\cdot)$. Thus, the hazard rate of the scaled VOWT $\hat{V}_k^n(\cdot)$ (per time unit), denoted by $\hat{\beta}_k^n(\cdot)$, is characterized by scaling $\beta_k^n(\cdot)$ as follows:

$$\hat{\beta}_k^n(w) = \sqrt{2^n}\beta_k^n\left(\frac{\lfloor\sqrt{2^n}w\rfloor}{2^n}\right), \ w \geq 0; \tag{4.24}$$

see Appendix B.1.1 for the derivation. We assume that as $n \to \infty$, $\hat{\beta}_k^n(\cdot)$ converges to $\hat{\beta}_k(\cdot)$, where $\hat{\beta}_k(\cdot)$ is the hazard rate of (the steady-state distribution of) the limiting VOWT $\hat{V}_k(\cdot)$. Thus, equation (4.23) characterizes the mapping from the limiting (scaled) abandonment rate $\hat{q}(\cdot)$ to the limiting (scaled) hazard rate $\hat{\beta}(\cdot)$, where $\hat{q} = (\hat{q}_k)$ and $\hat{\beta} = (\hat{\beta}_k)$.

The rest of this subsection characterizes the mapping from the hazard rate $\hat{\beta}_k(\cdot)$ of the (limiting) VOWT to the (limiting) abandonment rate $\hat{q}_k(\cdot)$. Recall that $F_k^n(\cdot)$ denotes the cdf of $\varepsilon_k^n(1) - \varepsilon_k^n(0)$ for $k = 1, \ldots, K$. We have not imposed any assumptions on how $F_k^n(\cdot)$

varies with n so far, i.e. how $\varepsilon_k^n(\cdot)$ scales with n. To this end, recall from (2.9) that

$$q_k^n(w) = \bar{F}_k^n\left(-c_k^n + \alpha^n\left[\beta_k^n(w) r_k^n + (1-\beta_k^n(w)) J_k^n(w+1)\right]\right),\ w \geq 0, \tag{4.25}$$

where $\alpha^n = (\alpha)^{1/2^n}$ is the discount factor for one period in the n^{th} system. Under scaling, we derive from equations (4.19) and (4.24) that

$$\hat{q}_k^n(w) = 2^n \bar{F}_k^n\left(-\frac{c_k}{\sqrt{2^n}} + \alpha^n\left[\frac{\hat{\beta}_k^n(w)}{\sqrt{2^n}} r_k + \left(1 - \frac{\hat{\beta}_k^n(w)}{\sqrt{2^n}}\right) J_k^n\left(\frac{\lfloor\sqrt{2^n}(w+1/\sqrt{2^n})\rfloor}{2^n}\right)\right]\right). \tag{4.26}$$

Throughout our analysis, we assume $F_k^n(\cdot)$ is such that $\hat{q}_k^n(\cdot)$ converges to a continuous function $\hat{q}_k(\cdot)$, $k = 1, \ldots, K$. Existence of the limit of $\hat{q}_k^n(\cdot)$ requires that the cdf $F_k^n(\cdot)$ converges at an appropriate rate. For this to happen, the value of the function $\bar{F}_k^n(\cdot)$ in equation (4.26) should be of order $1/2^n$. In particular, because we assume $\hat{\beta}_k^n(\cdot) \to \hat{\beta}_k(\cdot)$, we have that the value of the argument of the function $\bar{F}_k^n(\cdot)$ in equation (4.26) has the same order of magnitude as the expected discounted utility $J_k^n(\cdot)$. Moreover, $J_k^n(\cdot)$ has the same order of magnitude as r_k^n, which is of order 1. Therefore, we need the value of $\bar{F}_k^n(\cdot)$ be of order $1/2^n$ for the arguments of order 1 for $\hat{q}_k^n(\cdot)$ to converge.

To facilitate the definition of the equilibrium in heavy traffic[5], let Φ denote the mapping from $\hat{q}(\cdot)$ to $\hat{\beta}(\cdot)$ which is characterized by equation (4.23) and Γ_k denote the mapping from $\hat{\beta}_k(\cdot)$ to $\hat{q}_k(\cdot)$, which, for instance, is characterized by equation (4.32) under Assumption 6. Next, we provide the formal equilibrium definition.

Definition 4. *We say that $e^* = (\hat{\beta}^*, \hat{q}^*)$, where $\hat{\beta}^* = (\hat{\beta}_1^*, \ldots, \hat{\beta}_K^*)$ and $\hat{q}^* = (\hat{q}_1^*, \ldots, \hat{q}_K^*)$, is an equilibrium in the heavy traffic limit if the following conditions are both satisfied: $\hat{\beta}^* =$*

5. To speak of an equilibrium in system n, the hazard rate $q_k^n(\cdot)$ of the abandonment time and the hazard rate $\beta_k^n(\cdot)$ of the (steady-state) distribution of the VOWT should be consistent with each other. Recall that the abandonment probability function $q_k^n(\cdot)$ is characterized by equation (4.25). In particular, it depends on $\beta_k^n(\cdot)$. Also recall that the hazard rate function $\beta_k^n(\cdot)$ of the (steady-state) distribution of the VOWT depends on class k customers' abandonment probabilities $q_k^n(\cdot)$ as well as the service discipline used. Because the exact analysis of the VOWT for the various class is not tractable analytically, we follow Armony and Maglaras [8, 9] and analyze the system equilibrium in heavy traffic.

$\Phi(\hat{q}^*)$ *and* $\hat{q}_k^* = \Gamma_k(\hat{\beta}_k^*)$ *for* $k = 1, \ldots, K$.

In Section 4.1.2, we make a specific assumption on the distribution function $F_k^n(\cdot)$, $n \geq 1$ and $k = 1, \ldots, K$ and characterize the equilibrium in the heavy traffic limit under that assumption.

4.1.2 Characterization of the Equilibrium in the Heavy Traffic Limit

In this section, we characterize the equilibrium by analyzing a one-dimensional workload process. In particular, we obtain a closed-form characterization of the hazard rate $\hat{\beta}(\cdot)$ of the (vector-valued) VOWT in steady-state, given the abandonment rate $\hat{q}(\cdot)$ and vice versa. To this end, let $\hat{W}_k(t)$ denote the backlog of class k workload in the system at time t, defined as

$$\hat{W}_k(t) = \frac{1}{\mu_k}\hat{Q}_k(t), \ t \geq 0. \tag{4.27}$$

Then the workload in the system, denoted by $\hat{W}(t)$, is given by

$$\hat{W}(t) = \sum_{k=1}^{K} \hat{W}_k(t), \ t \geq 0. \tag{4.28}$$

Substituting equations (4.21) and (4.27) into (4.28) characterizes the evolution of the workload as follows:

$$\hat{W}(t) = \hat{X}(t) - \int_0^t \sum_{k=1}^{K} \rho_k \int_0^{\frac{\hat{W}_k(s)}{\rho_k}} \hat{q}_k(u)\,\mathrm{d}u\,\mathrm{d}s + \hat{I}(t), \tag{4.29}$$

where $\hat{X}(t) = \sum_{k=1}^{K} \hat{X}_k(t)/\mu_k$. Note that $\hat{X}(t)$ is a one-dimensional Brownian motion that starts from the origin with drift $\theta = \sum_{k=1}^{K} \theta_k/\mu_k$ and variance $\sigma^2 = \sum_{k=1}^{K} \frac{1}{\mu_k^2}[\lambda_k^3(\sigma_k^A)^2 + \rho_k\mu_k^3(\sigma_k^S)^2]$. Since we only consider the underloaded case, the drift $\theta < 0$.

We restrict attention to the class of scheduling policies under which a state-space collapse result holds in the heavy traffic limit in the following sense: There exists a vector-valued

function $\gamma(\cdot) = (\gamma_1(\cdot), \ldots, \gamma_K(\cdot)) : \mathrm{R}_+ \to \mathrm{R}_+^K$ such that

$$\hat{W}_k(t) = \gamma_k(\hat{W}(t)),\ t \geq 0 \ \text{ and } \ k = 1, \ldots, K. \tag{4.30}$$

We make the following assumption on the function $\gamma(\cdot)$.

Assumption 4. *For $k = 1, \ldots, K$, we assume that*

1. *$\gamma_k(\cdot)$ is strictly increasing and continuously differentiable.*
2. *There exists a constant $\bar{\gamma}_k'$ such that $\lim_{w\to\infty} \gamma_k'(w) = \bar{\gamma}_k'$, where $\gamma_k'(\cdot)$ is the derivative of $\gamma_k(\cdot)$ for $k = 1 \ldots, K$*[6].

By the assumption that $\gamma_k(\cdot)$ is strictly increasing and continuously differentiable, $\gamma_k(\cdot)$ is invertible with its inverse denoted by $\gamma_k^{-1}(\cdot)$, $k = 1, \ldots, K$.

Remark 1. The class of policies considered include dynamic index policies, and hence, a large class of dynamic priority policies. However, it does not include static priority policies, e.g. the $c\mu$-rule. Nonetheless, under static priority policies all backlog is kept in a single-class at all times (in the limit). Therefore, the problem effectively reduces to a one-dimensional problem and the results to follow apply in that case trivially.

We also make the following assumption on the reward and the cost of waiting[7].

Assumption 5. *Assume that the reward and the cost of waiting satisfy the following:*

$$r_k + c_k \frac{\sigma^2}{2\theta\rho_k} \sup_{t\geq 0} \gamma_k'(t) > 0,\ k = 1, \ldots, K.$$

6. It follows from (4.30) that $\gamma_k(x) \leq x$ for $x \geq 0$. Thus $\gamma'(\cdot)$ cannot diverge. This assumption excludes that case that $\gamma_k'(\cdot)$ oscillates.

7. Recall that $\theta < 0$. Then note that $-\sigma^2 \sup_{t\geq 0} \gamma_k'(t)/2\theta\rho_k$ is an upper bound of the expected waiting time of class k customers; see Corollary 7. Thus, under Assumption 5, the expected utility of waiting (ignoring the random shocks) is greater than the expected utility of abandoning. This assumption ensures that the scaled expected utility $\hat{J}_k(\cdot)$ $(k = 1, \ldots, K)$ of waiting is bounded away from zero. Thus, the hazard rate $\hat{q}_k(\cdot)$ of the abandonment time, given in equation (4.32), is well defined.

In addition, we assume that $F_k^n(\cdot)$ (for $n \geq 1$ and $k = 1, \ldots, K$) follows a symmetric Pareto distribution with its cumulative distribution function provided in the following assumption. Under this assumption, we are able to provide a closed form characterization of the scaled expected utility $\hat{J}_k(\cdot)$ of waiting and the hazard rate $\hat{q}_k(\cdot)$ in the heavy traffic limit.

Assumption 6. *Choose $\delta > 1$ and $x_n = 1/2^n$ for $n \geq 1$. Then we assume for $k = 1, \ldots, K$ and $n \geq 1$ that*

$$F_k^n(x) = \begin{cases} \frac{x_n}{2(-x)^\delta}, & x \leq -(x_n)^{1/\delta}, \\ 1/2, & -(x_n)^{1/\delta} < x < (x_n)^{1/\delta}, \\ 1 - \frac{x_n}{2x^\delta}, & x \geq (x_n)^{1/\delta}. \end{cases} \tag{4.31}$$

Under Assumptions 5 and 6, the abandonment rate $\hat{q}_k(w)$ and the limiting scaled expected utility of waiting[8] $\hat{J}_k(w)$ for $k = 1, \ldots, K$ and $w \geq 0$ are given as follows[9](see Appendix B.1.1 for the derivation):

$$\hat{q}_k(w) = \frac{1}{2(\hat{J}_k(w))^\delta}, \tag{4.32}$$

$$\hat{J}_k(w) = r_k - c_k \int_w^\infty \exp\left(\int_w^s -\hat{\beta}_k(u)\,\mathrm{d}u\right)\,\mathrm{d}s. \tag{4.33}$$

Remark 2. Although the analysis to follow characterizes the equilibrium and establishes its uniqueness under Assumption 6, we believe that the existence and uniqueness result remains valid more generally. The existence proof requires that the mapping from $\hat{\beta}_k(\cdot)$ to $\hat{J}_k(\cdot)$ (for $k = 1, \ldots, K$), provided in (4.33), is continuous[10]. The uniqueness proof, which is built on

8. The scaled expected utility of waiting is defined as $\hat{J}_k^n(w) = J_k^n\left(\lfloor\sqrt{2^n}w\rfloor/2^n\right)$ for $w \geq 0$ and $k = 1, \ldots, K$. We assume that $\hat{J}_k^n \to \hat{J}_k$ as $n \to \infty$ for $k = 1, \ldots, K$.

9. We show in Appendix B.1.1 that under Assumption 5, $\hat{J}_k(\cdot)$, $k = 1, \ldots, K$ is bounded away from zero; see equation (B.16). Thus, $\hat{q}_k(\cdot)$ is well defined.

10. Later on we use suitable transformations of $\hat{\beta}_k(\cdot)$ and $\hat{J}_k(\cdot)$, denoted by $\hat{\beta}_W(\cdot)$ and $\tilde{J}_k(\cdot)$, to characterize the equilibrium. To put it formally, we require that the mapping L_1 defined in (4.51), which maps from $\hat{\beta}_W(\cdot)$ to $\tilde{J}_k(\cdot)$, is continuous. The continuity of the mapping L_1 helps prove property (i) of Lemma 10, which in turn establishes the existence.

contradiction, relies on three essential properties of the mappings characterized by (4.32) and (4.33). First, the abandonment rate $\hat{q}_k(w)$ is characterized by a (strictly) decreasing function of $\hat{J}_k(w)$ (for $w \geq 0$ and $k = 1, \ldots, K$) as shown in (4.32). This property enables us to characterize the equilibrium using $\hat{J}_k(\cdot)$ instead of $\hat{q}_k(\cdot)$. Second, equation (4.33) shows that $\hat{J}_k(w) \to r_k$ as $w \to \infty$[11]. We show in Appendix B.1.1 that $\hat{J}_k(\cdot)$ is the solution to an ordinary differentiation equation, rewritten here for convenience:

$$\hat{J}_k'(w) = c_k - \hat{\beta}_k(w)(r - \hat{J}_k(w)), \ w \geq 0.$$

The third property used in the uniqueness proof is that the right-hand side of this ODE is decreasing in $\hat{\beta}_k(w)$ and increasing in $\hat{J}_k(w)$[12]. We believe that these properties carry over to a more general class of distributions.

To facilitate the analysis to follow, define the aggregate abandonment rate, denoted by $\hat{q}_W(\cdot)$, as follows:

$$\hat{q}_W(x) = \sum_{k=1}^{K} \gamma_k'(x) \hat{q}_k \left(\frac{\gamma_k(x)}{\rho_k} \right), \quad x \geq 0. \tag{4.34}$$

Substituting equations (4.30) and (4.34) into (4.29) allows us to rewrite the evolution of $\hat{W}(\cdot)$ as follows:

$$\hat{W}(t) = \hat{X}(t) - \int_0^t H(\hat{W}(s)) \, \mathrm{d}s + \hat{I}(t), \quad t \geq 0, \tag{4.35}$$

where

$$H(x) = \int_0^x \hat{q}_W(s) \, \mathrm{d}s, \ x \geq 0. \tag{4.36}$$

Equations (4.35)-(4.36) characterize the dynamics of the workload process $\hat{W}(\cdot)$ by incorpo-

11. To be specific, we require that the transformation $\tilde{J}_k(w)$ of $\hat{J}_k(w)$ converges r_k as $w \to \infty$. This property is provided in Lemma 11, which builds one side of the contradiction.

12. To put it formally, we require that the right-hand side of the ODE (4.46) is decreasing in $\hat{\beta}_W(w)$ and increasing in $\tilde{J}_k(w)$. This property is used to prove Lemmas 13 and 14, which in turn build the other side of the contradiction.

rating customers' abandonment behavior through the aggregate abandonment rate $\hat{q}_W(\cdot)$. Proposition 6.2 of Reed and Ward [93] characterizes the steady-state distribution of the process $\hat{W}(\cdot)$, which we state next for completeness.

Proposition 6. *(Proposition 6.2 of Reed and Ward [93]). Let $\hat{W}(\cdot)$ denote the solution to the SDE in equation* (4.35)*. The density of its steady-state distribution is given as follows:*

$$\pi(w) = \zeta \exp\left\{ \frac{2}{\sigma^2} \left(\theta w - \int_0^w H(s)\,\mathrm{d}s \right) \right\},\ w \geq 0,$$

where ζ is a normalizing constant such that $\int_0^\infty \pi(s)\,\mathrm{d}s = 1$.

Let $\Pi(\cdot)$ denote the cumulative distribution function of the steady state distribution of the workload process. The hazard rate of $\Pi(\cdot)$, denoted by $\hat{\beta}_W(\cdot)$, is given as follows: For $w \geq 0$,

$$\hat{\beta}_W(w) = \frac{\pi(w)}{1 - \Pi(w)} = \left(\int_w^\infty \exp\left[\frac{2}{\sigma^2} \left(\theta(s-w) - \int_w^s H(u)\,\mathrm{d}u \right) \right] \mathrm{d}s \right)^{-1}. \tag{4.37}$$

The following corollary characterizes the hazard rate $\hat{\beta}(\cdot)$ of the VOWT given the hazard rate $\hat{\beta}_W(\cdot)$ of the workload; see Appendix B.1.2 for the proof.

Corollary 6. *The hazard rate $\hat{\beta}_k(\cdot)$ of the VOWT of class k is given as follows:*

$$\hat{\beta}_k(w) = \hat{\beta}_W\left(\gamma_k^{-1}(\rho_k w) \right) (\gamma_k^{-1})'(\rho_k w)\,\rho_k,\ k = 1, \ldots, K. \tag{4.38}$$

The following corollary provides a lower bound for $\hat{\beta}_k$ (for $k = 1, \ldots, K$), and hence, an upper bound for the (steady-state) expected VOWT, denoted by $\mathbb{E}\hat{V}_k(\infty)$; see Appendix B.1.2 for the proof.

Corollary 7. *We have that $\hat{\beta}_W(w) \geq -2\theta/\sigma^2$ for $w \geq 0$. In addition, the following holds:*

For $k = 1, \ldots, K$,

$$\hat{\beta}_k(w) \geq -\frac{2\theta}{\sigma^2} \frac{\rho_k}{\sup_{t\geq 0} \gamma_k'(t)}, \ w \geq 0, \tag{4.39}$$

$$\mathbb{E}\hat{V}_k(\infty) \leq -\frac{\sigma^2}{2\theta} \sup_{t\geq 0} \gamma_k'(t). \tag{4.40}$$

Note that equations (4.34) and (4.36)-(4.38) provide closed-form characterization of the mapping Φ from the abandonment rate $\hat{q}(\cdot)$ to the hazard rate of the VOWT $\hat{\beta}(\cdot)$, which leads to the following Proposition.

Proposition 7. *The equilibrium in the heavy traffic limit is characterized by equations* (4.32)-(4.34) *and* (4.36)-(4.38).

It follows from Proposition 8 that the nonnegative functions $(\hat{\beta}_W^*, H^*, \hat{q}_W^*, \hat{J}^*, \hat{q}^*, \hat{\beta}^*)$ in an equilibrium satisfy equations (4.32)-(4.34) and (4.36)-(4.38) simultaneously. Also, it follows from (4.32) that there is a one-to-one mapping between the abandonment rate $\hat{q}_k(w)$ and the expected discounted utility of waiting $\hat{J}_k(w)$ for $w \geq 0$. In addition, it follows from (4.32) and (4.34) that $\hat{q}_W(w)$ is fully characterized by $\hat{J}(w) = (\hat{J}_1(w), \ldots, \hat{J}_K(w))$, $w \geq 0$. Thus, to simplify the equilibrium characterization, we substitute equations (4.32)-(4.34) into equation (4.36). This leads to a direct characterization of H in terms of $\hat{J}$ as shown next:

$$H(w) = \int_0^w \sum_{k=1}^K \gamma_k'(s) \left[2 \left(\hat{J}_k \left(\frac{\gamma_k(s)}{\rho_k} \right) \right)^\delta \right]^{-1} \mathrm{d}s, \ w \geq 0. \tag{4.41}$$

To facilitate the analysis to follow, we define a suitable transformation of $\hat{J}_k(w)$ through a time change as follows: For $k = 1, \ldots, K$ and $w \geq 0$, let

$$\tilde{J}_k(w) = \hat{J}_k \left(\frac{\gamma_k(w)}{\rho_k} \right). \tag{4.42}$$

Substituting this and (4.38) into equation (4.33), we obtain the following equation that

characterizes $\tilde{J}_k$ from $\hat{\beta}_W$ directly:

$$\tilde{J}_k(w) = r_k - c_k \int_w^\infty \exp\left(\int_w^s -\hat{\beta}_W(u)\, \mathrm{d}u\right) \frac{\gamma_k'(s)}{\rho_k}\, \mathrm{d}s,\ w \geq 0, k = 1, \ldots, K. \tag{4.43}$$

In addition, by substituting (4.42) into equation (4.41), we obtain that

$$H(w) = \int_0^w \sum_{k=1}^K \frac{\gamma_k'(s)}{2(\tilde{J}_k(s))^\delta}\, \mathrm{d}s,\ w \geq 0. \tag{4.44}$$

Thus, we characterize the equilibrium through the nonnegative functions $(\hat{\beta}_W^*, \tilde{J}^*, H^*)$ as stated in the next corollary.

Corollary 8. *The equilibrium quantities* $(\hat{\beta}_W^*, \tilde{J}^*, H^*)$ *are characterized by equations* (4.37) *and* (4.43)-(4.44) *simultaneously.*

In addition, by differentiating equations (4.37) and (4.43)-(4.44), we obtain a system of ordinary differential equations (ODE) with conditions on the initial values[13].

Corollary 9. *The equilibrium is characterized by the following system of ordinary differential equations: For* $w \geq 0$*,*

$$\hat{\beta}_W'(w) = \hat{\beta}_W(w)\left(\hat{\beta}_W(w) + \frac{2}{\sigma^2}(\theta - H(w))\right), \tag{4.45}$$

$$\tilde{J}_k'(w) = c_k \frac{\gamma_k'(w)}{\rho_k} - \hat{\beta}_W(w)(r_k - \tilde{J}_k(w)),\ k = 1, \ldots, K, \tag{4.46}$$

$$H'(w) = \sum_{k=1}^K \frac{\gamma_k'(w)}{2(\tilde{J}_k(w))^\delta}, \tag{4.47}$$

13. Note, however, that this system of ODEs is not a system of initial value problems, because the initial values are determined endogenously and depend on the entire paths of the solutions.

where the initial values satisfy

$$\hat{\beta}_W(0) = \left(\int_0^\infty \exp\left[\frac{2}{\sigma^2} \left(\theta s - \int_0^s H(u)\,\mathrm{d}u \right) \right] \mathrm{d}s \right)^{-1}, \tag{4.48}$$

$$\tilde{J}_k(0) = r_k - c_k \int_0^\infty \exp\left(\int_0^s -\hat{\beta}_W(u)\,\mathrm{d}u \right) \frac{\gamma_k'(s)}{\rho_k}\,\mathrm{d}s, \tag{4.49}$$

$$H(0) = 0. \tag{4.50}$$

4.1.3 Existence and Uniqueness of the Equilibrium

This section establishes the existence and uniqueness of the equilibrium.

Theorem 2. *There exists a unique equilibrium e^*.*

To prove this theorem, we first provide useful properties of any potential equilibrium. Then we use these properties to prove the existence of an equilibrium (Proposition 8), followed by the proof of the uniqueness (Proposition 9).

To facilitate the analysis to follow, let $C[0,\infty)^k$ and $C^1[0,\infty)^k$ denote the set of continuous functions and the set of continuously differentiable functions on $[0,\infty)^k$ (for $k \geq 1$), respectively. In light of equations (4.37) and (4.43)-(4.44), we define the following mappings:

$$L_1 : C[0,\infty) \to C^1[0,\infty)^K, \quad L_2 : C^1[0,\infty)^K \to C^1[0,\infty), \quad L_3 : C^1[0,\infty) \to C^1[0,\infty),$$

which are given as follows: For $x \in C[0,\infty)$, $y \in C^1[0,\infty)^K$, $z \in C^1[0,\infty)$ and $w \geq 0$,

$$[L_1 \circ x]_k(w) = r_k - c_k \int_w^\infty \exp\left(\int_w^s -x\,(u)\ \mathrm{d}u \right) \frac{\gamma_k'(s)}{\rho_k}\,\mathrm{d}s,\ k = 1, \ldots, K, \tag{4.51}$$

$$[L_2 \circ y](w) = \int_0^w \sum_{k=1}^K \frac{\gamma_k'(s)}{2(y_k(s))^\delta}\,\mathrm{d}s, \tag{4.52}$$

$$[L_3 \circ z](w) = \left(\int_w^\infty \exp\left[\frac{2}{\sigma^2} \left(\theta(s-w) - \int_w^s z(u)\,\mathrm{d}u \right) \right] \mathrm{d}s \right)^{-1}. \tag{4.53}$$

For $x \in C[0,\infty)$, define $\tilde{J}_x = (\tilde{J}_{x,1}, \ldots, \tilde{J}_{x,K})$, H_x and $\hat{\beta}_x$ as follows:

$$\tilde{J}_x = L_1 \circ x,\ H_x = L_2 \circ L_1 \circ x \ \text{ and } \ \hat{\beta}_x = L_3 \circ L_2 \circ L_1 \circ x. \tag{4.54}$$

In addition, let Ψ denote the composition of the mappings L_1, L_2 and L_3. That is, $\Psi = L_3 \circ L_2 \circ L_1$. Thus, the equilibrium hazard rate $\hat{\beta}^*_W$ is the fixed point of the mapping Ψ. That is, $\hat{\beta}^*_W = \Psi \circ \hat{\beta}^*_W$. By studying the properties of the mapping Ψ, we derive useful properties of the equilibrium.

To facilitate the analysis to follow, we next define auxiliary functions which provides upper and lower bounds for the equilibrium quantities. The following functions give the lower and upper bounds for H^*: For $w \geq 0$, let

$$\underline{H}(w) = \int_0^w \sum_{k=1}^K \frac{\gamma'_k(s)}{2r_k^\delta}\,\mathrm{d}s, \tag{4.55}$$

$$\bar{H}(w) = \int_0^w \sum_{k=1}^K \left(2\left(r_k + \frac{c_k\sigma^2}{2\theta\rho_k}\sup_{u\geq 0}\gamma'_k(u)\right)^\delta\right)^{-1} \gamma'_k(s)\,\mathrm{d}s. \tag{4.56}$$

By substituting these bounds for H^* into the mapping L_3 gives the following functions: For $w \geq 0$,

$$\underline{\beta}(w) = \left(\int_w^\infty \exp\left[\int_w^s \frac{2}{\sigma^2}(\theta - \underline{H}(u))\,\mathrm{d}u\right]\,\mathrm{d}s)\right)^{-1}, \tag{4.57}$$

$$\bar{\beta}(w) = \left(\int_w^\infty \exp\left[\int_w^s \frac{2}{\sigma^2}(\theta s - \bar{H}(u))\,\mathrm{d}u\right]\,\mathrm{d}s)\right)^{-1}. \tag{4.58}$$

These provide lower and upper bounds for $\hat{\beta}^*_W$, respectively. The following function gives a lower bound for $\tilde{J}^*$: For $w \geq 0$, and $k = 1, \ldots, K$,

$$\underline{J}_k(w) = r_k - \frac{c_k}{\underline{\beta}(w)\rho_k}\sup_{t\geq 0}\gamma'_k(t). \tag{4.59}$$

Note that for $w \geq 0$ and $k = 1, \ldots, K$,

$$\underline{\beta}(w) > -\frac{2\theta}{\sigma^2} \quad \text{and} \quad \underline{J}_k(w) > r_k + \frac{c_k \sigma^2}{2\theta\rho_k} \sup_{t \geq 0} \gamma_k'(t). \tag{4.60}$$

The following lemma shows useful properties of the mapping Ψ; see Appendix B.1.2 for its proof.

Lemma 10. *The following holds:*

(i) The mapping Ψ is continuous in the topology induced by the uniform convergence on compact sets (u.o.c.). In particular, if the sequence of functions $x_n : [0, \infty) \to [-2\theta/\sigma^2, \infty)$ converges to x u.o.c. as $n \to \infty$, then $\Psi \circ x_n \to \Psi \circ x$ u.o.c. as $n \to \infty$.

(ii) Let $x \in C[0, \infty)$ be such that $x(w) \geq -2\theta/\sigma^2$ for $w \geq 0$. Then, the following holds: For $w \geq 0$,

$$\hat{\beta}_x(w) \in \left[\underline{\beta}(w), \bar{\beta}(w)\right], \ \tilde{J}_x(w) \in \prod_{k=1}^{K} \left[r_k + \frac{c_k \sigma^2}{2\theta\rho_k} \sup_{t \geq 0} \gamma_k'(t), r_k \right)$$

and $H_x(w) \in \left[\underline{H}(w), \bar{H}(w)\right]$.

(iii) Let $x \in C[0, \infty)$ be such that $x(w) \in \left[\underline{\beta}(w), \bar{\beta}(w)\right]$ for $w \geq 0$. Then the following holds: For $w \geq 0$, $\tilde{J}_x(w) \in \prod_{k=1}^{K} [\underline{J}_k(w), r_k)$. Moreover, $\hat{\beta}_x(w) + 2(\theta - H_x(w))/\sigma^2 \geq 0$.

(iv) Let $x \in C[0, \infty)$ be such that $x(w) \in \left[\underline{\beta}(w), \bar{\beta}(w)\right]$ for $w \geq 0$. Then the following hold: For $k = 1, \ldots, K$,

$$\lim_{w \to \infty} \tilde{J}_{x,k}(w) = r_k, \ \lim_{w \to \infty} H_x(w) = \infty \quad \textit{and} \quad \lim_{w \to \infty} \hat{\beta}_x(w) + \frac{2}{\sigma^2}\left(\theta - H_x(w)\right) = 0.$$

Throughout the rest of this section, we suppress the superscript '$*$' and denote the equilibrium quantities by $(\hat{\beta}_W, \tilde{J}, H)$. We use the properties of the mappings L_1, L_2, L_3 and Ψ

given in Lemma 10 to calculate the set in which the equilibrium quantities live, which is given in the following lemma; see Appendix B.1.2 for its proof.

Lemma 11. *We have that* $(\hat{\beta}_W(w), \tilde{J}(w)) \in \mathcal{A}(w)$ *for* $w \geq 0$*, where*

$$\mathcal{A}(w) = [\underline{\beta}(w), \bar{\beta}(w)] \times \prod_{k=1}^{K} [\underline{J}_k(w), r_k), \ w \geq 0. \tag{4.61}$$

In addition, we have that

$$\lim_{w\to\infty} \hat{\beta}_W(w) + \frac{2}{\sigma^2}\left(\theta - H(w)\right) = 0, \tag{4.62}$$

and $\lim_{w\to\infty} \tilde{J}_k(w) = r_k$ *for* $k = 1, \ldots, K$*.*

The following is immediate from Lemma 11 because $\underline{\beta}(w) \to \infty$ as $w \to \infty$.

Corollary 10. *As* $w \to \infty$*,* $\hat{\beta}_W(w) \to \infty$*.*

Next, we show that the equilibrium exists. The following lemma, which is a special case of the Schauder-Tychonoff fixed-point theorem [35], facilities the proof of the existence.

Lemma 12. *(Page 9, Coppel, 1965) Let* $C[0, \infty)$ *denote the set of all functions which are continuous on* $[0, \infty)$ *and let* F *be the subset formed by those functions* $x(t)$ *such that*

$$\mu_1(t) \leq x(t) \leq \mu_2(t), \ t \in [0, \infty), \tag{4.63}$$

where $\mu_1(t)$ *and* $\mu_2(t)$ *are fixed positive continuous functions*[14]*. Let* T *be a mapping of* F *into itself with the properties:*

(i) T *is continuous, in the sense that if* $x_n \in F$ *for* $n = 1, 2, \ldots$ *and* $x_n \to x$ *u.o.c., then* $T(x_n) \to T(x)$ *u.o.c.*

14. In the original statement of this theorem in Coppel [35], F is the subset formed by functions $x(t)$ such that $|x(t)| \leq \mu_1(t)$. The proof still goes through if we let F be any closed subset of $\{x(t) : |x(t)| \leq \mu_1(t)\}$. Thus, we modify the theorem so that we can apply it most conveniently.

(ii) The functions in the image set $T(F)$ are equicontinuous at every point of $[0,\infty)$[15].

Then the mapping T has at least one fixed point in F.

Thus, we show the existence of the equilibrium, stated in the following proposition, by showing that the mapping Ψ satisfies the conditions in Lemma 12.

Proposition 8. *There exists an equilibrium e^*.*

Proof. By Corollary 8, it suffices to show that there exists a fixed point of the mapping Ψ. Let $\mu_1(t) = \underline{\beta}(t)$ and $\mu_2(t) = \bar{\beta}(t)$ for $t \geq 0$. Note that both $\mu_1(\cdot)$ and $\mu_2(\cdot)$ are positive and continuous. In addition, let F be the subset of $C[0,\infty)$ defined as follows:

$$F = \{x \in C[0,\infty) : \mu_1(t) \leq x(t) \leq \mu_2(t),\ t \geq 0\}.$$

In addition, denote the image of F under Ψ by $\mathcal{B}_W$, i.e. $\mathcal{B}_W = \Psi(F)$.

To show that Ψ has a fixed point, we need to show that Ψ maps F into itself and that the two properties in Lemma 12 are both satisfied. It follows from property (ii) of Lemma 10 that for any $x \in F$, $(\Psi \circ x)(t) \subseteq [\mu_1(t), \mu_2(t)]$ for $t \geq 0$ because $x(t) \geq \mu_1(t) = -2\theta/\sigma^2$ for $t \geq 0$. Thus, $\mathcal{B}_W \subseteq F$. In addition, we observe the following:

(i) It follows from property (i) of Lemma 10 that Ψ is continuous.

(ii) We need to show that $\mathcal{B}_W$ is equicontinuous at t for $t \geq 0$. Let H_x and $\hat{\beta}_x$ be the functions defined in equation (4.54). Thus, it follows from (4.53) that for $t \geq 0$,

$$\hat{\beta}_x(t) = \left(\int_t^\infty \exp\left[\frac{2}{\sigma^2} \left(\theta(s-t) - \int_t^s H_x(u)\,\mathrm{d}u \right) \right] \mathrm{d}s \right)^{-1}.$$

15. In Coppel [35], the functions in the image set $T(F)$ are bounded and equicontinuous at every point of $[0,\infty)$. Note that F is bounded at every point of $[0,\infty)$ by assumption; see equation (4.63). Since T maps F to itself, $T(F)$ is bounded at every point of $[0,\infty)$ by assumption as well. Thus, we omit the boundedness condition in the statement.

By taking the derivative of both sides of this equation with respect to t, we have that for $t \geq 0$,

$$\hat{\beta}'_x(t) = \hat{\beta}_x(t)\left(\hat{\beta}_x(t) + \frac{2}{\sigma^2}(\theta - H_x(t))\right) \leq \left(\hat{\beta}_x(t)\right)^2 \leq (\mu_2(t))^2,$$

where the first inequality follows from $\theta < 0$ and property (ii) of Lemma 10 that $H_x(t) \geq \underline{H}(t) \geq 0$. It follows from properties (ii) and (iv) of Lemma 10 that $\hat{\beta}_x(t) \geq 0$ and $\hat{\beta}_x(t) + 2(\theta - H_x(t))/\sigma^2 \geq 0$, so $\hat{\beta}'_x(t) \geq 0$.

Since $\mu_2(t)$ is increasing in t, for any $|\delta| < 1$, we have that $\hat{\beta}'_x(t+\delta) < (\mu_2(t+1))^2$. For $\epsilon > 0$, we let $\delta_t = \min\left\{\epsilon/(\mu_2(t+1))^2, 1\right\}$. Thus, for any $t' \in [t - \delta_t, t + \delta_t]$, it follows from the mean value theorem that

$$|\hat{\beta}_x(t') - \hat{\beta}_x(t)| = |\hat{\beta}'_x(\xi)||t' - t| \leq (\mu_2(t+1))^2\delta_t \leq \epsilon, \tag{4.64}$$

where ξ is in-between t' and t. The first inequality follows from $\hat{\beta}'_x(\xi) \leq (\mu_2(t+1))^2$ and $|t' - t| \leq \delta_t$. Equation (4.64) holds for all $x \in F$, so $\mathcal{B}_W$ is equicontinuous at t.

Since both conditions in Lemma 12 are satisfied, we conclude that equation $\hat{\beta}_W = \Psi \circ \hat{\beta}_W$ has at least one fixed point, i.e. there exists an equilibrium. □

The rest of this section proves uniqueness[16], which we show by contradiction. To this end, suppose $(\hat{\beta}^1_W, \tilde{J}^1, H^1) \neq (\tilde{\beta}^2_W, \tilde{J}^2, H^2)$ are two different equilibria. For $w \geq 0$, define

$$\delta_{\hat{\beta}}(w) = \hat{\beta}^1_W(w) - \hat{\beta}^2_W(w), \tag{4.65}$$

$$\delta_{\tilde{J}_k}(w) = \tilde{J}^1_k(w) - \tilde{J}^2_k(w), \ k = 1, \ldots, K, \tag{4.66}$$

$$\delta_H(w) = H^1(w) - H^2(w). \tag{4.67}$$

It follows from Corollary 10 that $\hat{\beta}^i_W(w) \to \infty$ as $w \to \infty$ for $i = 1, 2$. Therefore, for

16. Appendix B.3 summarizes the logic flow of the proof. A diagram showing the dependence of the auxiliary lemmas is also provided.

analytical convenience, we define $\tilde{\beta}^i(w)$, $i = 1, 2$ as follows:

$$\tilde{\beta}^i(w) = \hat{\beta}^i_W(w) + \frac{2}{\sigma^2}(\theta - H^i(w)), \ w \geq 0. \tag{4.68}$$

In addition, define

$$\delta_{\tilde{\beta}}(w) = \tilde{\beta}^1(w) - \tilde{\beta}^2(w). \tag{4.69}$$

Lemma 11 shows that $\tilde{\beta}^i(w) \to 0$ as $w \to \infty$ for $i = 1, 2$. In what follows, we use $\tilde{\beta}^i(\cdot)$, $i = 1, 2$ to characterize the equilibrium, instead of $\hat{\beta}^i_W(\cdot)$, which is unbounded. Substituting (4.68) into equations (4.45)-(4.50) yields the equilibrium characterization given in the next corollary.

Corollary 11. *The equilibrium quantities* $(\tilde{\beta}^i, \tilde{J}^i, H^i)$, $i = 1, 2$ *satisfy the following: For* $w \geq 0$ *and* $i = 1, 2$,

$$\left(\tilde{\beta}^i(w)\right)' = \tilde{\beta}^i(w)\left(\tilde{\beta}^i(w) - \frac{2}{\sigma^2}(\theta - H^i(w))\right) - \frac{2}{\sigma^2}\sum_{k=1}^{K}\frac{\gamma_k'(w)}{2(\tilde{J}^i_k(w))^\delta}, \tag{4.70}$$

$$\left(\tilde{J}^i_k(w)\right)' = c_k\frac{\gamma_k'(w)}{\rho_k} - (\tilde{\beta}^i(w) - \frac{2}{\sigma^2}(\theta - H^i(w)))(r_k - \tilde{J}^i_k(w)), \tag{4.71}$$

$$\left(H^i(w)\right)' = \sum_{k=1}^{K}\frac{\gamma_k'(w)}{2(\tilde{J}^i_k(w))^\delta}, \tag{4.72}$$

where the initial values satisfy

$$\tilde{\beta}^i(0) = \left(\int_0^\infty \exp\left[\frac{2}{\sigma^2}\left(\theta s - \int_0^s H^i(u)\,\mathrm{d}u\right)\right]\,\mathrm{d}s\right)^{-1}, \tag{4.73}$$

$$\tilde{J}^i_k(0) = r_k - c_k\int_0^\infty \exp\left(\int_0^s -\tilde{\beta}^i(u) + \frac{2}{\sigma^2}(\theta - H^i(u))\,\mathrm{d}u\right)\frac{\gamma_k'(s)}{\rho_k}\,\mathrm{d}s, \tag{4.74}$$

$$H^i(0) = 0. \tag{4.75}$$

The following technical lemma characterizes the relationship between $\delta_{\tilde{\beta}}$, $\delta_{\tilde{J}}$ and δ_H, which plays a key role in proving Lemma 14.

Lemma 13. *There exist nonnegative functions $g_i(w)$, $i = 0, 1, \ldots, K$ that satisfy the following relationship:*

$$\delta_H(w) = -g_0(w)\delta_{\tilde{\beta}}(w) - \sum_{k=1}^{K} g_k(w)\delta_{\tilde{J}_k}(w), \ w \geq 0. \tag{4.76}$$

In addition, there exist constants w_0 and M such that for $w \geq w_0$,

$$g_i(w) \leq M \quad \textit{for} \quad i = 0, 1, \ldots, K. \tag{4.77}$$

The proof of Lemma 13 proceeds in three major steps. First, we construct an auxiliary function $\zeta^w(\cdot)$ that characterizes $H(\cdot)$ in terms of $\hat{\beta}_W(\cdot)$ and $\tilde{J}(\cdot)$. Next, we show that the partial derivatives of $\zeta^w(\cdot)$ are bounded. The last step characterizes equilibrium quantities using $\zeta^w(\cdot)$ and arrives at (4.76); see Appendix B.2 for the proof.

To facilitate the statement of Lemma 14, define the $(K+1)\times(K+1)$ matrix A as follows:

$$A = -\frac{2\theta}{\sigma^2} I + \begin{bmatrix} 0 & \frac{\delta\bar{\gamma}_1'}{\sigma^2 r_1^{\delta+1}} & \cdots & \frac{\delta\bar{\gamma}_K'}{\sigma^2 r_K^{\delta+1}} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix},$$

where I is the $(K+1)\times(K+1)$ identity matrix. We also define the functions $l_k(\cdot)$, $m_k(\cdot)$ and $n_k(\cdot)$ for $k = 1, \ldots, K$ as follows: For $w \geq 0$,

$$l_k(w) = \frac{\delta}{\sigma^2}\left[\frac{\gamma_k'(w)}{(\bar{J}_k(w))^{\delta+1}} - \frac{\bar{\gamma}_k'}{r_k^{\delta+1}}\right] - \frac{2}{\sigma^2}\tilde{\beta}^2(w)g_k(w), \tag{4.78}$$

$$m_k(w) = \frac{2}{\sigma^2}(r_k - \tilde{J}_k^2(w))g_k(w), \tag{4.79}$$

$$n_k(w) = -(r_k - \tilde{J}_k^2(w))\left(1 - \frac{2}{\sigma^2}g_0(w)\right), \tag{4.80}$$

where the function $g_i(\cdot)$ (for $i = 0, 1, \ldots, K$) is defined so that (4.76) and (4.77) hold. In

addition $\bar{J}_k(w)$ in (4.78) is defined so as to satisfy the following:

$$\frac{1}{(\tilde{J}_k^1(w))^\delta} - \frac{1}{(\tilde{J}_k^2(w))^\delta} = -\frac{\delta}{(\bar{J}_k(w))^{\delta+1}}(\tilde{J}_k^1(w) - \tilde{J}_k^2(w)). \tag{4.81}$$

Then, define the matrix $B(w)$ (for $w \geq 0$) as follows:

$$B(w) = \begin{bmatrix} \tilde{\beta}^2(w)\left(1 - \frac{2}{\sigma^2}g_0(w)\right) & l_1(w) & \cdots & l_K(w) \\ n_1(w) & m_1(w) & \cdots & m_K(w) \\ \vdots & \vdots & \ddots & \vdots \\ n_K(w) & m_1(w) & \cdots & m_K(w) \end{bmatrix}.$$

The next two lemmas are used to prove the uniqueness. Lemma 14, which follows from Lemma 11, provides a system of ODEs that $\delta_{\tilde{\beta}}$ and $\delta_{\tilde{J}_k}$ (for $k = 1, \ldots K$) jointly satisfy. Lemma 15 provides a useful property of such system of ODEs. Their proofs are given in Appendix B.1.2. To facilitate the statement of the lemmas, we define the matrix norm $|||M|||_\infty$ of a $(K+1) \times (K+1)$ matrix M as follows:

$$|||M|||_\infty = \max_{1 \leq i \leq K+1} \sum_{j=1}^{K+1} \left|m_{ij}\right|.$$

Lemma 14. *The functions $\delta_{\tilde{\beta}}(w)$ and $\delta_{\tilde{J}_k}(w)$ for $k = 1, \ldots, K$ jointly satisfy the following system of ODEs:*

$$\begin{bmatrix} \delta'_{\tilde{\beta}}(w) \\ \delta'_{\tilde{J}_1}(w) \\ \vdots \\ \delta'_{\tilde{J}_K}(w) \end{bmatrix} = (A + c(w)I + B(w)) \begin{bmatrix} \delta_{\tilde{\beta}}(w) \\ \delta_{\tilde{J}_1}(w) \\ \vdots \\ \delta_{\tilde{J}_K}(w) \end{bmatrix}, \ w \geq 0,$$

where $c(w) = \hat{\beta}_W^1(w) + 2\theta/\sigma^2 \geq 0$ for $w \geq 0$. Moreover, $\lim_{w\to\infty} |||B(w)|||_\infty = 0$.

The following technical lemma is key to the proof of uniqueness.

Lemma 15. *Let $x(\cdot)$ be a solution to the following system of ODEs:*

$$x'(t) = \left(\tilde{A} + \tilde{c}(t)I + \tilde{B}(t)\right) x(t), \tag{4.82}$$

where $\tilde{c}(\cdot)$ is a nonnegative continuous function and $\tilde{A}$ and $\tilde{B}(t)$ are $(K+1) \times (K+1)$ matrices satisfying the following: First, the entries of $\tilde{B}(t)$ are functions of t such that $lim_{t\to\infty}|||B(t)|||_\infty = 0$. Second, I is the $(K+1) \times (K+1)$ identity matrix. Lastly, $\tilde{A}$ is an upper triangular matrix of the form

$$\tilde{A} = aI + \begin{bmatrix} 0 & b_1 & \cdots & b_K \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix},$$

where $a > 0$ and $b_1, \ldots, b_K$ are constants.

If $x(t)$ is nonzero everywhere, i.e. $x(t) \neq 0$ for $t \geq 0$, then $x(t)$ cannot converge to zero, i.e. $x(t) \not\to 0$ as $t \to \infty$.

The uniqueness result follows from Lemmas 14 and 15.

Proposition 9. *The equilibrium is unique.*

Proof. We argue by contradiction. Suppose there exist two equilibria $e_1 = (\tilde{\beta}^1, \tilde{J}^1)$ and $e_2 = (\tilde{\beta}^2, \tilde{J}^2)$ such that $e_1 \neq e_2$. Let H^1 and H^2 be the functions defined by substituting $\tilde{J}^1$ and $\tilde{J}^2$ into the right-hand side of (4.44), respectively. Thus, we have that $(\tilde{\beta}^1, \tilde{J}^1, H^1) \neq (\tilde{\beta}^2, \tilde{J}^2, H^2)$. In addition, we define $\delta_{\tilde{\beta}}$, $\delta_{\tilde{J}_k}$ and δ_H as in (4.66), (4.67) and(4.69).

We first show that $(\delta_{\tilde{\beta}}(w), \delta_{\tilde{J}}(w)) \not\to 0$ as $w \to \infty$ by verifying that $(\delta_{\tilde{\beta}}(w), \delta_{\tilde{J}}(w))$ satisfies the conditions in Lemma 15. It follows from Lemma 14 that the matrices A, $B(w)$ and the function $c(w)$ satisfy the conditions in Lemma 15. Thus, the difference of the

equilibrium quantities $(\delta_{\tilde{\beta}}(w), \delta_{\tilde{J}}(w))$, $w \geq 0$ is a solution to the system of ODEs that satisfies (4.82). The only condition we need to verify is $(\delta_{\tilde{\beta}}(w), \delta_{\tilde{J}}(w)) \neq 0$ for $w \geq 0$. In other words, $\delta_{\tilde{\beta}}(w)$, $\delta_{\tilde{J}_1}(w)$,..., $\delta_{\tilde{J}_K}(w)$ cannot be zero simultaneously for $w \geq 0$.

We show that $(\delta_{\tilde{\beta}}(w), \delta_{\tilde{J}}(w)) \neq 0$ for $w \geq 0$ in two steps. In the first step, we show that $\delta_{\tilde{\beta}}(w)$, $\delta_{\tilde{J}_1}(w)$,..., $\delta_{\tilde{J}_K}(w), \delta_H(w)$ cannot be zero simultaneously for $w \geq 0$, i.e.

$$(\delta_{\tilde{\beta}}(w), \delta_{\tilde{J}}(w), \delta_H(w)) \neq 0, \ w \geq 0. \tag{4.83}$$

Using this and Lemma 13, we show in the second step that $(\delta_{\tilde{\beta}}(w), \delta_{\tilde{J}}(w)) \neq 0$ for $w \geq 0$. We show (4.83) by contradiction. Suppose there exists w_0 such that $(\delta_{\tilde{\beta}}(w_0), \delta_{\tilde{J}}(w_0), \delta_H(w_0)) = 0$. In other words, the following holds: $(\tilde{\beta}^1(w_0), \tilde{J}^1(w_0), H^1(w_0)) = (\tilde{\beta}^2(w_0), \tilde{J}^2(w_0), H^2(w_0))$. Note from Lemma 11 and equation (4.60) that for any potential equilibrium quantity $\tilde{J}_k$, the following holds: For $k = 1, \ldots, K$ and $w \geq 0$,

$$\tilde{J}_k(w) \geq \underline{J}_k(w) \geq r_k + \frac{c_k \sigma^2}{2\theta\rho_k} \sup_{t \geq 0} \gamma_k'(t) > 0, \tag{4.84}$$

where the last inequality follows from Assumption 5.

If we restrict the initial values such that (4.84) holds, then the right-hand sides of equations (4.70)-(4.72) are continuously differentiable and thus locally Lipschitz continuous. Note also that the initial value problem (4.70)-(4.72) with $(\tilde{\beta}(w_0), \tilde{J}(w_0), H(w_0)) = (\tilde{\beta}^1(w_0), \tilde{J}^1(w_0), H^1(w_0))$ satisfies the condition imposed in (4.84). Thus its solution is unique; see Theorem 1.4.1 in Kong [76]. Since $(\tilde{\beta}^1, \tilde{J}^1, H^1)$ and $(\tilde{\beta}^2, \tilde{J}^2, H^2)$ are both solutions to this initial value problem, they must be the same. This contradicts the assumption that they are two different equilibria. Thus, (4.83) holds.

In the second step, we show that $(\delta_{\tilde{\beta}}(w), \delta_{\tilde{J}}(w)) \neq 0$ for $w \geq 0$. It follows from (4.83) that we only need to exclude the case when $(\delta_{\tilde{\beta}}(w), \delta_{\tilde{J}}(w)) = 0$ whereas $\delta_H(w) \neq 0$. We show that this cannot happen by contradiction. Suppose there exists w_1 such that $(\delta_{\tilde{\beta}}(w_1), \delta_{\tilde{J}}(w_1)) = 0$

whereas $\delta_H(w_1) \neq 0$. It follows from Lemma 13 that

$$\delta_H(w_1) = -g_0(w_1)\delta_{\tilde{\beta}}(w_1) - \sum_{k=1}^{K} g_k(w_1)\delta_{\tilde{J}_k}(w_1) = 0.$$

However, this contradicts the assumption that $\delta_H(w_1) \neq 0$ and leads to $(\delta_{\tilde{\beta}}(w), \delta_{\tilde{J}}(w)) \neq 0$ for $w \geq 0$. Thus, it follows from Lemma 15 that $(\delta_{\tilde{\beta}}(w), \delta_{\tilde{J}}(w)) \not\to 0$ as $w \to \infty$.

On the one hand, we have that $(\delta_{\tilde{\beta}}(w), \delta_{\tilde{J}}(w)) \not\to 0$ as $w \to \infty$. On the other hand, it follows from Lemma 11 that

$$\begin{aligned}
&\lim_{w\to\infty} \delta_{\tilde{\beta}}(w) = \lim_{w\to\infty} \tilde{\beta}^1(w) - \lim_{w\to\infty} \tilde{\beta}^2(w) = 0, \\
&\lim_{w\to\infty} \delta_{\tilde{J}_k}(w) = \lim_{w\to\infty} \tilde{J}_k^1(w) - \lim_{w\to\infty} \tilde{J}_k^2(w) = r_k - r_k = 0, \ k = 1, \ldots, K.
\end{aligned}$$

In other words, $(\delta_{\tilde{\beta}}(w), \delta_{\tilde{J}}(w)) \to 0$ as $w \to \infty$, which leads to a contradiction. Thus, the equilibrium is unique. □

Thus, the main result of Theorem 2 follows immediately from Propositions 8 and 9.

4.2 Numerical Characterization of the Equilibrium

This section provides an algorithm to compute the equilibrium. In addition, we provide a numerical example to illustrate the algorithm in Section 4.2.2. This example uses data from an Israeli bank call center and shows the value of modeling abandonment endogenously.

4.2.1 The Algorithm for Computing the Equilibrium

This section provides an algorithm to compute the equilibrium of the original multiclass system by approximating it with the equilibrium of the limiting system studied in the previous section. The approach implicitly assumes that the equilibrium exists and is unique in the pre-limit system. The results of the simulations conducted in Section 4.2.2 makes this

plausible. However, we are unable to prove this because there are no closed-form expressions to characterize the VOWT in the pre-limit system. Therefore, we resort to the heavy traffic approximation and show the existence and uniqueness for the limiting system; see Theorem 2. The premise of the heavy traffic approximation is that for a large system index n, the multiclass queueing model and the limiting diffusion model are close. Thus, we expect the equilibrium of the two to be close as well. This approach is motivated by and follows that of Armony and Maglaras [8]. Letting $e^n = (\beta^n, q^n)$ denote the equilibrium of the n^{th} system and $e = (\hat{\beta}, \hat{q})$ denote the equilibrium of the diffusion model, and using the scaling relations in equations (4.19) and (4.24), we approximate e^n by e as follows: For $w \in \{j/2^n : j = 1, \ldots\}$ and $k = 1, \ldots, K$,

$$q_k^n(w) = \frac{1}{2^n}\hat{q}_k^n\left(\sqrt{2^n}w\right) \approx \frac{1}{2^n}\hat{q}_k\left(\sqrt{2^n}w\right), \tag{4.85}$$

$$\beta_k^n(w) = \frac{1}{\sqrt{2^n}}\hat{\beta}_k^n\left(\sqrt{2^n}w\right) \approx \frac{1}{\sqrt{2^n}}\hat{\beta}_k\left(\sqrt{2^n}w\right). \tag{4.86}$$

Prior to describing the algorithm, we state the following corollary that will be helpful for computing the equilibrium (see Appendix A.1.4 for the proof):

Corollary 12. *We have that* $\lim_{w\to\infty} \hat{\beta}_W(w)/w = \tilde{\beta}_\infty$*, where*

$$\tilde{\beta}_\infty = \frac{2}{\sigma^2}\left(\sum_{k=1}^{K} \frac{\bar{\gamma}_k'}{2r_k^\delta}\right).$$

We approximate the equilibrium by a truncated one[17]. Namely, we fix a large $T > 0$ and replace the equilibrium quantities with exogenously given values for $w \geq T$. To this end, recall from Corollary 8 that the equilibrium is characterized by the nonnegative functions $(\hat{\beta}_W^*, \tilde{J}^*, H^*)$. Also, note from Corollary 12 and Lemma 11 that for large values of w, we have $\hat{\beta}_W(w) \approx w\tilde{\beta}_\infty$ and $\tilde{J}_k(w) \approx r_k$ for $k = 1, \ldots, K$. Thus, we look for a truncated

17. In this sense, our algorithm is similar in spirit to that proposed in Chapter 3 that computes a truncated equilibrium as an approximation. They show that the truncated approximation is accurate.

equilibrium in which $\hat{\beta}_W(w) = \tilde{\beta}_\infty w$ and $\tilde{J}_k(w) = r_k$ for $k = 1, \ldots, K$ and $w \geq T$[18].

In light of Corollary 9, we characterize the equilibrium quantities for $w \leq T$ by the ODEs in (4.45)-(4.47). Thus, the truncated equilibrium is defined formally as follows:

Definition 5. *The truncated equilibrium with $T > 0$ as characterized by $(\hat{\beta}_W(w), \tilde{J}(w), H(w))$ satisfy the following conditions:*

1. *$\hat{\beta}_W(w) = \tilde{\beta}_\infty w$ and $\tilde{J}_k(w) = r_k$ for $w \geq T$ and $k = 1, \ldots, K$.*

2. *$(\hat{\beta}_W(w), \tilde{J}(w), H(w))$ satisfy equations (4.45)-(4.47) and (4.50) for $w \leq T$.*

Viewing $H(T)$ as a parameter, note that equations (4.45)-(4.47) with $(\hat{\beta}_W(T), \tilde{J}(T), H(T))$ given defines an initial value problem. By the uniqueness of the solution to this initial value problem, the values of equilibrium quantities for $w \leq T$ are uniquely determined by the values of $(\hat{\beta}_W(T), \tilde{J}(T), H(T))$. Since the values of $\hat{\beta}_W(T)$ and $\tilde{J}(T))$ are given exogenously, the truncated equilibrium is fully characterized by $H(T)$. Equation (4.50), $H(0) = 0$, provides the consistency condition that $H(T)$ needs to satisfy. The following lemma shows that $H(T)$ is uniquely determined by (4.50), proving that the truncated equilibrium is also unique. Its proof is given in Appendix B.1.3.

Lemma 16. *The truncated equilibrium is unique for $T > 0$.*

To compute the truncated equilibrium, we start with a guess of $H(T)$, and compute the value of $(\hat{\beta}_W(w), \tilde{J}(w), H(w))$ for $w \leq T$. Given a guess of $H(T)$, $(\hat{\beta}_W(w), \tilde{J}(w), H(w))$ for $w \leq T$ is the solution to the initial value problem defined by equations (4.45)-(4.47) with $(\hat{\beta}_W(T), \tilde{J}(T), H(T))$. We can then apply various numerical algorithms to compute the solution to the initial value problem numerically. One of the simplest algorithms is Euler's method. To this end, we discretize time with a small step size $\Delta > 0$, and compute

18. The numerical computation shows that if the truncation time T is large enough, the computed result is insensitive to the choice of the truncation time T and the terminal values $\hat{\beta}(T)$ and $\tilde{J}_k(T)$ (for $k = 1, \ldots, K$). In the numerical example provided in Section 4.2.2, we use one, two and five times the maximum waiting time observed in the data as the value of T to compute the truncated equilibrium. The computed results are the same under these choices.

$(\hat{\beta}_W(w), \tilde{J}(w), H(w))$ recursively from equations derived from equations (4.45)-(4.47)[19]:

$$\hat{\beta}_W(w) = \max\left\{-\frac{2\theta}{\sigma^2}, \hat{\beta}_W(w+\Delta) - \Delta\hat{\beta}_W(w+\Delta)\left(\hat{\beta}_W(w+\Delta) + \frac{2}{\sigma^2}(\theta - H^+(w+\Delta))\right)\right\}, \tag{4.87}$$

$$\tilde{J}_k(w) = \tilde{J}_k(w+\Delta) - \Delta\left[c_k\frac{\gamma_k'(w)}{\rho_k} - \hat{\beta}_W(w)(r_k - \tilde{J}_k(w+\Delta))\right],\ k = 1, \ldots, K, \tag{4.88}$$

$$H(w) = H(w+\Delta) - \Delta\sum_{k=1}^{K}\frac{\gamma_k'(w)}{2(\tilde{J}_k(w))^\delta}, \tag{4.89}$$

where $x^+ = \max(x, 0)$ and $H(0) = 0$. The idea behind the computational scheme is to start with a guess of $H(T)$ and to recursively calculate $\hat{\beta}_W(w)$, $\tilde{J}(w)$ and $H(w)$ for $w < N$. If the guess of $H(T)$ is correct, then the value of $H(0)$ calculated recursively must equal zero. Lemma 17 shows that $H(0)$ is a monotone function of $H(T)$ (see Appendix B.1.3 for its proof) and this observation leads to a simple algorithm.

Lemma 17. *If $H^1(T) > H^2(T)$, then $H^1(0) > H^2(0)$, where $H^1(0)$ and $H^2(0)$ are the values obtained from equations* (4.87)-(4.89) *recursively by substituting $H(T) = H^1(T)$ and $H(T) = H^2(T)$, respectively.*

It follows from Lemma 17 that if $H(0) < 0$, then the true value of $H(T)$ is larger than the guessed value. Therefore, the initial guess must be increased. Otherwise, i.e. $H(0) > 0$, we should lower the initial guess. This observation is key to the algorithm provided in Table 4.1.

Note that given $(\hat{\beta}_W, \tilde{J}, H)$, the other (equilibrium) quantities of interest can be com-

19. If we start with a random guess of $H(T)$ and compute $(\hat{\beta}_W(w), \tilde{J}(w), H(w))$ recursively using (4.87)-(4.89), the value of $H(w)$ can be negative. Thus, we truncate the value of $H(w)$ in (4.87) by zero. In addition, since it follows from Corollary 7 that $\hat{\beta}_W(w) \geq -2\theta/\sigma^2$, we truncate the computed value of $\hat{\beta}_W(w)$ by $-2\theta/\sigma^2$ in (4.87).

Table 4.1: The algorithm for calculating the truncated equilibrium in the multiclass case.

Algorithm 2: The truncated equilibrium in the multiclass case.

Initialize: $H(T) \leftarrow h^0$ and $\bar{h} \leftarrow \infty$ and $\underline{h} \leftarrow 0$.
Update the value of $H(T)$**:**
while $\bar{h} - \underline{h} > \epsilon$
    Calculate $\hat{\beta}_W(\cdot)$, $\hat{q}_W(\cdot)$ and $H(\cdot)$ via equations (4.87)-(4.89).
    if $H(0) = 0$
        stop
    else
        if $H(0) > 0$
            $\bar{h} \leftarrow H(T)$
        else
            $\underline{h} \leftarrow H(T)$
        end if
    end if
    Pick $h \in (\underline{h}, \bar{h})$ and $H(T) \leftarrow h$
end while

puted. In particular, it follows from Corollary 6 that for $w \geq 0$,

$$\hat{\beta}_k\left(\frac{\gamma_k(w)}{\rho_k}\right) = \hat{\beta}_W(w)\left(\gamma_k^{-1}\right)'(\gamma_k(w))\,\rho_k,\ k = 1, \ldots, K.$$

Moreover, it is immediate from equations (4.32) and (4.34) that for $w \geq 0$,

$$\tilde{q}_k(w) = \hat{q}_k\left(\frac{\gamma_k(w)}{\rho_k}\right) = \frac{1}{2(\tilde{J}_k(w))^\delta} \quad \text{and} \quad \hat{q}_W(w) = \sum_{k=1}^{K} \gamma_k'(w)\tilde{q}_k(w),\ k = 1, \ldots, K.$$

4.2.2 A Numerical Example

This section uses a data set from an Israeli bank call center to study the performance of the call center under various service disciplines and endogenous abandonments. In this analysis, we first use the algorithm proposed in Section 4.2.1 to compute the waiting and abandonment time distributions in equilibrium. We then propose an approach that combines the numerical computation with an iterative simulation to characterize the system performance. Lastly, we compare the simulation results with those of the model with exogenous abandonments.

The data set contains individual call level data for a six-month period from April to

September in 2008. The data set is the same one used in Aksin et al. [3][20]; we refer the readers to the Data section of Aksin et al. [3] for a detailed description of the data set.

Currently, the call center operates under a point-update priority policy. Incoming customers are categorized into four priority groups: High, medium, low, and no priority groups. Upon arrival, each customer receives an initial priority point depending on her priority group. The higher priority the group has, the higher the initial point is. The priority point of the customer is then increased every 60 seconds as the customer waits in the queue. Figure 4.1 shows how the priority point is updated as the function of waiting time. When a server

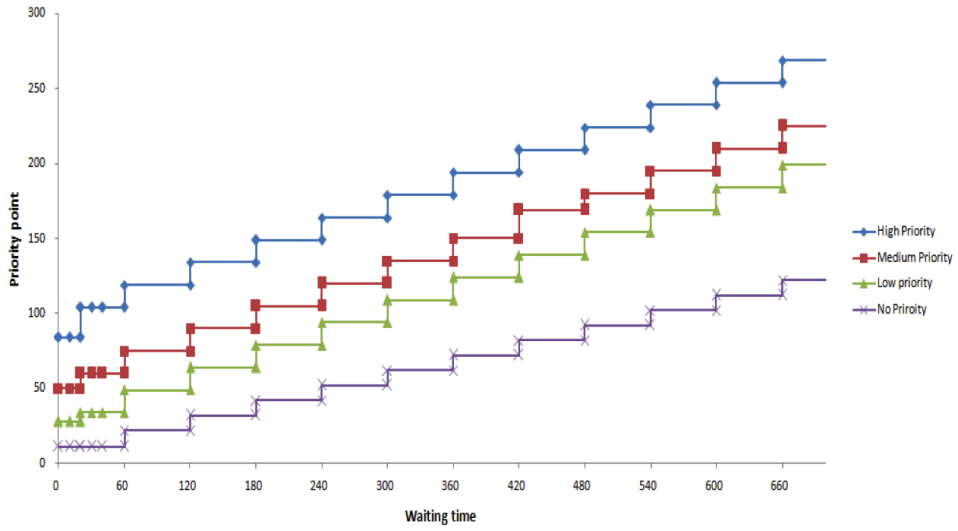


Figure 4.1: The priority points as a function of the priority group of the customer and her waiting time under the current policy.

becomes available, she picks the customer with the highest priority point among four queues. We refer to this policy as the current policy and study the system performance under different service disciplines. We consider the same service disciplines as the ones considered in Aksin et al. [3]: the FCFS policy, a static (and non-preemptive) priority policy, a threshold policy and the reversed strict priority policy[21]. Besides these policies, we also consider a rearranged point-update priority policy. The policy swaps the labels of different groups and

20. The data set is made available by the Service Enterprise Engineering (SEE) lab at Technion (http://ie.technion.ac.il/Labs/Serveng/).

21. The readers can refer to Section 6 of Aksin et al. [3] for a detailed description of the policies.

uses the current policy with the new labels. To be specific, the no-priority group becomes the new high-priority group, and all other groups are downgraded by one, i.e. the high, medium and low priority groups are now labeled as the medium, low and no priority, respectively. The rationale for this policy is that the no-priority group consists largely of new customers, and we would like their first experience with the call center to be positive.

As a preliminary to replicating the current performance of the call center, we first estimate the empirical distributions of the interarrival, service and abandonment times of all four priority groups from the data set. We also estimate the reward (r_k) and the waiting cost (c_k) (for $k = 1, \ldots, K$) of the abandonment model described in Section 2.3.2; see Appendix B.4.1 for this estimation problem and the estimates. In addition, we calibrate the number of agents to replicate the current performance of the call center, as done in Section 6 of Aksin et al. [3]. Namely, using the empirical distributions of the interarrival, service and abandonment times and the current policy, we simulate the system and vary the number of agents. We compare the simulated average waiting time of each group and pick the number of agents to be 133, for which the waiting time and abandonment statistics are closest to those in the data[22].

We compute the waiting time and abandonment time in equilibrium under all service disciplines mentioned above by substituting the estimates into the algorithm proposed in Section 4.2.1. To be specific, we use the approximation with parameter $n = 6$ and $2^n = 64$. As we note in Section 4.2.1, the algorithm seeks only one parameter $H(T)$ and thus is computationally efficient. Table 4.2 shows the computed average VOWTs and the fractions of abandoning customers of all classes.

As Kelly and Laws [68] note "... the important features of good control policies are displayed in sharpest relief" in the heavy traffic limit. As such the numerical computation gives zeros for the average waiting time of the classes which enjoy strict priority. This is due to the state-space collapse result of the heavy traffic approximation. Nonetheless, the

22. This choice is consistent with Aksin et al. [3].

Table 4.2: The computed mean of the VOWTs and the fractions of abandoning customers

	High Priority		Medium Priority		Low Priority		No Priority	
	sec.	% Aban.	sec.	% Aban.	sec.	% Aban.	sec.	% Aban.
Current Policy	0.754	0.02	15.481	1.04	70.182	5.49	182.991	37.22
FCFS	75.412	8.99	75.412	5.89	75.412	5.77	75.412	35.12
Static Priority	-	-	-	-	-	-	251.464	51.57
Threshold (th = 75 sec.)	-	-	-	-	294.784	30.72	75.000	26.85
Threshold (th = 25 sec.)	-	-	-	-	394.442	46.69	25.000	7.23
Reversed strict priority	177.841	60.47	-	-	-	-	-	-
Reversed point-update	50.426	6.07	103.441	11.42	194.817	24.42	6.904	1.52

numerical computation still gives good predictions on the average waiting time of the entire system.

Instead of computing the waiting and the abandonment time in equilibrium numerically, Aksin et al. [3] propose an iterative simulation approach to study the system with endogenous abandonment behavior. This approach starts with a guess of the abandonment time distributions in equilibrium under the new service discipline and simulates the system. Then it uses the realized empirical distributions of the VOWT from the simulation as the input to the abandonment model (of Section 2.3.2), computes the resulting abandonment probabilities in that model and updates the abandonment time distributions accordingly. It then re-runs the simulation with the updated distributions of the abandonment time and keeps updating the abandonment decisions and the VOWT until the simulation converges.

However, the convergence of this approach is not guaranteed. Even when it converges, the rate of convergence depends on the initial guess of the abandonment time distributions. To illustrate these, we simulate the system under three different initial guesses: the results generated by the algorithm developed in Section 4.2.1, the exogenous abandonment distributions observed in the data, and another exogenous distribution where each customer abandons with probability 5% in each period, i.e. a geometric distribution. The stopping criteria of each simulation is that the difference in the average VOWTs of two consecutive iterations is less than or equal to $\epsilon = 1\%$. Table 4.3 shows the number of iterations needed in the simulation for each initial guess under various policies. As shown in Table 4.3, under the reversed strict priority policy, the exogenous model and the endogenous model give significantly different predictions on the waiting time and abandonment behavior of the system.

Table 4.3: Number of iterations of the iterative simulation with various initial guesses.

	Numerical Computation	Exogenous Aban.	Geometric Aban.
Current Policy	2	2	40
FCFS	2	2	14
Static Priority	3	2	N/A
Threshold (th = 75 sec.)	3	3	N/A
Threshold (th = 25 sec.)	3	3	N/A
Reversed strict priority	3	18	33
Reversed point-update	2	6	15

This leads to a high number of iterations unless we use the results of the numerical computation as the initial guess. Moreover, the simulation does not converge under the static priority policy and threshold policies when we use the geometric distribution as the initial guess.

Thus, we propose the approach of combining the numerical computation and the iterative simulation approaches described above. This approach uses the abandonment time distributions from the numerical computation as the initial guess of the iterative simulation. There are two advantages of this approach. First, it fixes the shortcoming of the state-space collapse of the heavy traffic approximations for strict priority policies mentioned above. Second, it speeds up the convergence of the iterative simulation approach significantly.

Using the approach proposed above, we simulate the system performance and abandonment behavior in the new equilibrium under various service policies. We compare the results with those under the exogenous model. The simulation with the exogenous abandonment assumption uses the empirical distributions of the abandonment time observed in the data and assumes that they do not change with the service discipline change. Table 4.4 provides the means of the VOWT and the fraction of customers that abandon in equilibrium under the various policies.

The comparison shows that the predictions of the exogenous and endogenous models are different[23]. This observation suggests that one must be cautious when modeling the

23. We conducted a Kolmogorov-Smirnov test (K-S test) to compare the predicted (steady-state) distributions of the VOWTs of all priority groups from the endogenous and exogenous models. The K-S test shows (at the significance level of 5%) that the endogenous and exogenous distributions are different in all but four cases. The results of the K-S test are summarized in Appendix B.4.2.

Table 4.4: The mean of the VOWT and the fractions of customers that abandon under the exogenous model and the iterative simulation.

Simulation with Exogenous Abandonments								
	High Priority		Medium Priority		Low Priority		No Priority	
	sec.	% Aban.	sec.	% Aban.	sec.	% Aban.	sec.	% Aban.
Current Policy	5.556	0.36	22.950	1.13	72.685	5.44	182.354	36.66
FCFS	80.702	9.77	82.657	6.56	83.977	6.61	83.553	20.61
Static Priority	5.543	0.22	8.510	0.38	26.653	2.26	177.667	40.51
Threshold (th = 75 sec.)	5.511	0.36	19.492	1.12	243.324	27.00	79.832	19.54
Threshold (th = 25 sec.)	5.534	0.37	23.480	1.43	315.836	37.64	31.077	10.58
Reversed strict priority	355.139	59.08	43.950	3.10	8.858	0.41	5.840	2.25
Reversed point-update	59.866	6.06	136.382	11.85	223.539	20.68	7.330	2.44
Simulation with Endogenous Abandonments								
	High Priority		Medium Priority		Low Priority		No Priority	
	sec.	% Aban.	sec.	% Aban.	sec.	% Aban.	sec.	% Aban.
Current Policy	5.442	0.47	19.303	1.27	67.133	5.36	171.958	35.70
FCFS	78.096	10.67	79.435	6.98	80.82	6.89	79.627	18.98
Static Priority	5.536	0.40	8.407	0.66	26.719	1.78	189.652	40.41
Threshold (th = 75 sec.)	5.513	0.76	19.292	1.49	258.741	26.72	78.675	18.36
Threshold (th = 25 sec.)	5.496	0.41	23.370	1.82	336.691	38.42	31.680	8.53
Reversed strict priority	204.252	59.26	42.818	3.38	8.944	0.61	5.846	1.42
Reversed point-update	55.214	5.19	124.106	11.12	209.478	24.15	7.685	1.61

underlying abandonment behavior and predicting the system performance under major operational changes. If customers make rational abandonment decisions, ignoring endogenous abandonments can lead to significant errors in predicting performance. That said, we are not claiming that the endogenous (or the exogenous) model is the correct model. We merely claim that their predictions can be very different. The usefulness and accuracy of either model may depend on the context and can be determined by randomized experiments or by out-of-sample tests for data sets from settings that exhibit sufficiently rich variation in the operating conditions. As such further empirical research is needed to answer this question.

Our findings are consistent with the findings in Aksin et al. [3]. However, our approach, which combines the numerical computation and the iterative simulation, not only provides a theoretical foundation for the study of the system with endogenous abandonments, but also significantly speeds up the computation of the equilibrium.

4.3 Concluding Remarks

In this paper, we study a multiclass queueing system with customers making forward-looking abandonment decisions dynamically. We conduct an approximate analysis of the multiclass

queueing system under the hazard rate scaling and characterize the equilibrium of the heavy traffic limit of such a system. We show that the equilibrium exists and is unique. We also propose an algorithm to compute the system equilibrium.

Our analysis points to several future research directions that worth exploring. For example, we assume that customers receive no information about the system status. An interesting direction is to analyze the systems in which various delay announcements are made. In such systems, customers' abandonment decisions depend not only on the system congestion but also on the announcements they receive. Investigating such systems via the equilibrium approach can help the characterization of the impact of the delay announcement on the system performance measure and customers abandonment behavior simultaneously.

CHAPTER 5

MANAGING CALLBACK OPTION UNDER ARRIVAL RATE UNCERTAINTY

5.1 Introduction

One key challenge in managing call centers is dealing with the uncertainty in call volumes. The arrival process is commonly modeled as a (non-homogeneous) Poisson process. To be specific, the arrival process follows a Poisson process whose intensity, also referred to as the arrival rate process, is a time-dependent deterministic process. Under such arrival models, the uncertainty can be effectively managed by using the square root staffing rule. In particular, the number of agents (and thus the service capacity) equals to the sum of the offered load and a square root safety staffing against uncertainty[1]. Because the coefficient of variation of a Poisson random variable vanishes as its mean gets large, the uncertainty of the Poisson model is negligible for large call centers. Consequently, under this arrival model one predicts that the square-root staffing rule will lead to excellent performance for large call centers; see Gans et al. [44] and Aksin et al. [2] and the references therein.

Recent literature on call centers emphasizes the importance of modeling the arrival rate process as an autocorrelated nonstationary stochastic process by analyzing call arrival data; see discussions in Kim and Whitt (2014a, b) and also Ibrahim et al. [64] for an overview. In particular, Glynn et al. [50] provide a novel perspective on modeling call arrivals. The authors discuss three important time scales in modeling call arrivals and managing them: the microscopic, the macroscopic and the mesoscopic time scales. The microscopic time scale corresponds to minutes, the macroscopic time scale corresponds to hours, whereas the mesoscopic time scale is from a few minutes to hours. Indeed, the uncertainty on microscopic scale can be attributed to the uncertainty due to the Poisson process (as opposed to the

1. The time-of-day effect can be addressed by modifying the staffing level appropriately throughout the day due to the deterministic model of the arrival rate process.

uncertainty in the arrival rate process) whereas the variation on the macroscopic scale corresponds to the time-of-day effect. Thus, the non-homogenous Poisson model can capture the uncertainty on the microscopic and macroscopic time scales well, but ignores the mesoscopic time scale. Glynn et al. [50] observes that when call arrivals are observed on the mesoscopic time scale, they exhibit much higher variability and have no clear seasonality patterns, i.e. the call arrival rate may experience unpredictable fluctuations, i.e. temporary dips or surges, on this time scale. Moreover, the uncertainty on this time scale is not negligible in large call centers as the Poisson model predicts; see Figure 5.1 for the example provided in Figure 1 of Zhang et al. [120][2]. Given the unpredictable and temporary nature of the surge in the arrival rate in the mesoscopic time scale, it is hard to adjust the staffing level with a short notice as needed on this time scale to provide a satisfactory service quality. Although the call center manager can increase its staffing level to account for the arrival rate variablity on the mesoscopic time scale, it can be much costlier than what is anticipated by the square-root staffing.

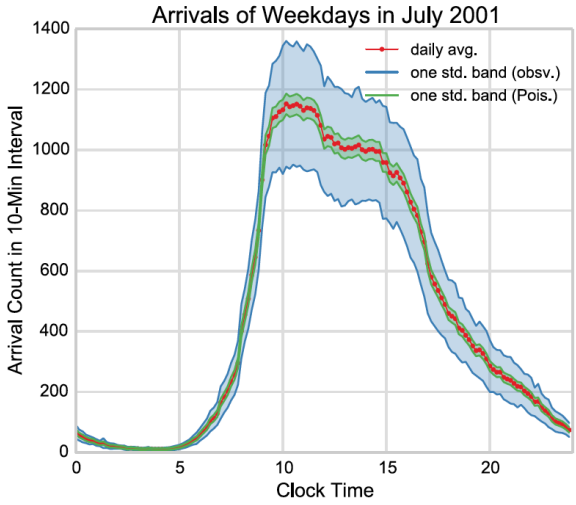


Figure 5.1: The mean and the one-standard deviation bands of the arrival count in 10-minute time periods of the calls arriving at the US bank call center in July 2001. The one-standard deviation band of the observed data is wider than the predicted one-standard deviation band of Poisson model.

2. We refer the readers to Zhang et al. [120] for more analysis of the mean and variance of the observed data.

Instead of a supply-side intervention, e.g. staffing, this chapter studies a demand-side intervention by offering the callback option to manage such temporary and unpredictable surges on the mesoscopic time scale. To be more specific, the callback option works as follows: when the system is congested, the call center manager notifies the arriving customer and presents him with the option to hang up to be called back later within a reasonable time window. When the system congestion decreases, outbound calls are initiated for such customers. Consequently, the callback option allows the call center manager to shift some of the calls arriving when the system is undergoing temporary arrival surges to a period when the system is less busy. A survey conducted by Software Advise[3] shows that more than half of the customers are willing to wait for more than 30 minutes offline for the call center to get back to them. Thanks to the customers who are willing to accept the callback option, the call center manager can improve the performance metrics of the online queue significantly using the callback option; the simulation study in Section 5.6 shows that the average waiting time of the online queue reduces for more than 50% when 5% of customers are served by the callback option.

This chapter studies how to manage the callback option effectively to mitigate the effects of the arrival rate uncertainty using a canonical queueing model. The call center consists of two queues: An online queue and an offline queue. When a customer arrives, the call center manager examines the system state and decides whether to offer the incoming customer the callback option or not. If the call center offers the callback option to the customer, he may decide to accept or reject the offer. A customer is routed to the offline queue (to be called back later) only if he is offered the callback option and he accepts it. Otherwise, he is routed to the online queue. The call center incurs a waiting cost of h per unit time for each customer waiting in the online queue, whereas it incurs a one-time penalty of p if the customer is routed to the offline queue. We refer the call center manager's decision as the callback decision. In

3. More results in the survey can be found at http://csi.softwareadvice.com/3-ways-to-offer-callback-0614/.

what follows, we assume that the time commitment for the callback option is sufficiently large so that we can relax the delay commitment in our theoretical model. Under this assumption, the call center manager makes the callback decision by solving an admission control problem which allows customers to reject or follow the call center manager's admission decisions. This admission control problem is interesting in its own right. However, what makes it relevant to the callback option and the call center operations is the mean-reverting nature of call arrival rate processes on the mesoscopic time scale as observed by Glynn et al. [50]. To be more specific, although the results proved below hold for general point process, we pay particular attention to the Poisson process with stochastic intensity as a call arrival model because of its practical importance. In particular, we model the stochastic intensity as a Cox-Ingersoll-Ross (CIR) process, a mean reverting diffusion process, following Glynn et al. [50]. The mean-reversion feature of the CIR process makes it a good candidate for modeling temporary surges in arrivals in the mesoscopic time scale. When a temporary surge occurs, the call center offers the callback option. After the surge ends, which may last from a few minutes to an hour, the call center uses its excess service capacity to serve customers waiting for the callback. Given the mean-reverting nature of the CIR process (and that the system is stable), the delay commitment made to the offline queue will be met with high probability. Indeed, our simulation study shows that the average delays for the offline queue are around 30 minutes (unless the abandonment rate is high)[4]. The simulation study thus provides a guidance as to how the callback time commitment should be given.

Our theoretical results focus on the complete foresight policies which allow the call center manager to observe all future arrival and service times. First, we consider the setting where all customers are willing to accept the callback option. The callback decision made by the call center manager relies on comparing the marginal costs of routing a customer to the online and offline queues. The marginal cost comparison involves both the direct cost to the

4. The simulation study also allows abandonments. It shows that average delays as well as the 90^{th} percentile of the offline waiting time is less than one hour for most cases unless the abandonment rate is high.

customer in question and the negative externality imposed on others by him if he joins the online queue. The presence of such externalities have long been recognized in the queueing literature; see Hassin [57] and Section 1.8.1 of Hassin [58]. Following Hassin [57], we account for this externality by considering an auxiliary system, which operates under the last-come-first-served (LCFS) discipline and its variations within the online queue. We show that a simple limited-lookahead policy is optimal in this case. This policy keeps track of only the total number of customers in the system (the sum of the numbers of customers in both the online and offline queues) which is insensitive to the routing and priority schemes for work conserving policies. When a new customer arrives, this policy looks at whether the total number of customers in the system will fall back to the level before his arrival in the next p/h time units. If not, he is put in the offline queue to be called back. This policy has several virtues. First, it only needs to keep track of the total number of customers in the system, which is easy to study because its evolution is independent of the call center manager's policy. Second, it requires only limited future information (over the next p/h time units) although it is optimal among policies that can look into the entire future. Third, it is pathwise optimal. Thus, the optimality result carries over to more general settings with an essentially arbitrary call arrival process provided the system is stable.

Although the lookahead policy is not directly applicable because it needs to look into the future, it provides useful insights and helps motivate an effective control policy. For example, it suggests not waiting until the queue builds up to start offering the callback option, motivating a preemptive intervention as opposed to a reactive one. In particular, it suggests offering the callback option as soon as one anticipates a sufficiently high arrival rate, i.e. offering the callback option when the arrival rate exceeds a threshold. The extension of this model, which allows customers to turn down the callback offers, helps us sharpen the insights. In that setting, we show that the modified p/h-lookahead policy is optimal. This policy looks at how long it would take for the total number of customers in the system to drop below the current level minus the current number of customers in the system who previously

rejected the callback offer. The modified p/h-lookahead policy shares the simplicity and the other virtues of the aforementioned lookahead policy. In addition, it highlights the importance of taking into account the time to serve the customers in the system who have rejected the callback offer[5]. Interpreting this policy in the context of a fluid model lends itself to a natural non-anticipating policy that offers the callback option when a linear combination of the online queue length and the current arrival rate exceeds a certain threshold, referred to as the line policy. Interestingly, the line policy also arises as (near) optimal when the externalities imposed on others by the customers admitted to the online queue are taken into account in the fluid model. Thus, we propose and test the line policy as well as the lookahead policies against the optimal policy obtained by solving the associated Markov decision process (under suitable Markovian assumptions) numerically in a simulation study using a data set from a US bank call center.

In our simulation study, we assume that arrival rate process follows the CIR process and calibrate the system using a dataset from a US bank call center. The lookahead policies far outperform the non-anticipating policies in the simulation study. This is to be expected because they use the future information effectively. Interestingly, the line policy has an excellent performance. In the numerical examples, calibrated by the US Bank's call center data, the maximum optimality gap is about 1%. We also observe that routing a small fraction of customers to the offline queue results in excellent system performance, i.e. small waiting times in the online queue and manageable callback delays (for the offline queue). Moreover, the results are robust to customers' accept/reject decisions of the callback option unless the fraction of customers willing to accept the callback offer is low. Lastly, we check how abandonments impact the performance and find that as long as the abandonments are low, the lookahead and line policies perform about as well as they would without any abandonments (despite the increases in the system load due to the reduced number of abandonments).

5. Our analysis in Section 5.4.2 reveals that the system alternates between the callback and no-callback episodes. Moreover, just before the beginning of a callback episode, the online queue consists only of those customers who have previously rejected a callback offer.

However, if the abandonment rate is high, the callback delays may become excessive. This is because the callback option significantly lowers the number of abandonments, thereby increasing the system load significantly. Consequently, we conclude that the callback option under the line policy can be an effective way to mitigate the arrival rate uncertainty provided that the abandonment rate is not too high.

The rest of this chapter is organized as follows. Section 5.2 reviews the related literature. Section 5.3 introduces a two-class queueing model to study a single-class queueing system with the callback option. Section 5.4 studies the complete foresight policies. Section 5.4.1 proves the optimality of the p/h-lookahead policy when all customers accept the callback offer, whereas Section 5.4.2 proves that the modified p/h-lookahead policy is optimal for the case when customers may reject the callback offer. Section 5.5 studies a fluid model and advances a non-anticipating policy, the line policy, combining the insights from the fluid model and the lookahead policies. Section 5.6 presents the numerical study and Section 5.7 concludes. Appendix C.1 provides the system equations characterizing the evolution of the queue length processes. Appendix C.2 provides the proofs of the lemmas in Section 5.4. Appendix C.3 describes the approach to estimate the parameters characterizing the arrival process.

5.2 Literature Review

This chapter is related to three streams of literature. The first stream models the nonstationary arrivals to a call center. The second stream studies call blending, focusing on managing different types of jobs, the inbound and outbound calls, in a call center. This stream includes the study of managing the callback option, which is a specific way of offering outbound calls. The last stream studies admission control decisions for queueing systems. There is a growing literature on modeling the nonstationary arrivals to a call center[6]; see Ibrahim et al. [64] for

6. Nonstationary demand models are relevant in other operations management applications as well; see for example, Besbes and Maglaras [26] for a revenue management application, Shi et al. [99] for a study of a

a recent review. Green and Kolesar [52] propose the pointwise stationary approximations (PSA) to model the time-varying demand. The underlying assumption is that the system reaches its steady state at each time t instantaneously (with the current arrival rate). Using PSA, Harrison and Zeevi [56] propose a stochastic programming approach to study the daily staffing decisions under arrival rate uncertainty. Bassamboo et al. [24] extend this work to multiple customer classes and server pools. Jennings et al. [66] consider a multiserver system with nonstationary arrival and service time processes. The authors choose the staffing level to meet the projected demand subject to a certain service level guarantee. Feldman et al. [41] build on Jennings et al. [66] and propose a flexible simulation-based iterative staffing algorithm to determine the staffing level with abandonments. The authors show that their algorithm yields time-stable delay probabilities across a wide range of target probabilities. Liu and Whitt [83] extend Feldman et al. [41] and propose an algorithm using closed-form approximations they derive for various performance characteristics. The authors show that their algorithm stabilizes the abandonment probability across the full spectrum of target abandonment probabilities. He et al. [62] propose a staffing algorithm, which extends the square-root staffing formula to account for the non-Poisson stochastic variability in the arrival process. Whitt and Zhao [114] develop an effective time-varying staffing strategy to stabilize blocking probabilities at target levels in a loss model with non-Poisson time-varying arrivals and flexible staffing. Whitt [113] provides an extensive review of research on the performance analysis of queueing systems with time-varying arrival rate; see also Gans et al. [44], Green et al. [53] and Aksin et al. [2] and the references therein. The recent work of Gans et al. [46] proposes a framework that integrates the forecasting of the arrival rate and the stochastic programming approach to manage the staffing level in a call center setting. Their model accounts for the time-dependent, stochastic and autocorrelated features of the arrival process as well as the forecast updates on the arrival rate process. The extant litera-

discharge policy in an emergency department, Kim et al. [70] and Kim et al. [71] for studies of endocrinology and outpatient clinics and Ata et al. [16] for an empirical analysis of time-based pricing in electricity supply chains.

ture that focuses on modeling and managing the nonstationary arrivals mainly focus on the staffing and scheduling problem. In contrast, we take the staffing level as given and focus on managing demand using the callback option, i.e. we focus on deciding which arriving customers should be given the callback option.

The second stream of literature studies call blending, a process managing both the inbound and outbound calls in a call center. Inbound calls originate from outside customers, whereas the outbound calls are initiated by the call center. There is infinite supply of the outbound calls. Bhulai and Koole [27] studies an optimal control problem in which the call center manager seeks to maximize the number of outbound calls under the average waiting time constraint of inbound calls. Assuming that the service rates of the two types of jobs are equal, they prove that the optimal control is a threshold-type policy. Gans and Zhou [45] relax this assumption and show that a threshold policy is optimal among all policies that give priority to inbound calls. Pang and Perry [91] study the combined problem of staffing and control in a call center with call blending in heavy traffic. They propose a logarithmic safety staffing rule, combined with a threshold control policy. The aforementioned three papers assume that both the inbound and outbound calls share the same set of servers. Deslauriers et al. [40] study five Markovian models of call centers with two sets of servers – inbound only and blend. In these Markovian models, the manager decides when to make outbound calls and how many as a function of the system state, guided under a threshold policy motivated by Bhulai and Koole [27].

Two key papers studying managing the callback option are Armony and Maglaras [8, 9]. Armony and Maglaras [8, 9] study a call center which offers the callback option to arriving customers who upon arrival choose among staying in the online queue, receiving a callback (i.e. joining an offline queue) and balking. In Armony and Maglaras [8], arriving customers only know the steady-state expected delay in the online queue and an (asymptotic) delay commitment for the offline queue, whereas the follow up paper Armony and Maglaras [9] considers the case where customers are provided with real time delay information. In both

papers, the authors provide a novel asymptotic analysis in the heavy traffic regime[7]. They characterize the (approximate) system equilibrium and derive asymptotically optimal staffing and priority rules in that asymptotic regime.Armony and Maglaras [8, 9] derive several insights. Most importantly, they observe that offering multiple channels of service (such as the callback option) can improve the call centers performance substantially. This is supported by our findings as well although our model and the style of analysis differ significantly from theirs. First, we consider a model of the arrival rate process that allows bursty arrivals. Second, we study a stable i.e. an underloaded system, whereas they focus on the critically loaded regime, i.e. the heavy traffic regime. Focusing on a system with a traffic intensity strictly less than one enables us to provide an exact analysis of the lookahead polices. In particular, in our setting the callback option is only relevant when the system experiences temporary (and unpredictable) surges in the call volume. Thus, the call center manager exercises the callback option judiciously as opposed to offering it throughout the day. To repeat, the question of when to offer the callback option is our main focus.

The recent paper of Legros et al. [78] studies when to offer the callback option using a Markov decision process model. The authors propose a threshold policy whereby the callback option is triggered whenever the online queue length exceeds the threshold. An arriving customer receives a delay announcement, and the callback offer (if the online queue length exceeds the threshold). Then he chooses between balking, joining the online queue and possibly the callback option (if offered) according to a probabilistic choice model. The customers in the online queue may abandon. The objective is to minimize the sum of waiting and abandonment costs. They show that the threshold policy is optimal when there are two servers and characterize a switching curve numerically that describes when to offer the callback option more generally. One important difference of our work from Legros et al. [78] and Armony and Maglaras [8, 9] is that we incorporate the time-varying and

7. Because Armony and Maglaras [8, 9] provide real time delay estimates, their analysis incorporates state-dependent arrival rate as well.

uncertain nature of the call arrival rate in our model. Whitt [111] studies a system in which some customers can tolerate substantial delays, similar to the system offering the callback option. In such system, demand can be smoothed by partitioning the service requests into separate priority classes according to their response-time requirement. Whitt [111] explores the benefit of demand intervention by estimating the capacity savings that can be obtained from partitioning time-varying demand into priority classes.

Instead of using the staffing level as a mechanism to manage the nonstationary random arrival rate, which may be too costly or even infeasible on the mesoscoptic time scale identified by Glynn et al. [50], we focus on managing the arrival process itself. The latter mechanism is related to the admission control literature; see Stidham [103] and Stidham [104] for overviews and references therein. The majority of the admission control literature formulate the problem as a Markov decision process. For example, Lewis et al. [80] consider an admission control problem for a finite buffer queue with Poisson arrivals and multiple exponential servers. Serving each admitted customer yields a reward, which is random ex-ante but observed upon arrival, before making the admission control decision. Moreover, the authors consider the more refined objectives of Bias and Blackwell optimality criteria, which can distinguish among policies that optimize the long-run average reward. Under these criteria, the authors show optimality of the trunk reservation policies which associate a separate threshold k_r for each r, and accepts a customer with reward r if the number of arrivals in the system is less than k_r. Ata and Shneorson [19] consider the admission and service rate control (by choosing the arrival and service rates as functions of the system state) for an $M/M/1$ queue to maximize the long-run average social welfare. The authors characterize the optimal admission and service rate control policy in closed form. Lewis et al. [81] generalize Lewis et al. [80] to the setting where the arrival rate, service rates, and system capacity are varying over time in a known fashion. The authors show the existence of optimal policies that are monotone in the number of customers in the system. Moreover, they also show the strong result that all bias optimal polices are of the (time-varying) threshold type. Yoon and Lewis

[116] also assume that the arrival and service rates are bounded, periodic functions of time. They show that under the infinite horizon discounted and average reward optimality criteria, for each fixed time, optimal pricing and admission control strategies are nondecreasing in the number of customers in the system. They propose an easily implementable pointwise stationary approximation (PSA) to approximate the optimal policies, suggest a heuristic to improve the implementation of the PSA and verify its usefulness via a numerical study.

Örmeci et al. [90] study an admission control problem for a two class loss system with periodically varying parameters. They show the existence of an optimal threshold policy, and that under certain conditions these thresholds have certain monotonicity properties. Zayas-Cabán and Lewis [117] extend this work to consider the system with abandonments. They show the optimality of a threshold policy for each fixed time and that simpler admission policies that ignore non-stationarity or abandonments lead to significant losses in average rewards. Hampshire and Massey [55] develop optimal control methods for non-stationary systems using the calculus of variations machinery. Cudina and Ramanan [37] consider a fluid control problem for a related single-class time-varying queue. They show that for a set of performance measures, any sequence of policies whose performance approach the infimum in the fluid control problem is asymptotically optimal (in a certain uniform acceleration regime).

There are several papers that study the admission control problem using diffusion approximations. Ward and Kumar [109] consider a G/G/1 queue with impatient customers in heavy traffic, and establish asymptotic optimality of a threshold policy; also see Rubino and Ata [95] for a similar analysis (via singular control) and Ata and Olsen [17] for a closed-form characterization of the admission control thresholds in the heavy traffic regime. Ata [12] considers an admission control problem for a multiclass queueing system with "thin" classes in the heavy traffic regime. The author shows the optimality of a nested threshold policy and characterizes the thresholds in closed form. Kocaga and Ward [75] consider an admission control problem for an $M/M/N+M$ queue when there are costs associated with

abandonments, server idleness and rejecting arrivals. The authors solve this problem using both the Markov decision processes and diffusion approximations in the Halfin-Whitt heavy traffic regime; and they show the (asymptotic) optimality of a threshold policy. Kocaga et al. [74] consider the optimal staffing and admission control for a queueing system where the arrival rate is random but not time varying. The authors propose a square-root staffing rule and a threshold policy for admission control, and show that this policy is asymptotically optimal. Our work takes a different approach and first studies the problem when the future information is available. Using a sample path analysis, we prove that a simple lookahead policy is pathwise optimal. Based on the insights glenned from this analysis, we propose a non-anticipating policy, the line policy, and show that it is effective in a simulation study calibrated by a data set from a US Bank's call center.

Another related paper in the admission control literature is Spencer et al. [101]. The authors study an admission control problem for an overloaded $M/M/1$ queue under the assumption that the future information is available. The objective is to minimize the long-run average queue length subject to a lower bound on the throughput. The authors propose the no-job-left-behind policy, which effectively rejects those jobs with "excessive" delay (hence, left behind) by looking into the future. The authors show that the no-job-left-behind policy is asymptotically optimal in heavy traffic. They also study a class of policies with a finite lookahead window, where the lookahead window grows logarithmically, and establish its asymptotic optimality. Xu and Chan [115] consider a similar model to manage the admission decisions into an emergency department using the knowledge of future arrivals. The authors propose policies that exercise control when one anticipates a high number of arrivals by looking into the future. They also enhance their policies using the thresholds on the queue length, which diverts arrivals if either the queue length is large or a high number of patients will arrive in future periods. The authors show that the proposed policy provides delay improvements over standard policies used in practice. They also consider the impact of errors in future information. They show that even with noisy predictions, the proposed

policies can still perform well. Our model differs from Spencer et al. [101] and Xu and Chan [115] in several ways. First, instead of studying an overloaded system, we assume that the system is stable, i.e. underloaded. However, it may experience temporary surges in the arrival rate due to the novel model of the arrival rate process we assume. The stability assumption simplifies the analysis, allowing us to conduct an exact analysis as opposed to an asymptotic analysis as done in Spencer et al. [101]. Indeed, we prove that the p/h-lookahead policy is pathwise optimal, a strong notion of optimality. In addition, our objective function is different. More importantly, we consider the arrival rate model that incorporates nonstationarity, randomness and autocorrelations, which are not modeled in Spencer et al. [101] and Xu and Chan [115]. Lastly, our model allows the customers to reject the callback offers. This feature of our model can be viewed as a job joining the system despite the system manager's desire to not admit it in an admission control problem. To the best of our knowledge, this important feature of our model is not considered in the admission control literature previously, including the aforementioned papers.

5.3 The Model

We consider a canonical single-server queueing model with a single class of homogenous customers[8]. The arrival times of customers are given by the (increasing) sequence $\{\tau_i : i \geq 1\}$, where τ_i denotes the i^{th} customer's arrival time. Letting $A(t)$ denote the cumulative number of arrivals by time t, it is given as follows:

$$A(t) = \sup\{i \geq 1 : \tau_i \leq t\}, \ t \geq 0,$$

where we set $\sup \emptyset = 0$ for notational convenience. As the reader will see below, our results hold for general point processes provided that the system is stable. However, because of its practical importance, particular attention will be paid to the recent arrival model proposed

8. Our model can be generalized in various directions. For example, we expect most of our results continue to hold for multiserver systems.

by Glynn et al. [50]. To be specific, the arrival process is a Poisson process with the stochastic intensity process $\{\lambda(t) : t \geq 0\}$. The stochastic intensity process will also be referred to as the arrival rate process. Glynn et al. [50] models the arrival rate process as a Cox-Ingersoll-Ross (CIR) process, and provides empirical support for using it as a model of arrivals to a call center. Following Glynn et al. [50][9], we assume that the arrival rate process $\{\lambda(t) : \ t \geq 0\}$ follows a CIR process. In particular, it satisfies the following stochastic differential equation:

$$\mathrm{d}\lambda(t) = a(b - \lambda(t))\,\mathrm{d}t + \sigma\sqrt{\lambda(t)}\,\mathrm{d}W(t), \ t \geq 0, \tag{5.1}$$

where a, b and σ are positive constants and $\{W(t) : \ t \geq 0\}$ is a standard Brownian motion. As mentioned above, the arrival process $A(\cdot)$ is a Poisson process with the stochastic intensity process $\{\lambda(t) : t \geq 0\}$ given in Equation (5.1); see page 23 of Bremaud [32] or page 476 of Jeanblanc et al. [65] for the formal definition. The service times of the customers are given by the sequence $\{\nu_j : j \geq 1\}$ of i.i.d. exponential random variables with rate μ, where ν_j denotes the time between two consecutive service completions. Note that the service times are associated with the server not with the customers. That is, the time ν_j (for $j \geq 1$) is independent of the index of the customer served. We assume that the service times and the arrival process are independent. Letting $S(t)$ denote the total number of customers served by time t if the server works continuously over $[0, t]$, it is given as follows:

$$S(t) = \sup\left\{n \geq 1 : \sum_{i=1}^{n} \nu_i \leq t\right\}, \ t \geq 0.$$

We restrict attention to work-conserving policies. We assume that $b < \mu$ so that the system is stable under the work-conserving policies. Nonetheless, it may experience temporary surges in the customer arrival rate, because the arrival rate changes over time as a stochastic process.

9. For simplicity, our model of the arrival rate process $\{\lambda(t) : t \geq 0\}$ ignores the time-of-day effect. Although it is straightforward to incorporate the time-of-day effect into the CIR model for our (numerical) analysis, we do not see that as central to the paper, because we are mainly concerned with managing the variations in the arrival rate on the mesoscopic time scale. Moreover, as noted above, our results on the lookahead policies hold for general arrival processes, subsuming those with the time-of-day effect.

To accommodate such surges in arrivals, the system offers the callback option to arriving customers. The call center manager's problem can be described as follows: When a customer arrives at the system, she observes the system and decides whether to offer the callback option to the incoming customer or not. The incoming customer receiving no callback option is routed to the online queue. If the customer receives the callback option, she decides whether to accept the offer or not. If she accepts the offer, she is routed to the offline queue to be called back; otherwise, she joins the online queue.

We refer the call center manager's decision of offering the callback option as the callback policy in the foregoing development. We represent this callback policy with the sequence $\mathcal{I} = \{i_k : k \geq 1\}$ of the indices of the customers receiving the callback option. Let the sequence $\mathcal{A} = \{j_k : k \geq 1\}$ denote the indices of customers who are willing to accept the callback offer, whereas its complement $\mathcal{R} = \{1, 2, \ldots, \}\backslash\mathcal{A}$ denotes the sequence of customers who will reject it (if offered) and stay in the online queue. The sequence $\mathcal{A}$ (and thus its complement $\mathcal{R}$) is not known ex ante. The call center manager only observes the customers' accept/reject decisions when a callback offer is made. In addition, let the sequences $\mathcal{I}_1$ and $\mathcal{I}_2$ denote the indices of the customers routed to the online and offline queue, respectively. Note that $\mathcal{I}_2 = \{1, 2, \ldots\}\backslash\mathcal{I}_1$ and that only the customers in set $\mathcal{A}$ may join the offline queue, which happens if they are offered the callback option, i.e.

$$\mathcal{I}_2 = \mathcal{I} \cap \mathcal{A}. \tag{5.2}$$

This state of affairs can equivalently be described as the call center manager routing the arriving customers to one of two queues: the online queue versus the offline queue; see Figure 5.2. Given a callback policy $\mathcal{I}$, let $A_1(t)$ and $A_2(t)$ denote the cumulative numbers of

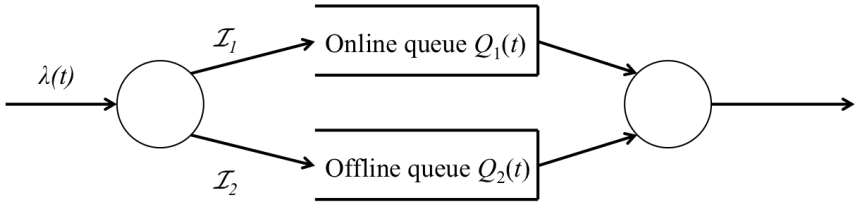


Figure 5.2: A two-class queue model for the queue with the callback option.

customers routed to the online and offline queues up to time t, respectively. The cumulative number of customers routed to the online queue up to time t, i.e. $A_1(t)$, is given as follows:

$$A_1(t) = \sup\left\{k : i_k \in \mathcal{I}_1,\ \tau_{i_k} \leq t\right\},\ t \geq 0, \tag{5.3}$$

where we set $\sup \emptyset = 0$ for notational convenience. In addition, we have that $A_2(t) = A(t) - A_1(t)$ for $t \geq 0$.

If the customer is routed to the online queue, he waits for service in the online queue. The waiting cost per time unit is h. If the customer is routed to the offline queue, he will hang up and wait for the system to call him back. There is a one-time penalty of p for sending a customer to the offline queue. In particular, the call center may provide a callback time window or a delay commitment D to the customers. That is, the system promises that he will wait at most D time units in the offline queue. In what follows, we assume D is sufficiently large that given the mean-reverting nature of the CIR process (see Equation (5.1)) and that $b < \mu$, this requirement will be met with high probability. Thus, we will relax the delay commitment below. However, our numerical analysis will provide guidance as to how D should be set so that this relaxation is reasonable in practice. For example, as can be seen from Tables 5.2 and 5.3, delays in the offline queue are in the order of 30 minutes to an hour; see Section 5.6 for further details.

In addition to the callback policy, the call center manager also decides on how to prioritize the customers waiting in the online and offline queues. Given the cost structure described above, it is more or less obvious that for any reasonable callback policy, it is optimal to give strict (preemptive) priority to the online queue. Thus, we focus on the work-conserving service policy that gives strict priority to the online customers. For simplicity, we also allow preemption. In addition, we assume that both the online and offline queues are served in a FCFS fashion. Let $Q_1(\cdot)$ and $Q_2(\cdot)$ denote the online and offline queue length processes, respectively. In addition, let $Q(t)$ denote the total number of customers in the system at time t. We note that the evolutions of $Q(\cdot)$, $Q_1(\cdot)$ and $Q_2(\cdot)$ are fully characterized by

the primitive processes $\{A(t) : t \geq 0\}$ and $\{S(t) : t \geq 0\}$, the callback policy $\mathcal{I}$ and the strict priority service policy; see Appendix C.1 for the system equations characterizing their evolutions.

Thus, the long-run average cost of a callback policy $\mathcal{I}$, denoted by $C^{\mathcal{I}}$, is given as follows:

$$C^{\mathcal{I}} = \overline{\lim_{t\to\infty}} \frac{1}{t}\mathbb{E}\left[pA_2(t) + h\int_0^t Q_1(s)\,\mathrm{d}s\right].$$

The system manager strives to choose a policy $\mathcal{I}$ that minimizes $C^{\mathcal{I}}$. In what follows, we consider the complete foresight policies, where the call center manager knows the sequences of arrival times $\{\tau_i : i \geq 1\}$ and service times $\{\nu_j : j \geq 1\}$ in Section 5.4. Based on the insights gleaned from the optimal lookahead policies and a fluid model advanced in Section 5.5, we also propose a non-anticipating policy, the line policy, and show that it has excellent performance via a simulation study presented in Section 5.6 using data from a US bank call center.

5.4 The Optimal Policy with Complete Foresight

This section considers the system with complete foresight. To be specific, the call center manager knows the sequences of arrival times $\{\tau_i : i \geq 1\}$ and service times $\{\nu_j : j \geq 1\}$. In Section 5.4.1, we assume for simplicity that all customers who are offered the callback option will accept it, i.e. $\mathcal{R} = \emptyset$. We show that a simple lookahead policy that looks only into the next p/h time units is optimal. In Section 5.4.2, we relax this assumption to allow customers to turn down the callback offers and show that a modified lookahead policy that also looks into the next p/h time units only is optimal.

Fixing a sample path, one can view the arrival and service completion processes $\{A(t) : t \geq 0\}$, $\{S(t) : t \geq 0\}$ and $\{T(t) : t \geq 0\}$ as known functions. Without loss of generality, it suffices to analyze a busy period of the system. Because the system is stable, the number of arrivals in a busy period is finite almost surely. Fix a busy period and assume there are n

arrivals in that busy period, whose arrival times are given by $\tau_1, \tau_2, \ldots, \tau_n$.

Recall that $\mathcal{A} \subseteq \{1, 2, \ldots, n\}$ is the set of customers who are willing to accept the callback offer, whereas its complement $\mathcal{R} = \{1, 2, \ldots, n\} \backslash \mathcal{A}$ is the set of customers who will reject it and stay in the online queue. The set of customer receiving the callback option is denoted by the set $\mathcal{I}$, referred as the callback policy. Note that the customers in set $\mathcal{I} \cap \mathcal{A}$ will be removed from the online queue and join the offline queue. We denote $Q_1^{\mathcal{I}} = \Phi(Q, \mathcal{I} \cap \mathcal{A})$ the (online) queue length process resulting from the removal of customers in the set $\mathcal{I} \cap \mathcal{A}$ from the online queue; the operator $\Phi(\cdot, \cdot)$ is defined suitably to represent this operation. Thus, the optimal callback policy seeks the set $\mathcal{I} \subseteq \{1, 2, \ldots, n\}$ of customers to offer the callback option so as to minimize the total cost during this busy period, i.e.

$$\min_{\mathcal{I} \subseteq \{1,2,\ldots,n\}} p|\mathcal{I} \cap \mathcal{A}| + hH(\Phi(Q, \mathcal{I} \cap \mathcal{A})), \tag{5.4}$$

where $|\mathcal{I}|$ denotes the number of elements in the set $\mathcal{I}$ and $hH(\Phi(Q, \mathcal{I} \cap \mathcal{A})$ denotes the holding cost during the busy period associated with the resulting queue length process $\Phi(Q, \mathcal{I} \cap \mathcal{A})$.

We first consider the special case when all customers receiving the callback option take the offers, i.e. $\mathcal{A} = \{1, 2, \ldots, n\}$. Under this assumption, the set of customers routed to the offline queue is equivalent to the set of those receiving the callback option, i.e. $\mathcal{I} \cap \mathcal{A} = \mathcal{I}$. We construct a simple lookahead policy and prove its pathwise optimality for this special case. Building on this result, we also prove the optimality of a modified lookahead policy for a general set $\mathcal{A}$.

5.4.1 The optimal policy when all customers accept the callback option

This subsection studies the special case of $\mathcal{A} = \{1, 2, \ldots, n\}$. Thus, Eqaution (5.4) simplifies to the following:

$$\min_{\mathcal{I} \subseteq \{1,2,\ldots,n\}} p|\mathcal{I}| + hH(\Phi(Q, \mathcal{I})). \tag{5.5}$$

To facilitate the analysis to follow, define a sequence of times $\{s_i : \ i \geq 1\}$ as follows:

$$s_i = \inf\{t \geq \tau_i : Q(t) = Q(\tau_i-)\}, \ i \geq 1. \tag{5.6}$$

Recall that τ_i is the arrival time of the i^{th} customer. Then, s_i denotes the next time (after customer i's arrival) when the number of customers in the system $Q(t)$ falls back to the level just before his arrival[10]. As an aside note that the evolution of $Q(t)$ is independent of the callback policy or the service policy as long as it is work conserving. We next define a class of policies that the optimal policy belongs to.

Definition 6. *(w-lookahead policy) A w-lookahead policy routes customer i to the online queue, i.e. $i \in \mathcal{I}$, if $s_i \geq \tau_i + w$.*

Figure 5.3 provides the routing decisions of two customers when the system is operated under the p/h-lookahead policy for a given sample path. The main result of this section,

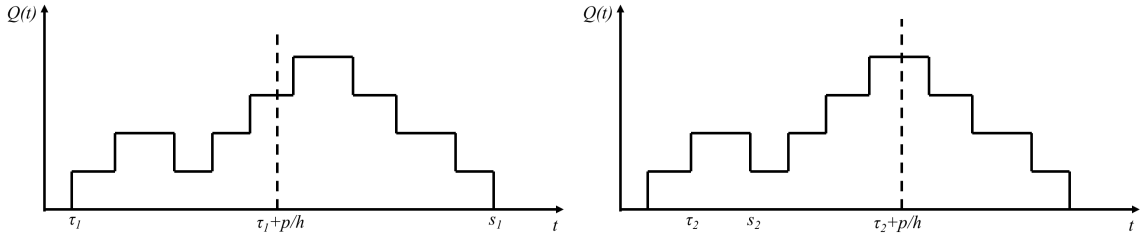


Figure 5.3: The routing decisions of two customers when the system is operated under the p/h-lookahead policy for a given sample path. Left panel: Customer 1 is routed to the offline queue because $s_1 \geq \tau_1 + p/h$. Right panel: Customer 2 stays in the online queue because $s_2 < \tau_2 + p/h$.

stated next, proves that such a policy is optimal.

Theorem 3. *The p/h-lookahead policy is optimal.*

The rest of this subsection is dedicated to prove Theorem 3. In what follows, we first construct a policy that determines the sequence of customers to be routed to the offline queue in a greedy fashion (Definition 7). Second, we show that this greedy policy is optimal

10. Note that $Q(\tau_i-) = Q(\tau_i) - 1$.

(see Lemma 18). Third, we analyze a virtual system in which customers in the online queue are served under last-come-first-served (LCFS) policy. We show a key property of the greedy policy, which leads to a much simpler characterization of it (see Lemma 20) in this virtual system. Indeed, this step is the "secret sauce" of our proof. Fourth, this property and the optimality of the greedy policy is used to deduce the optimality of the p/h-lookahead policy. Lastly, we argue that the p/h-lookahead policy is optimal in the original system (where customers in the online queue are served under first-come-first-served (FCFS) fashion as well.

The greedy policy is defined iteratively. First, assume that all customers are routed to the online queue. In this system, denoted by superscript $\emptyset$, the online queue length process $Q_1^{\emptyset}(t)$ coincides with the total number of customers in the system. That is,

$$Q_1^{\emptyset}(t) = Q(t),\ t \geq 0.$$

Next, the K-greedy policy, which is defined formally as follow, gives the indices of K customers who are routed to the offline queue.

Definition 7. *(K-greedy policy). The K-greedy policy chooses K customers to be routed to the offline queue as follows*[11]*:*

$$i_k^* = \text{argmax}_{i \in \mathcal{I}_{k-1}^C} \left[H(Q_1^{k-1}) - H(\Phi(Q_1^{k-1}, \{i\})) \right],\ k = 1, \ldots, K, \tag{5.7}$$

where i_k^* denotes the index of the k^{th} customer routed to the offline queue, and $\mathcal{I}_k = \mathcal{I}_{k-1} \cup \{i_k^*\}$, $Q_1^k = \Phi(Q_1^{k-1}, \{i_k^*\})$ with $\mathcal{I}_0 = \emptyset$ and $Q_1^0 = Q_1^{\emptyset}$.

The K-greedy policy first picks customer i_1^* whose removal results in the largest holding cost savings. Removing customer i_1^* from the online queue results in the new online queue length process Q_1^1. The policy repeats this process until K customers are removed from the queue. The following lemma shows that the K-greedy policy maximizes the total holding

11. If there are multiple indices that achieve the maximum, we pick the smallest index.

cost savings among all policies that remove K customers; see Lemma 6 of Spencer et al. [101] and its proof for a proof of this result.

Lemma 18. *For any $K \leq n$ and $\mathcal{I} \subseteq \{1, 2, \ldots, n\}$ with $|\mathcal{I}| = K$, the following holds:*

$$H(Q_1^{\emptyset}) - H(Q_1^K) \geq H(Q_1^{\emptyset}) - H(\Phi(Q_1^{\emptyset}, \mathcal{I})).$$

That is, $H(Q_1^K) \leq H(\Phi(Q_1^{\emptyset}, \mathcal{I}))$.

Thus, the problem simplifies to finding the optimal greedy policy. That is,

$$\min_{K \in \{1,2,\ldots,n\}} pK + hH(Q_1^K). \tag{5.8}$$

Finding the optimal value K boils down to comparing the marginal cost p of routing a customer to the offline queue and the marginal saving of waiting cost of the line queue, which equals to $H(Q_1^{K-1}) - H(Q_1^K)$. The marginal saving $H(Q_1^{K-1}) - H(Q_1^K)$ includes two parts: The waiting cost customer i_K^* would have incurred if he were not routed to the offline queue and the negative externalities he imposes on others in the online queue. To characterize the potential marginal saving, one needs to compute the (intermediate auxiliary) queue length processes $Q_1^1, \ldots, Q_1^K$. Surprisingly, we next provide an explicit construction of the optimal greedy policy that only uses the initial online queue-length process $Q_1^{\emptyset}(t)$, or equivalently, the total queue length for the system $Q(t)$. This crucially relieves us from computing the (intermediate auxiliary) queue length processes $Q_1^1, \ldots, Q_1^K$.

Given a callback policy, (and that the server gives strict priority to the online queue), the evolution of the online queue $Q_1(t)$ is independent of the service discipline used to prioritize customers within the online queue. Therefore, although first-come-first-served (FCFS) is the natural service discipline, any other (work-conserving) service discipline will result in the same evolution of the online queue length $Q_1(t)$, because the service times are associated with the server not with the customers as discussed above. As observed by Hassin [57], under the last-come-first-served (LCFS) service discipline, a customer joining the (online) queue does not impose any externalities on others. Rather, the LCFS service discipline

ensures that he internalizes all such externalities that would have been incurred under the FCFS service discipline. Therefore, in what follows we study a virtual system that serves the online queue in a LCFS fashion, which will clearly have the same performance as the original system. Then for this system, we identify the optimal routing policy and argue that the same routing policy is also optimal for the original system which serves the online queue in a FCFS fashion. Thus, we next focus on this virtual system that uses LCFS to prioritize the customers in the online queue and show that the externality is accounted in the waiting time (under LCFS) of the customer i_K^* routed to the offline queue in the K^{th} iteration.

As a preliminary to finding the optimal routing policy, we first characterize the waiting times of the customers in the system where no one is routed to the offline queue. To this end, recall that the time s_i (for $i \geq 1$), defined in Equation (5.6), denotes the next time (after customer i joins the system) when the total number of customers in the system $Q(t)$ falls back to the level just before customer i's arrival. Then the next lemma helps characterize the waiting times of customers in the system with no callbacks.

Lemma 19. *In the system with no callbacks, customer i completes his service at time s_i under the LCFS discipline. Thus, the waiting time of customer i, denoted by w_i, is given by*

$$w_i = s_i - \tau_i, \ i = 1, \ldots, n. \tag{5.9}$$

This result has long been recognized in the queueing theory literature; see for example, page 321 of Grass and Harris [51]. Recall that $H(Q)$ denote the total area under $Q(t)$ during the busy period. It is easy to show that

$$H(Q_1^{\emptyset}) = \sum_{i=1}^{n} w_i.$$

which can be proved geometrically by exploiting the equivalence of tracing the area either vertically or horizontally in calculating it. To this end, consider the k^{th} iteration of the K-greedy policy ($k \leq K$) and the resulting online queue length process Q_1^k. Let w_i^k denote the waiting time of the customer $i \in \mathcal{I}_k^C$, i.e. those customers who are still in the online

queue after the k^{th} iteration. Then the holding cost for the online queue Q_1^k after the k^{th} iteration is given as follows:

$$H(Q_1^k) = \sum_{i \in \mathcal{I}_k^C} w_i^k \quad \text{for } k = 1, \ldots, n. \tag{5.10}$$

The following lemma is key to our results and shows that under the K-greedy policy, the waiting times of the customers remaining in the online queue are the same as those in the original system; see Appendix C.2.1 for its proof.

Lemma 20. *Given $K \leq n$, we have that $w_i^k = w_i$ for $i \in \mathcal{I}_k^C$ and $k = 0, 1, \ldots, K$.*

We conclude from Lemma 20 that choosing $K^* = |\mathcal{I}^*|$, where

$$\mathcal{I}^* = \{i = 1, 2, \ldots, n : w_i > p/h\}, \tag{5.11}$$

the K^*-greedy policy is optimal. To see this, note from Lemma 20 and Equation (5.10) that

$$H(Q_1^k) = \sum_{i \in \mathcal{I}_k^C} w_i \quad \text{for } k = 1, \ldots, n.$$

Substituting this into Equation (5.7) of Definition 7 gives that

$$i_k^* = \text{argmax}_{i \in \mathcal{I}_k^C} w_i, \ k = 1, \ldots, K. \tag{5.12}$$

Thus, the K-greedy policy chooses the customers with the largest K waiting times w_i in the original system to route to the offline queue. Given the formulation in Equation (5.8), i.e. the penalty p for removing customer i from the online queue and the holding cost savings of hw_i, the optimal policy should remove customer i as long as his waiting time w_i exceeds p/h. This leaves us with the ultimate conclusion that the customer in the set $\mathcal{I}^*$ should be sent to the offline queue. A few important observations are in order. First, all that is relevant to determine the set $\mathcal{I}^*$ is the set of waiting times $\{w_i : i = 1, \ldots, n\}$ which are determined by

$$w_i = s_i - \tau_i, \ i = 1, \ldots, n,$$

as in Equations (5.6)-(5.9). Second, the calculation of w_i depends only on the total queue

length process $\{Q(t) : t \geq 0\}$ (see Equation (5.6)), which is invariant to the routing, priority and service discipline (within each queue) decisions. Third, what matters ultimately is that the waiting times of customers who remain in the system are less than or equal to p/h. These observations prove that not only the order of removal from the online queue does not matter (in particular, the removal decision can be made upon a customer's arrival) but also the optimal set $\mathcal{I}^*$ of customers routed to the offline queue is precisely the same as the one under the p/h-lookahead policy.

The last observation is that the evolution of the resulting online queue length process $Q_1(\cdot)$ is invariant to the service discipline used to prioritize customers within the online queue. To be specific, if we fix a routing policy $\mathcal{I}$, the evolution of the resulting online queue length process $Q_1(\cdot)$ (using LCFS) is identical to that of the original system, which serves the customers within the online queue in a FCFS fashion, because the service times are associated with the server not with the customers as discussed above. Thus, with a fixed routing policy $\mathcal{I}$, the average cost (see Equation (5.5)) of the system that uses LCFS is the same as the average cost of the original system. Since the p/h-lookahead policy minimizes the average cost of the system that uses LCFS among the complete foresight policies, it also minimizes the average cost of the original system. This completes the proof of Theorem 3.

The next section extends our model, allowing the customers to turn down the callback offer.

5.4.2 The optimal policy when customers can reject the callback option

One important assumption made in the preceding analysis is that all customers accept the callback offer. In practice, some customers may reject the callback option and prefer to wait online. As such, this section relaxes the earlier assumption, allowing customers to reject the callback offer. In this setting, we propose a suitably modified lookahead policy and prove its pathwise optimality for a general set $\mathcal{A} \subseteq \{1, 2, \ldots, n\}$.

To facilitate the policy description, recall that $Q(t)$ denotes the total number of customers

in the system (either in the online or offline queues) at time t. Also, let $Q_r(t)$ denote the number of customers in the online queue at time t who have rejected the callback offer. Then defining the random time

$$s_i^r = \inf \{t \geq \tau_i : Q(t) = Q(\tau_i-) - Q_r(\tau_i-)\}, \ i = 1, 2, \ldots, n, \tag{5.13}$$

the following definition introduces the proposed policy.

Definition 8. *The modified w-lookahead policy offers the callback option to customer i if $s_i^r \geq \tau_i + w$ for $i = 1, 2, \ldots, n$.*

The modified w-lookahead policy reduces to the w-lookahead policy if no customer rejects the callback option, in which case $Q_r(t) = 0$ for all $t \geq 0$. Another key property of the modified w-lookahead policy is that it only uses the revealed preferences of customers regarding the callback offers. Namely, it does not assume that we know the preferences of those customers who are not offered the callback option. Neither does it require the knowledge of future customers' preferences. In other words, when a customer arrives, the call center manager does not know if he belongs to set $\mathcal{A}$ or set $\mathcal{R}$. She only finds out if she offers the callback option, because customers in the set $\mathcal{A}$ would accept, whereas those in set $\mathcal{R}$ would reject. Although the call center manager does not know customers' preference ex ante, she adjusts her callback decisions after observing rejections of the callback option under the modified w-lookahead policy. To be specific, Equation (5.13) implies that the call center manager is more likely to offer the callback option to incoming customers after observing more rejections.

The following theorem establishes that the modified p/h-lookahead policy under rejection is optimal. The rest of this subsection is dedicated to proving Theorem 4.

Theorem 4. *The modified p/h-lookahead policy is optimal under possible rejections of the callback offers.*

To facilitate our analysis, we let $Q_n(t)$ denote the number of customers in the online

queue who did not receive the callback offer, i.e.

$$Q_n(t) = Q_1(t) - Q_r(t),\ t \geq 0. \tag{5.14}$$

In what follows, we consider two auxiliary systems and the optimal lookahead policies for them. In both auxiliary systems, the online queue is split into two sub-queues: queue r and queue n. In the first auxiliary system, the sets $\mathcal{A}$ and $\mathcal{R}$ are known and all customers in set $\mathcal{R}$ are sent to queue r and they have the highest priority. In particular, they do not "see" the rest of the system. Using this and focusing on the rest of the system, one can use the earlier results to show the optimality of the modified lookahead policy, which also serves as a lower bound for the original system because it assumes the sets $\mathcal{A}$ and $\mathcal{R}$ are observable.

The main difference between the two auxiliary systems is how they route the customers in set $\mathcal{R}$. Nonetheless, we show that in either system once the call center manager stops offering the callback option, she never offers it again until the online queue becomes empty. Using this, we show that in both systems the same customers join the online queue; and the online and offline queues have identical sample paths (Lemma 26). Thus, they incur the same cost.

The second auxiliary system serves as a bridge between the original system and the first auxiliary system. The main difference between the original system and the second auxiliary system is how customers in the online queue are prioritized. However, in both systems, we show that when the system is empty, the call center manager either offers the callback option to the next customer or not. In the latter case, no arriving customer receives the callback offer until the online queue becomes empty again. In the former case, the call center manager can continue to offer the callback option. However, once she stops, she does not offer it to any future customers until the online queue becomes empty. Using this structure, we show that both systems end up giving strict priority to the customers who rejected the callback offer over others, which in turn, leads to identical sample paths for the queue length processes (Lemma 25).

Combining these observations, we conclude that the three lookahead policies (described below) will result in identical sample paths for the online and offline queue length processes in the three systems. Hence, they have the same cost, establishing the optimality of the modified lookahead policy in the original system.

Next we describe the first auxiliary system in detail. We attach a ' ~ ' (' ^ ') to various quantities of interest in the first (second) auxiliary system. Consider a system with three queues: Queue r, queue n and queue 2, where $\tilde{Q}_r(t)$, $\tilde{Q}_n(t)$ and $\tilde{Q}_2(t)$ denote the numbers of customers in each queue at time t, respectively. We assume that the sets $\mathcal{A}$ and $\mathcal{R}$ are known to the call center manager ex ante in the first auxiliary system; and all customers in set $\mathcal{R}$ are routed to queue r. In contrast, a customer in set $\mathcal{A}$ joins queue n if he does not receive the callback offer, whereas he joins queue 2 (the offline queue) if he receives the callback offer. (Recall that all customers in set $\mathcal{A}$ accept the callback offer). In the first auxiliary system, the customers are served under a work-conserving static priority rule (with preemption) where queue r has the highest priority, queue n has the second-highest priority whereas queue 2 has the lowest priority. For notational convenience, we define $\tilde{Q}_1(t)$ as the total number of customers in the online queue (consisting of queue r and queue a) and $\tilde{Q}_a(t)$ as the total number of $\mathcal{A}$-customers (combining queue a and queue 2) in the system at time t. That is,

$$\tilde{Q}_1(t) = \tilde{Q}_r(t) + \tilde{Q}_n(t) \quad \text{and} \quad \tilde{Q}_a(t) = \tilde{Q}_n(t) + \tilde{Q}_2(t), \ t \geq 0. \tag{5.15}$$

Because both the original system and the first auxiliary system operate under work-conserving policies, the total numbers of customers are the same in both system, i.e.

$$Q(t) = \tilde{Q}_r(t) + \tilde{Q}_a(t), \ t \geq 0. \tag{5.16}$$

One critical observation is that the dynamics of queue r, i.e. $\{\tilde{Q}_r(t), t \geq 0\}$ is independent of the routing policy $\mathcal{I} \subseteq \mathcal{A}$, because its arrival process $\{\tau_i : i \in \mathcal{R}\}$ is independent of the routing policy and queue r enjoys the highest priority among all queues.

Next we define the first auxiliary w-lookahead policy and show that it is optimal for the

first auxiliary system, i.e. it results in a set $\tilde{\mathcal{I}} \subseteq \mathcal{A}$ of customers to be routed to the offline queue that minimizes the objective given in Equation (5.4). To facilitate that definition, we first introduce the random time $\tilde{s}_i$ as follows:

$$\tilde{s}_i = \inf\{t \geq \tau_i : \tilde{Q}_a(t) = \tilde{Q}_a(\tau_i-)\}. \tag{5.17}$$

Definition 9. *The first auxiliary w-lookahead policy offers the callback option to customer i if $i \in \mathcal{A}$ and $\tilde{s}_i \geq \tau_i + w$.*

This policy can be viewed as applying the optimal lookahead policy (see Definition 6) to the process $\{\tilde{Q}_a(t) : t \geq 0\}$ instead of the process $\{Q(t) : t \geq 0\}$. Combining this with the fact that the queue r essentially does not "see" the rest of the system because it enjoys the highest (preemptive) priority makes it more-or-less obvious that the first auxiliary p/h-lookahead policy is optimal for the first auxiliary system as proved in the next lemma; see Appendix C.2.2 for its proof.

Lemma 21. *The first auxiliary p/h-lookahead policy is optimal for the first auxiliary system.*

We also provide an additional characterization of the random time $\tilde{s}_i$ (for $i \in \mathcal{A}$) in the next lemma; see also Appendix C.2.2 for its proof.

Lemma 22. *The following holds:*

$$\tilde{s}_i = \inf\left\{t \geq \tau_i : Q(t) = Q(\tau_i-) - \tilde{Q}_r(\tau_i-)\right\} \quad \textit{for} \;\; i \in \mathcal{A}. \tag{5.18}$$

Combining Lemmas 21 and 22 with Definitions 8 and 9 shows that the modified p/h-lookahead policy is optimal for the first auxiliary system.

Although the first auxiliary p/h-lookahead policy is optimal, it assumes that (and crucially uses) the knowledge of the set $\mathcal{A}$ and $\mathcal{R}$ apriori. That is, the call center manager knows who will accept and reject the callback option before offering it in the first auxiliary system. Next, we consider the second auxiliary system which relaxes this assumption. To be specific, the call center manager does not know the set $\mathcal{A}$ and $\mathcal{R}$ ex ante in the second auxiliary system. As the reader will see below, the second auxiliary system (and its optimal

policy) serve as a bridge between the original system (and its proposed policy) and the first auxiliary system (and its optimal policy).

The second auxiliary system also has three queues: Queue r, queue n and queue 2 (the offline queue), and the number of customers in them at time t are denoted by $\hat{Q}_r(\cdot)$, $\hat{Q}_n(\cdot)$ and $\hat{Q}_2(\cdot)$, respectively. The system operates under a work-conserving static priority rule (with preemption) where queue r has the highest priority, queue n has the second highest priority and queue 2 (the offline queue) has the lowest priority. Moreover, customers within queue r are served in a FCFS fashion, whereas customers within queue n are served in a LCFS fashion.

As mentioned earlier, the second auxiliary system serves as a bridge between the original system and the first auxiliary system. On the one hand, the key difference between the first and the second auxiliary systems is how the customers in set $\mathcal{R}$ are routed. In the second auxiliary system, they join queue r if and only if they are offered the callback option. In particular, they join queue n if they are not offered the callback option. On the other hand, the major difference between the second auxiliary system and the original system is the service discipline used to prioritize customers within the online queue. In the second auxiliary system, customers in queue r have (preemptive) priority over those in queue n, whereas all customers in the online queue (which consists of queue n and queue r) are served in FCFS fashion in the original system.

To facilitate the analysis of the second auxiliary system, we also refer to the combinations of the queue r and queue n as the online queue, or queue 1. In particular, we have that

$$\hat{Q}_1(t) = \hat{Q}_r(t) + \hat{Q}_n(t),\ t \geq 0. \tag{5.19}$$

To facilitate the description of the second auxiliary w-lookahead policy (for the second auxiliary system), we define the random time $\hat{s}_i$ as follows:

$$\hat{s}_i = \inf\left\{t \geq \tau_i : Q(t) = Q(\tau_i-) - \hat{Q}_r(\tau_i-)\right\},\ i = 1, 2, \ldots, n. \tag{5.20}$$

Definition 10. *The second auxiliary w-lookahead policy offers the callback option to cus-*

tomer i if $\hat{s}_i \geq \tau_i + w$.

Remark. If $\hat{s}_i < \tau_i + w$, then customer i joins queue n. Otherwise, i.e. $\hat{s}_i \geq \tau_i + w$, the callback option is offered. Then the customer rejects it and joins queue r if $i \in \mathcal{R}$, whereas he accepts it and joins queue 2 (the offline queue) if $i \in \mathcal{A}$. □

To repeat, letting $\hat{\mathcal{I}}$ denote the set of customers who are offered the callback option, i.e.

$$\hat{\mathcal{I}} = \{i = 1, 2, \ldots, n : \hat{s}_i \geq w + \tau_i\}. \tag{5.21}$$

The customers who join the offline queue (queue 2) are given by $\hat{\mathcal{I}} \cap \mathcal{A}$. All customers in $\hat{\mathcal{I}} \cap \mathcal{R}$ joins queue r. Customer $i \notin \hat{\mathcal{I}}$ joins queue n. In particular, a customer $i \in \mathcal{R}$ may join queue n (in contrast to the first auxiliary system) if he is not offered the callback option.

The rest of this section is dedicated to proving that all three policies result in the same set of customers routed to the offline queue. Consequently, they all have the same cost. Moreover, because the first auxiliary p/h-lookahead policy minimizes the objective in Equation (5.4), the modified p/h-lookahead policy is optimal.

As a preliminary, we first prove useful properties of the first and second auxiliary w-lookahead policies in Lemmas 23 and 24; see Appendix C.2.2 for their proofs.

Lemma 23. *Consider the first auxiliary system under the first auxiliary p/h-lookahead policy. If customer $i \in \mathcal{A}$ is routed to the online queue, i.e. $i \notin \tilde{\mathcal{I}}$, then any other customer $j \in \mathcal{A}$ arriving after him but before his departure, i.e. $\tau_j \in (\tau_i, \tilde{s}_i)$, is also routed to the online queue, i.e. $j \notin \tilde{\mathcal{I}}$. Moreover, queue r is empty when customer i departs the system, i.e. $\tilde{Q}_r(\tilde{s}_i) = 0$.*

Lemma 24. *Consider the second auxiliary system under the second auxiliary p/h-lookahead policy. The following holds:*

(i) If customer i is routed to the online queue, i.e. $i \notin \hat{\mathcal{I}}$, then any customer j arriving after him but before his departure, i.e. $\tau_j \in (\tau_i, \hat{s}_i)$, is also routed to the online queue, i.e. $j \notin \hat{\mathcal{I}}$.

(ii) If customer i is routed to the online queue, i.e. $i \notin \hat{\mathcal{I}}$, then he leaves the system at time $\hat{s}_i$. Moreover, queue r is empty at his departure time, i.e. $\hat{Q}_r(\hat{s}_i) = 0$ and $\hat{Q}_n(\hat{s}_i) = \hat{Q}_n(\tau_i-)$.

(iii) If customer i is offered the callback option, i.e. $i \in \hat{\mathcal{I}}$, then queue n is empty upon his arrival, i.e. $\hat{Q}_n(\tau_i-) = \hat{Q}_n(\tau_i) = 0$.

Property (i) of Lemma 24 shows that if a customer is not offered the callback option, the call center manager does not offer it to any other customer arriving after him until his departure[12]. In addition, property (ii) of Lemma 24 implies that when this customer leaves the system, queue r is empty and the number of customers in queue n ($\hat{Q}_n$) falls back to the level just before his arrival. Lastly, property (iii) of Lemma 24 shows that if the system offers the callback option, the queue n must be empty.

These three observations imply that the system alternates between the callback and no-callback episodes. In a callback episode, the call center manager offers the callback option to every arriving customer. In contrast, she does not offer the callback option to any arriving customer during a no-callback episode. When the system is in a callback episode, queue n is empty (i.e. $\hat{Q}_n = 0$) and all customers currently in the online queue are those who rejected the callback option; see property (iii) of Lemma 24. Thus, we arrive at the following key observation: If both queues n and r are non-empty (i.e. $\hat{Q}_n > 0$ and $\hat{Q}_r > 0$), then the system is in the no-callback episode. Moreover, all customers in queue r currently must have arrived during the callback episode prior to the current no-callback episode. In other words, all customers in queue r have arrived before those in queue n. That is, the strict priority rule is never really enforced to reverse what would have happened under FCFS service discipline. Therefore, although the second auxiliary system gives priority to queue r over queue n, the evolution of the queue length processes would have been the same if the online queue

12. In particular, when both queues r and n are non-empty, the call center manager no longer offers the callback option to any incoming customers. This main reason explaining this decision is that with future information, the call center manager has already exercised preventive intervention by offering the callback option to customers arriving earlier.

(the combination of queues r and n) was served in a FCFS fashion (as done in the original system). This is because customers in queue r arrived before customers in queue n due to the aforementioned structure of the callback and no-callback episodes. This observation is the main intuition behind the equivalence of the second auxiliary system and the original system, in which the online queue is served in a FCFS fashion.

To elaborate further on the evolution of the second auxiliary system under the second auxiliary lookahead policy, note that once the system is in the no-callback episode, it stays in the no-callback episode until the online queue is empty. In particular, once the system is in the no-callback episode, no arriving customer is offered the callback option until the online queue is empty; see properties (i) and (ii) of Lemma 24. Moreover, once the online queue is empty, the system can enter either a no-callback episode or a callback episode. In the former case, the callback episode lasts until the online queue becomes empty. In the latter case, either the callback episode lasts until the online queue is empty or it can be followed by a no-callback episode, which lasts until the online queue becomes empty. To summarize, there are three possible combinations of callback and no-callback episodes for a busy period of the online queue: just one no-callback episode, just one callback episode, or a callback episode followed by a no-callback episode.

The following lemma formally establishes the equivalence of the second auxiliary system and the original system; see Appendix C.2.2 for its proof.

Lemma 25. *The evolutions of the queue length processes in the original system and the second auxiliary system are the same. That is,*

$$Q_r(t) = \hat{Q}_r(t),\ Q_n(t) = \hat{Q}_n(t) \quad \textit{and} \quad Q_2(t) = \hat{Q}_2(t),\ t \geq 0. \tag{5.22}$$

It is immediate from Lemma 25 and Definitions 8 and 10 that the second auxiliary p/h-lookahead policy results in the same callback offers as using the modified p/h-lookahead policy in the second auxiliary system.

Next, we show that the same set of customers are routed to the offline queue in the two auxiliary systems in the following lemma; see Appendix C.2.2 for its proof. Hence, the two auxiliary systems incur the same cost. Consequently, it follows from Lemma 25 that the original system has the same cost as well.

Lemma 26. *The same set of customers are routed to the offline queue in the two auxiliary systems, i.e.* $\tilde{\mathcal{I}} = \tilde{\mathcal{I}} \cap \mathcal{A} = \hat{\mathcal{I}} \cap \mathcal{A}$. *Consequently, we have that*

$$\tilde{Q}_1(t) = \hat{Q}_1(t) \quad \textit{and} \quad \tilde{Q}_2(t) = \hat{Q}_2(t), \ t \geq 0.$$

Lemmas 25-26 show that fixing a lookahead window $w > 0$, the three lookahead policies defined in this subsection results in identical sample paths for the online and offline queue length processes. Also recall from Lemma 21 that the first auxiliary p/h-lookahead policy minimizes the objective in Equation (5.4). Thus, so does the modified p/h-lookahead policy in the original system, which completes the proof of Theorem 4.

5.5 The Line Policy: A Non-anticipating Policy Based on the Insights from the Lookahead Policy

The key idea behind the lookahead policies is to compare the benefits and costs of routing an arriving customer to either the offline queue or the online queue. Routing a customer to the online queue results in an increase in the holding costs while avoiding the penalty p that would have been incurred if he were routed to the offline queue. The crux of the analysis in Section 5.4 is the characterization of the marginal increase in the holding cost by routing an arriving customer to the online queue.

In this section, we consider the non-anticipating policies and focus attention on characterizing the marginal increase in the holding cost by routing an arriving customer to the online queue. In general, it is difficult to characterize the change of the average queue length

(and thus the marginal cost) analytically[13]. However, it is intuitive that the higher the current queue length and the arrival rate, the more likely that the marginal cost of routing an additional customer to the online queue will exceed the penalty p; and ultimately we propose a policy that is monotone in the queue length and the arrival rate.

To derive an effective non-anticipating policy, first consider the lookahead policies studied in Section 5.4 and the insights they offer. Recall that under the lookahead policies, the system alternates between the callback and no-callback episodes. Moreover, once the system enters a no-callback episode, it stays there until the online queue is empty. First, consider the setting of Section 5.4.1 where all customers accept the callback offer. In this setting, any time a callback offer is made the online queue is empty and the call center manager makes the offer in anticipation of future congestion. To be specific, the call center manager checks to see whether the time $s_i - \tau_i$ for the total number of customers in the system (both in the online and offline queue) to fall back to the current level is longer than p/h. Intuitively, the higher the current arrival rate, the longer it would take for the total queue length to fall back to the current level. This motivates a threshold policy on the arrival rate in the non-anticipating case. Although this is a valuable insight, it does not address the general case. In particular, in the setting of Section 5.4.2, where customers can reject the callback offer, the modified lookahead policy (see Definition 8 and Equation (5.13)) may offer the callback option while having the customers in the online queue. In that case, the modified lookahead policy looks at how long it would take the total backlog to drop below the current level minus the current number of customers in the online queue[14] to decide whether to offer the callback option. The expected time to serve the current number of customers in the online queue is $Q_1(t)/\mu$. Moreover, the higher the current arrival rate, the longer it would take to get back to the current level. Combining these two suggests a policy of the following

13. The difficulty comes from the negative externality imposed on others by each customer joining the queue, which has been long recognized in the queueing literature; see for example Section 1.8.1 of Hassin [58] for a more detailed discussion.

14. Recall that all such customers at the start of a callback episode would have rejected the callback offer previously.

form: Offer the callback option if

$$\frac{Q_1(t)}{\mu} + A_1\lambda(t) \geq B_1,\ t \geq 0,$$

where $A_1, B_1 > 0$ are tuning parameters. The policy can equivalently be expressed as follows: Offer the callback option if

$$Q_1(t) + A_2\lambda(t) \geq B_2,\ t \geq 0, \tag{5.23}$$

where $A_2, B_2 > 0$ are tuning parameters. Notice that if $Q_1(t) = 0$, then this policy reduces to a threshold policy on the arrival rate as discussed above.

Next, we derive the policy given in Equation (5.23) by interpreting the modified lookahead policy directly in the context of a fluid model, where the arrival rate process, denoted by $\bar{\lambda}(\cdot)$, follows a deterministic process given as follows:

$$\bar{\lambda}'(t) = a(b - \bar{\lambda}(t)),\ t \geq 0. \tag{5.24}$$

This differential equation ignores the volatility term in Equation (5.1). To be more specific, we focus on the case $\bar{\lambda}(0) = \lambda > b$, which is when the callback option is most useful[15]. Solving for $\bar{\lambda}(\cdot)$ yields:

$$\bar{\lambda}(t) = b\left(1 - e^{-at}\right) + \lambda e^{-at},\ t \geq 0. \tag{5.25}$$

Note that $\bar{\lambda}(t)$ is a unbiased estimate of $\lambda(t)$, i.e. $\bar{\lambda}(t) = \mathbb{E}[\lambda(t)|\lambda(0) = \lambda]$; and it decreases to b monotonically.

In addition, we assume that the total number of customers in the system, denoted by $\bar{q}(\cdot)$, evolves deterministically. Its evolution is governed by the following differential equation: For $t \geq 0$,

$$\bar{q}'(t) = \begin{cases} \bar{\lambda}(t) - \mu, & \text{if } \bar{q}(t) > 0, \\ 0, & \text{if } \bar{q}(t) = 0, \end{cases} \tag{5.26}$$

with $\bar{q}(0) = q_0 \geq 0$.

15. Otherwise, i.e. $\lambda(0) = \lambda < b < \mu$, and as can be seen from Equation (5.25) that $\lambda(t) < b < \mu$ for all $t \geq 0$ and the queue length steadily decreases.

Given that the arrival rate is decreasing to $b < \mu$, we expect the queue length $\bar{q}(t)$ to decrease eventually to zero. Letting T_0 denote the first time the queue length becomes zero, i.e.

$$T_0 = \inf\{t \geq 0 : \bar{q}(t) = 0\},$$

we have the following characterization for the queue length process:

$$\bar{q}(t) = \begin{cases} q_0 + (b-\mu)t + \frac{\lambda - b}{a}\left(1 - e^{-at}\right), & \text{if } t \leq T_0, \\ 0, & \text{if } t > T_0. \end{cases} \tag{5.27}$$

Consider the original system under the modified lookahead policy. Recall that the system alternative between the callback and no-callback episodes, and the once it enters a no-callback episode it stays there until the online queue becomes empty. Thus, let us consider the system in a callback episode and contemplate whether to offer the callback option to the next arriving customer, say customer i. Also note that all customers in the online queue currently are those who have previously rejected a callback offer, i.e.

$$Q_1(\tau_i-) = Q_r(\tau_i-). \tag{5.28}$$

The modified lookahead policy offers the callback option if $s_i^r < \tau_i + p/h$, where

$$s_i^r = \inf\{t \geq \tau_i : Q(t) = Q(\tau_i-) - Q_r(\tau_i-)\}.$$

Setting $t_0 = \tau_i-$ and $t^* = s_i^r$, and using Equation (5.28), we note that t^* satisfies the following in the fluid model:

$$\bar{q}(t^*) = \bar{q}(t_0) - Q_1(t_0).$$

Then using Equation (5.27) gives the following:

$$\bar{q}(t_0) + (b-\mu)(t^* - t_0) + \frac{\lambda(t_0) - b}{a}\left(1 - e^{-a(t^* - t_0)}\right) = \bar{q}(t_0) - Q_1(t_0).$$

Simplifying this and assuming that $a(t^* - t_0)$ is not small allow us to approximate t^* as

follows:

$$t^* - t_0 \approx \frac{Q_1(t_0) + (\lambda(t_0) - b)/a}{\mu - b}.$$

Thus, interpreting the modified lookahead policy in the context of the fluid model gives rise to the policy that offer the callback option if

$$\frac{Q_1(t_0) + (\lambda(t_0) - b)/a}{\mu - b} > p/h, \tag{5.29}$$

which can be expressed as in Equation (5.23).

Lastly, we consider an alternative approach to derive the same policy. This approach is motivated by the recent work of Wang [110] that proposes studying a particular priority rule to accurately characterize the congestion externalities imposed on others by a customer admitted to the system. To build intuition, consider the following simple example, where the arrival rate takes only two values, i.e. $\lambda(t) \in \{\lambda_L, \lambda_H\}$ with $\lambda_L < \lambda_H$. A question that facilities our analysis is whether an arriving customer should be admitted when the arrival rate is λ_H or he should be routed to the offline queue. One can view this as a two-class system, where class 1 customers arrive when the arrival rate is λ_L and class 2 customers arrive when the arrival rate is λ_H. We assume that the server gives class 1 priority (with preemption). Let w_k denote the average time for class k (for $k = 1, 2$).

One important observation is that the average waiting time of the customers in class 1 is invariant to the routing decisions of the customers in class 2, because class 1 customers enjoy strict priority (with preemption) over class 2 customers. Therefore, the marginal cost of routing class 2 customers to the online queue is $h\bar{w}_2$. Consequently, class 2 customers should be routed to the online queue if and only if $\bar{w}_2 < p/h$.

Although one can compute $\bar{w}_1$ and $\bar{w}_2$ in this simple example (under additional assumptions), to the best of our knowledge such (closed-form) expressions are not available for more general arrival models. Therefore, we explore this question further using a fluid model advanced above.

To be specific, we consider the question of whether to route the customers who arrive when $\lambda(t) \in (\lambda - \epsilon, \lambda)$ for ϵ small. Recall that the key insight from preceding two class example was to prioritize the admitted customers in a particular way to account for the externalities[16]. Therefore, to assess the marginal cost of routing them to the online queue, we consider the preemptive strict priority service policy that gives them the lowest priority as motivated by the two class example.

Thus, using the fluid model to approximate the evolution of the online queue, the marginal cost of routing them to the online queue is hT_0, and we should do so if $hT_0 < p$, where T_0 corresponds to the first time the online queue becomes empty. Assuming aT_0 is not small, we use Equation (5.27) to approximate T_0 as follows:

$$T_0 \approx \frac{q_0 + (\lambda - b)/a}{\mu - b}.$$

Thus, our proposed policy is offer the callback option at time t if any only if

$$T_0 \approx \frac{Q_1(t) + (\lambda(t) - b)/a}{\mu - b} \geq \frac{p}{h}, \tag{5.30}$$

which is of the same form as the policy proposed via Equation (5.29). Rearranging the terms, Equation (5.23) can be expressed as follows:

$$Q_1(t) + A\lambda(t) \geq B,$$

where $A = 1/a$ and $B = b/a + p(\mu - b)/h$. Thus, we propose the following non-anticipating policy formally, referred as the line policy.

Definition 11. *(The line policy) The call center manager offers the callback option to a customer arriving at time t if $Q_1(t) + c_1\lambda(t) \geq c_2$, where c_1, c_2 are positive constants viewed as tuning parameters.*

Throughout the paper, we assume that the current arrival rate is observed whereas it can

16. Incidentally, the use of LCFS service discipline in the online queue of various parts of the proof in Section 5.4 is due to the same intuition.

only be estimated from the observed customer arrivals. As a robustness check, we consider the performance of the line policy where the current arrival rate is estimated by the average number of arrivals over the past five minutes. As one would expect for large call volumes, this lead to a close approximation of the original findings for the case of observable arrival rate in a simulate study[17] (available from the authors).

The next section provides a numerical study calibrated using data from a US Bank call center, which shows that the line policy achieves excellent performance among non-anticipating policies.

5.6 Simulation Study

This section uses a data set from a US bank call center to study the performance of the call center under the lookahead and the line policies. In this study, we first assume that the customers always take the callback option if they are offered one. We compare the various performance metrics under the lookahead and line policies. We then study the system when the customers are allowed to reject the offered callback option and choose to wait in the online queue instead. Next, we assume that the customers waiting in the online queue may abandon. We study the impact of the abandonments on the performance of the lookahead and line policies.

We use the individual call level data of a US bank call center[18] to study the system with the callback option. To be specific, we analyze the call arrival data of brokerage customers in February 2003. We focus on those customers who arrive during the peak hours (9am-2pm) in the weekdays and request the service from the agents. The summary statistics for this portion of the data are given in Table 5.1.

We assume that the arrival process follows a Poisson process with its intensity following a CIR process as described in Section 5.3. We use the Bayesian approach to estimate the

17. The maximum difference in performance was 0.3%.

18. This data set is publicly available at Service Enterprise Engineering (SEE) Center, Technion.

# of observations	38,392
Average waiting time (sec)	21.38
Average service time (sec)	233.94
% of abandonments	2.77%

Table 5.1: The summary statistics of the data.

parameters that characterizing the arrival process via Markov Chain Monte Carlo (MCMC) method[19], i.e. the constants a, b and σ in Equation (5.1). To be specific, we assume that the prior distribution of the parameters a, b and σ follow a Gaussian, Gamma and inverse Gamma distribution, respectively. We first draw the initial samples of the parameters from the prior distributions. We then implement the Gibbs sampler to generate new samples of the parameters until the sampled parameters converge. Lastly, we use the mean of the sampled parameters as their estimates.

Next, we calibrate the number of agents to replicate the current performance of the call center without the callback option. We simulate the system and vary the number of agents using the arrival processes with the estimated parameters a, b and σ[20], the empirical service time and abandonment time distributions[21]. We compare the simulated average waiting time and fraction of abandoning customers with the data and pick the number of agents to be 30.

First, we analyze the performance of the proposed policies in a baseline model which assumes that the customers waiting in the online queue do not abandon. In addition, the customers always take the callback option and join the offline queue if they are offered one. Recall that Theorem 3 shows that the p/h-lookahead policy is optimal under these two assumptions. We also propose a line policy that can be implemented easily when the future

19. Zhang [119] uses the same approach to estimate the parameters for a similar model.

20. We use the exact simulation method proposed in Giesecke et al. (2011) for a Poisson process with stochastic intensity to simulate the arrival process.

21. To be specific, we use the Kaplan-Meier estimate to estimate the abandonment time distribution. As observed in Brown et al. (2005), the Kaplan-Meier estimate may be biased under heavy center. Therefore, we assume that the abandonment time follows the exponential distribution and follow Brown et al. (2005) to use the first quartile of the Kaplan-Meier estimate of the cumulative distribution function to estimate the hazard rate.

information is not available in Section 5.5. We run a discrete event simulation to study the performance of these two policies for various values of p/h (assuming $h = 1$). For a fixed p/h value, we consider the p/h-lookahead policy (which is optimal). For the line policy, we run a two-dimensional search to find the tuning parameters c_1 and c_2 in Definition 11 that minimizes the average cost. In addition, under Markovian assumptions, we formulate the problem as a Markov Decision Problem (MDP) and solve it numerically to obtain the best non-anticipating routing policy. This serves as a benchmark to analyze the performance of the line policy among all non-anticipating policies. We summarize the key performance metrics, e.g. the fraction of customers routed to the offline queue, average waiting times of the online and offline queues, in Table 5.2. The average waiting time is 79.43 seconds in the base case without offering the callback option.

Lookahead Policy (LH)					
p/h (min)	% of offline cust.	Online wait. (sec)	Offline wait. (min)	Ave. Cost	
0.5	20.09%	8.18	5.94	\$6.33	
1.0	12.65%	11.46	9.19	\$ 10.79	
1.5	9.31%	13.94	11.75	\$13.95	
2.0	7.20%	15.95	13.78	\$ 16.21	
3.0	5.08%	19.17	19.35	\$ 19.94	
4.0	3.88%	21.61	25.40	\$ 22.59	
5.0	3.05%	23.49	30.74	\$ 24.38	
MDP					
p/h (min)	% of offline cust.	Online wait. (sec)	Offline wait. (min)	Ave. Cost	Change from LH
0.5	26.01%	11.17	4.59	\$10.30	62.06%
1.0	13.61%	16.23	7.97	\$ 15.46	43.24%
1.5	9.81%	18.93	10.52	\$18.88	35.29%
2.0	6.68%	22.26	14.50	\$ 21.51	32.72%
3.0	4.71%	25.32	21.18	\$ 25.18	26.31%
4.0	3.75%	27.38	29.32	\$ 27.84	23.27%
5.0	2.64%	30.10	33.31	\$29.66	21.66%
Line Policy (LP)					
p/h (min)	% of offline cust.	Online wait. (sec)	Offline wait. (min)	Ave. Cost	Change from MDP
0.5	25.82%	11.25	4.56	\$10.31	0.05%
1.0	13.56%	16.28	7.98	\$ 15.47	0.06%
1.5	9.77%	19.00	10.55	\$18.91	0.16%
2.0	6.67%	22.33	14.54	\$ 21.57	0.25%
3.0	4.22%	26.31	21.27	\$ 25.33	0.58%
4.0	2.85%	29.66	29.37	\$ 28.08	0.87 %
5.0	2.44%	31.02	33.38	\$29.97	1.04%

Table 5.2: The fraction of customers sent to the offline queue, average waiting times of the online and offline queues under LH, LP and MDP for various p/h values (assuming $h = 1$).

Table 5.2 shows that the average waiting time of the online queue reduces significantly by routing a small fraction of the customers to the offline queue under all three policies. In other words, the callback option allows the call center manager to smooth the temporary arrival surges effectively. In addition, the waiting times of the offline queue under all three policies are reasonable by the industry standards; see the survey conducted by Software Advise[22]. This result is driven by the mean-reverting nature of the arrival process (and that the system is stable) that causes the arrival surges at the mesoscopic scale. After the temporary surges end which last from minutes to half-hour, the call center is able to use its excess capacity to serve the offline customers.

Table 5.2 also compares the average cost of the system under all three policies. The lookahead policy performs significantly better than the MDP solution for all p/h values because it uses the future information effectively. The line policy achieves the near-optimal performance among all non-anticipating policies for all p/h values.

In practice, it may be difficult to obtain the accurate estimate of the value of p/h. As shown in Table 5.2, the performance of the online queue improves as we send more customers to the offline queue (when the value of p/h decreases). Thus, choosing the p/h value is equivalent to finding the tradeoff between the fraction of offline customers and the performance of the online queue. Figure 5.4 provides three examples of such tradeoff curves. To be specific, it shows the tradeoffs between the average, 80-percentile and 90-percentile of the waiting time of the online queue under the three polices. The call center manager can pick the point that matches the specific service quality requirement of the online queue and find the fraction of customers that the call center needs to offer the callback option to. By deciding the point on the tradeoff curves, the call center manager can determine the p/h value and the parameters of the line policy.

Second, we study the performance of the lookahead policy and the line policy when the customers are allowed to reject the offered callback option. We fix $p/h = 3$ minutes and vary

22. The results can be found at http://csi.softwareadvice.com/3-ways-to-offer-callback-0614/.

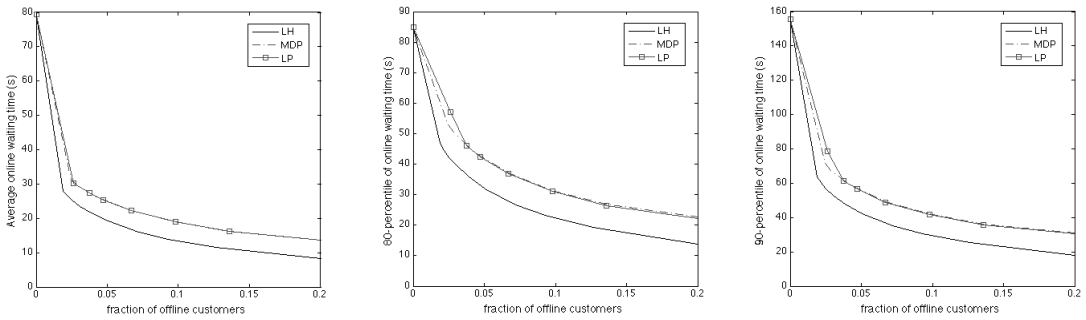


Figure 5.4: The trade-off curves of the fraction of customers sent to the offline queue and online waiting time under various policies.

the acceptance rate q, i.e. the fraction of customers that may accept the callback option if they are offered one, from 5% to 50%. In particular, we assume that if a customer accepts the callback offer with probability q if he is offered one in the simulation, though our model in Section 5.4.2 is more general. Table 5.3 summarizes the fraction of customers routed to the offline queue, average waiting times of the online and offline queues for various acceptance rates.

When the acceptance rate is not too small (i.e. $q \geq 20\%$), the fraction of offline customers and the average online waiting time under all policies are nearly the same as the ones of the no rejection case (the first study). In other words, when the acceptance rate is not too small, the system is insensitive to the acceptance rate. When the acceptance rate is very low (i.e. $q \leq 10\%$), the online waiting time increases significantly and the lookahead policy does not work as well as it does in the high acceptance cases. The main reason is that with low acceptance rate, the call center manager has to offer the callback option to many customers and has less control on who ends up in the offline queue.

Next, we study the impact of the abandonments on the performance of these policies. We consider three abandonment scenarios: High, medium and low abandonment scenarios. We assume that the abandonment time distribution and the service capacity are unchanged. The only difference of these three scenarios are the arrival processes. We use the data from December 2002, January 2003, and February 2003 of the system to estimate the parameters

Lookahead Policy				
% of Accepting	% of offline cust.	Online wait. (s)	Offline wait. (min)	Ave. Cost
5%	1.47%	48.88	37.08	$43.12
10%	2.47%	35.53	31.16	$31.50
20%	3.64%	22.59	25.51	$24.09
30%	4.25%	22.72	23.19	$21.93
50%	4.78%	20.36	21.42	$20.57
100%	5.08%	19.17	19.35	$19.94
MDP				
% of Accepting	% of offline cust.	Online wait. (s)	Offline wait. (min)	Ave. Cost
5%	1.03%	53.03	46.23	$46.62
10%	1.73%	40.38	39.98	$35.13
20%	2.54%	32.77	31.75	$28.92
30%	3.13%	30.05	26.73	$27.20
50%	3.65%	27.92	22.68	$25.96
100%	4.71%	25.32	21.18	$25.18
Line Policy				
% of Accepting	% of offline cust.	Online wait. (s)	Offline wait. (min)	Ave. Cost
5%	0.88%	53.12	48.65	$46.66
10%	1.45%	40.64	41.80	$35.27
20%	2.16%	32.95	32.93	$29.03
30%	2.57%	30.28	28.43	$27.30
50%	3.06%	28.16	23.79	$26.10
100%	3.73%	26.31	21.27	$25.33

Table 5.3: The fraction of customers routed to the offline queue, average waiting times of the online and offline queues under LH, LP and MDP for various rejection rates (assuming $p/h = 3$ minutes).

a, b and σ for the three scenarios[23]. The system is operated under the lookahead policy or the line policy calculated from the system with no abandonment. Table 5.4 summarizes the simulated performance metrics of the system under the lookahead and line policies in the low abandonment scenario.

Comparing the performance metrics shown in Tables 5.2 and 5.4, we conclude that the performance metrics of the no-abandonment system are close to those of a system with low abandonments for both the lookahead and line policies. Therefore, the lookahead and line policies continue to do well when the abandonment rate is low. Moreover, they also lower

23. The arrival volumes of the brokerage customers decreased significantly from December 2002 to February 2003 while the service capacity did not change significantly. Thus, we use the data of these three months to study systems with different abandonment scenarios. The arrival process and the service time in low abandonment case are the same as the no-abandonment study. We simulate the abandonments on top of the simulation of the baseline model. To study the medium and high abandonment scenarios, we conduct the same calibration using the data from January 2003 and December 2002, respectively.

Lookahead Policy (LH)				
p/h (min)	% of aban.	% of offline cust.	Online wait. (s)	Offline wait. (min)
N/A	2.91%	–	36.20	–
0.5	0.52%	20.08%	8.12	5.83
1.0	0.79%	12.66%	11.32	8.47
1.5	0.99%	9.31%	13.68	18.61
2.0	1.16%	7.24%	15.56	12.60
3.0	1.40%	5.06%	18.41	17.61
4.0	1.57%	3.86%	20.57	19.34
5.0	1.54%	3.08%	22.17	22.98
Line Policy (LP)				
p/h (min)	% of aban.	% of offline cust.	Online wait. (s)	Offline wait. (min)
0.5	0.66%	25.55%	11.14	4.11
1.0	1.11%	13.16%	15.97	6.56
1.5	1.34%	9.34%	18.54	9.18
2.0	1.62%	6.07%	21.56	10.91
3.0	1.92%	3.59%	25.01	14.39
4.0	2.16%	2.25%	27.76	19.77
5.0	2.25%	1.82%	28.73	22.87

Table 5.4: The fraction of abandoning customers, the fraction of customers sent to the offline queue, average waiting times of the online and offline queues under LH and LP for various p/h values under the low abandonment scenario (assuming $h = 1$).

the abandonments.

A related question is that how well the line policy does if one quantifies the cost of abandonments. Thus, we focus on minimizing the sum of holding cost, callback penalties and the abandonment costs under the long-run average criterion. Assuming each abandonment costs 500 (which is larger than the largest p/h value considered), we compare the cost of the line policy and the solution to the MDP formulation with abandonments. The comparison is summarized in Table 5.5, which shows that the line policy still achieves the near-optimal.

p/h (min)	0.5	1.0	1.5	2.0	3.0	4.0	5.0
Cost of MDP	\$18.97	\$26.97	\$31.43	\$34.95	\$38.77	\$41.87	\$43.01
Cost of LP	\$19.20	\$27.12	\$31.67	\$ 35.20	\$ 39.67	\$42.52	\$44.09
Relative Diff.	1.19%	0.57%	0.76%	0.71%	0.23%	0.15%	0.25%

Table 5.5: The cost of MDP and LP for various p/h values (assuming that $h = 1$ and the abandonment cost equals to 500).

Table 5.6 summarizes the performance metrics of the lookahead and line polices in the medium abandonment scenario. It shows that the average online waiting time does not

Lookahead Policy (LH)				
p/h (min)	% of aban.	% of offline cust.	Online wait. (s)	Offline wait. (hrs)
N/A	7.24%	–	88.42	–
0.5	0.47%	25.93%	7.79	2.09
1.0	0.72%	18.91%	11.48	2.04
1.5	0.92%	15.54%	13.44	2.04
2.0	1.08%	13.57%	15.59	2.01
3.0	1.33%	11.27%	18.76	2.04
4.0	1.56%	10.01%	21.42	2.09
5.0	1.70%	8.98%	23.47	2.10
Line Policy (LP)				
p/h (min)	% of aban.	% of offline cust.	Online wait. (s)	Offline wait. (hrs)
0.5	0.63%	32.16%	11.55	2.50
1.0	0.95%	23.10%	15.58	3.21
1.5	1.27%	17.93%	19.29	3.74
2.0	1.79%	12.73%	25.57	4.30
3.0	2.31%	9.45%	31.95	6.08
4.0	2.79%	7.37%	37.89	6.57
5.0	3.03%	6.59%	40.63	6.52

Table 5.6: The fraction of abandoning customers, the fraction of customers sent to the offline queue, average waiting times of the online and offline queues under LH and LP for various p/h values under the medium abandonment scenario (assuming $h = 1$).

change much from the no abandonment case. However, the fraction of customers routed to the offline increases significantly because the callback option keeps many customers in the system who would abandon in the system with no callback option. This leads to a significant reduction in the fraction of abandoning customers under both policies. However, it also leads to a significant increase in the waiting time of the offline queue.

Table 5.7 summarizes the performance metrics of the lookahead and line polices in the high abandonment scenario. On the one hand, the simulation shows that the average waiting time of the online queue under the high abandonment scenario does not change significantly compared to the no abandonment case. On the other hand, the fraction of customers routed to the offline queue increases significantly. Thus, the system is operated under the overloaded regime. In this scenario, although the callback option ensures excellent performance of the online queue, some customers in the offline queue will never be served. Therefore, unless the call center manager uses busy signals to divert some arriving customers or increases the capacity to accommodate the increase in the system load due to the reduction in

Lookahead Policy (LH)				
p/h (min)	% of aban.	% of offline cust.	Online wait. (s)	Offline wait. (hrs)
N/A	19.07%	–	244.58	–
0.5	0.40%	33.20%	7.34	∞
1.0	0.61%	27.41%	10.28	∞
1.5	0.75%	24.89%	12.46	∞
2.0	0.88%	23.21%	14.25	∞
3.0	1.08%	21.62%	16.93	∞
4.0	1.21%	20.71%	18.94	∞
5.0	1.33%	19.95%	20.63	∞
Line Policy (LP)				
p/h (min)	% of aban.	% of offline cust.	Online wait. (s)	Offline wait. (hrs)
0.5	0.59%	39.60%	11.95	∞
1.0	0.73%	35.68%	14.21	∞
1.5	1.10%	29.01%	19.33	∞
2.0	1.86%	22.67%	30.21	∞
3.0	2.54%	19.59%	39.75	∞
4.0	3.24%	17.55%	49.57	∞
5.0	3.59%	17.01%	54.72	∞

Table 5.7: The fraction of abandoning customers, the fraction of customers sent to the offline queue, average waiting times of the online and offline queues under LH and LP for various p/h values under the high abandonment scenario (assuming $h = 1$).

abandonments, the callback option will not be effective when the abandonment rate is high.

5.7 Concluding Remarks

In this chapter, we study a call center in which its manager can offer the callback option to incoming customers when the system is congested. We prove that a simple lookahead policy is optimal in the complete foresight system. We then extend the model to the setting where customers can choose to accept or reject the callback offer. We show that the modified lookahead policy is optimal. Building on the intuition and analysis of the optimal policy in the system with complete foresight, we propose the line policy, which achieves near optimal performance among non-anticipating policies.

Given the mean-reverting nature of the arrival process, our simulation study shows that offering callback option to incoming customers improves the performance of the online queue significantly with a reasonable delay of the offline queue. Although our model does not incorporate abandonments, the simulation study shows that the lookahead and line policies

continue to do well provided the abandonment rate is low. However, when the abandonment rate is high, the callback option is not effective, because the callback option lowers the abandonment rate significantly, thereby increasing the system load.

APPENDIX A

APPENDIX OF CHAPTERS 2-3

A.1 Proofs of Lemmas, Propositions, Corollary and Theorems in Chapters 2-3

A.1.1 Proofs of Results in Section 2.3.2

Proof of Lemma ??. Substituting equation (2.5) into equation (2.9) yields the following: For $k = 1, \ldots, K$ and $w \geq 1$,

$$\begin{aligned} q_k(w) &= \mathbb{P}(\varepsilon_k(1) - \varepsilon_k(0) \geq -c_k + \alpha \left\{\beta_k(w) r_k + (1 - \beta_k(w)) J_k(w+1)\right\}) \\ &= \bar{F}_k \left(-c_k + \alpha \left\{\beta_k(w) r_k + (1 - \beta_k(w)) J_k(w+1)\right\}\right). \end{aligned}$$

□

Proof of Proposition 1. Fixing $k = 1, \ldots, K$ and $\beta(\cdot)_k$, define $T_{k,\beta_k} : l^\infty \to l^\infty$ as follows: For $w \geq 1$,

$$\begin{aligned} &(T_{k,\beta_k} \circ x)(w) \\ &= \mathbb{E}_{\varepsilon_k}[\max\{\varepsilon_k(1), -c_k + \alpha[\beta_k(w) r_k + (1 - \beta_k(w)) x(w+1)] + \varepsilon_k(0)\}] \\ &= \mathbb{E}_{\varepsilon_k}[-c_k + \alpha[\beta_k(w) r_k + (1 - \beta_k(w)) x(w+1)] - (\varepsilon_k(1) - \varepsilon_k(0))]^+ + \mathbb{E}_{\varepsilon_k}[\varepsilon_k(1)] \\ &= \mathbb{E}_{\varepsilon_k}[-c_k + \alpha[\beta_k(w) r_k + (1 - \beta_k(w)) x(w+1)] - (\varepsilon_k(1) - \varepsilon_k(0))]^+, \end{aligned} \tag{A.1}$$

where l^∞ is the space of bounded sequences of real numbers. The last equality follows from Assumption 1 that $\mathbb{E}_{\varepsilon_k}[\varepsilon_k(1)] = 0$. Note that T_{k,β_k} is the right-hand side of equation (2.8). Therefore, the expected discounted utility function J_k is the fixed point of operator T_{k,β_k}, i.e. $J_k = T_{k,\beta_k}(J_k)$ for $k = 1, \ldots, K$.

We use the Blackwell's sufficient conditions for a contraction [Theorem 3.3, 105] to show

that the operator T_{k,β_k} is a contraction[1].

First, we check that if $J_k^1(w) \leq J_k^2(w)$ (for $w \geq 1$), then $T_{k,\beta_k} \circ J_k^1(w) \leq T_{k,\beta_k} \circ J_k^2(w)$ for $w \geq 1$. The following holds: For $w \geq 1$,

$$(1 - \beta_k(w))J_k^1(w+1) \leq (1 - \beta_k(w))J_k^2(w+1),$$

because $1 - \beta_k(w) \geq 0$ and $J_k^1(w+1) \leq J_k^2(w+1)$. Since the inequality is preserved by the max operator and the expectation, it follows that $T_{k,\beta_k} \circ J_k^1(w) \leq T_{k,\beta_k} \circ J_k^2(w)$ for $w \geq 1$.

Then we show that $T_{k,\beta_k} \circ (J_k + e)(w) \leq T_{k,\beta_k} \circ J_k(w) + \alpha e$ for $e > 0$ and $w \geq 1$. It follows that

$$\begin{aligned}
&(T_{k,\beta_k} \circ (J_k + e))(w) \\
&= \mathbb{E}_{\varepsilon_k}\left[\max\{\varepsilon_k(1), -c_k + \alpha[\beta_k(w)r_k + (1 - \beta_k(w))[J_k(w+1) + e] + \varepsilon_k(0)\}\right] \\
&\leq \mathbb{E}_{\varepsilon_k}\left[\max\{\varepsilon_k(1), -c_k + \alpha[\beta_k(w)r_k + (1 - \beta_k(w))J_k(w+1)] + \varepsilon_k(0) + \alpha e\}\right] \\
&\leq \mathbb{E}_{\varepsilon_k}\left[\max\{\varepsilon_k(1) + \alpha e, -c_k + \alpha[\beta_k(w)r_k + (1 - \beta_k(w))J_k(w+1)] + \varepsilon_k(0) + \alpha e\}\right] \\
&= \mathbb{E}_{\varepsilon_k}\left[\max\{\varepsilon_k(1), -c_k + \alpha[\beta_k(w)r_k + (1 - \beta_k(w))J_k(w+1)] + \varepsilon_k(0)\}\right] + \alpha e \\
&= T_{k,\beta_k} \circ J_k(w) + \alpha e.
\end{aligned}$$

The first inequality holds because $(1 - \beta_k(w))e \leq e$. Hence, the two sufficient conditions of Blackwell are satisfied, i.e. T_{k,β_k} is a contraction mapping. It follows from Banach fixed point theorem that there exists a unique fixed point of $x = T_{k,\beta_k} \circ x$; see Theorem 3.2 in Stokey and Lucas [105]. Since the solution to equation (2.8) is equivalent to the fixed point of $J_k = T_{k,\beta_k} \circ J_k$, the solution is unique.

□

Proof of Lemma 2. Fix $k = 1, \ldots, K$ and $\beta_k \in [0,1]^\infty$. Let $\mathcal{V}_k = \{\nu \in l^\infty : \nu(w) \in$

1. Stokey and Lucas [105] state the result for subsets of $\mathbb{R}^n$ for some integer n. The result and proof can be generalized to any subset of a Banach space, espectially $(l^\infty, ||\cdot||_\infty)$.

$[0, r_k],\ w \geq 1\}$. In addition, let $T_{k,\beta_k} : l^\infty \to l^\infty$ be the operator defined in equation (A.1). For any $J_k^0 \in \mathcal{V}_k$ and $w \geq 1$, the following holds:

$$\begin{aligned}(T_{k,\beta_k} \circ J_k^0)(w) &= \mathbb{E}_{\varepsilon_k}\left[\max\{\varepsilon_k(1), -c_k + \alpha[\beta_k(w) r_k + (1-\beta_k(w)) J_k^0(w+1) + \varepsilon_k(0)\}\right] \\ &\leq \mathbb{E}_{\varepsilon_k}\left[\max\{\varepsilon_k(1), -c_k + \alpha r_k + \varepsilon_k(0)\}\right] \leq r_k.\end{aligned}$$

The first inequality holds because $J_k^0 \in \mathcal{V}_k$. In particular, $J_k^0(w+1) \leq r_k$. The second inequality follows from Assumption 2. In addition, it is immediate from (A.1) that $(T_{k,\beta_k} \circ J_k^0)(w) \geq 0$ for $w \geq 1$. Therefore, $T_{k,\beta_k} \circ J_k^0 \in \mathcal{V}_k$. Since T_{k,β_k} is a contraction mapping and $\mathcal{V}_k$ is closed, $J_k = \lim_{n\to\infty} T_{k,\beta_k}^n J_k^0 \in \mathcal{V}_k$. In particular, $J_k(w) \leq r_k$ for $w \geq 1$.

□

A.1.2 Proofs of results in Section 3.1.1

Proof of Proposition 2.

It follows from equation (3.2) that $V(t)$ is a Markov process. Since we only consider the underloaded case, i.e. $a < b$, there is a unique stationary distribution of $V(t)$. Let $v(w)$ denote the stationary distribution of $V(t)$, i.e. $v(w) = \lim_{t\to\infty} \mathbb{P}(V(t) = w)$ for $w \geq 1$. The flow balance equations of the Markov process described in (3.2) is given as follows:

$$v(w) = \begin{cases} \sum_{i=0}^{w} v(i) a \bar{G}(i)(1-b)^{w-i} b + v(w+1)(1 - a\bar{G}(w+1)), & w \geq 1, \\ v(0)(1 - a\bar{G}(0)) + ab\bar{G}(0)) + v(1)(1 - a\bar{G}(1)), & w = 0, \end{cases} \tag{A.2}$$

where $\bar{G}(w) = 1 - G(w)$ for $w \geq 1$. We simplify these equations by defining $r(w) = v(w)(1-b)^{-w}$. By substituting the definition of $v(w)$ into (A.2) for $w = 0$ and rearranging the terms, we have that

$$(1 - a\bar{G}(1)) r(1) - a\bar{G}(0) r(0) = 0. \tag{A.3}$$

By substituting $r(w)$ into (A.2), we obtain the following: For $w \geq 1$,

$$\begin{aligned} r(w) =&(1-b)^{-w}\left[\sum_{i=0}^{w}(1-b)^{w}r(i)ab\bar{G}(i) + (1-b)^{w+1}r(w+1)(1-a\bar{G}(w+1))\right] \\ =&\sum_{i=0}^{w} r(i)ab\bar{G}(i) + (1-b)r(w+1)(1-a\bar{G}(w+1)) \end{aligned}$$

Rearranging the terms for $w = 1$, we obtain that

$$(1-b)[(1-a\bar{G}(2))r(2) - r(1)] = b[(1-a\bar{G}(1))r(1) - a\bar{G}(0)r(0)] = 0, \tag{A.4}$$

where the last equality follows from (A.3). Subtracting $r(w)$ from $r(w+1)$, we obtain that for $w \geq 1$,

$$\begin{aligned} r(w+1) - r(w) =&(1-b)(1-a\bar{G}(w+2))h(w+2) + r(w+1)ab\bar{G}(w+1) \\ &- (1-b)(1-a\bar{G}(w+1))r(w+1) \\ =&(1-b)(1-a\bar{G}(w+2))r(w+2) - (1-b-a\bar{G}(w+1))r(w+1). \end{aligned}$$

Rearranging the terms, we have that for $w \geq 1$,

$$(1-b)(1-a\bar{G}(w+2))r(w+2) - r(w+1) = [(1-a\bar{G}(w+1))r(w+1) - r(w)].$$

By substituting (A.4) into this equation recursively, we have that

$$(1-a\bar{G}(w+1))r(w+1) = r(w),\ w \geq 1.$$

This gives that

$$r(w) = r(1)\prod_{i=2}^{w}(1-a\bar{G}(i))^{-1}, w \geq 2.$$

By substituting $v(w) = (1-b)^w r(w)$ into this equation, we obtain the following:

$$v(w) = v(1) \prod_{i=2}^{w} \frac{(1-b)}{(1-a\bar{G}(i))}, w \geq 2.$$

Let $\beta(w)$ denote the probability of entering service in next period (in steady state) after waiting for w periods. Thus $\beta(w)$ is given as follows: For $w \geq 1$,

$$\begin{aligned}
\beta(w) &= \lim_{t\to\infty} \mathbb{P}(V(t) = w | V(t) \geq w) \\
&= \frac{v(w)}{\sum_{t=w}^{\infty} v(w)} \\
&= \frac{v(1) \prod_{i=2}^{w} (1-b)(1-a\bar{G}(i))^{-1}}{v(1) \sum_{t=w}^{\infty} \prod_{i=2}^{t} (1-b)(1-a\bar{G}(i))^{-1}} \\
&= \left(1 + \sum_{t=w+1}^{\infty} \prod_{i=w+1}^{t} (1-b)(1-a\bar{G}(i))^{-1} \right)^{-1}.
\end{aligned}$$

□

A.1.3 Proofs of results in Section 3.2

Proof of Lemma 1. For a given $\beta \in [0,1]^{\infty}$, let $\tilde{\beta} = \Phi(\Gamma(\beta))$. Note that the mapping $\Phi(\Gamma(\cdot))$ is characterized by equations (2.8), (2.10), (3.1) and (3.3).

Note from (3.1) that $\bar{G}(w) \geq 0$ for all $w \geq 1$. Thus, it follows from equation (3.3) that for $w \geq 1$,

$$\tilde{\beta}(w) \leq (1 + \sum_{i=1}^{\infty} (1-b)^i)^{-1} = b.$$

This gives the upper-bound of equation (3.5). In addition, it also follows from equation (3.3)

that for $w \geq 1$,

$$\begin{aligned}\tilde{\beta}(w) =& \left(1 + \sum_{t=w+1}^{\infty} \prod_{i=w+1}^{t} \frac{1-b}{1-a\bar{G}(i+1)}\right)^{-1} \\ \geq& \left(1 + \sum_{i=1}^{\infty} \left(\frac{1-b}{1-a\bar{G}(w+1)}\right)^{i}\right)^{-1} \\ =& 1 - \frac{1-b}{1-a\bar{G}(w+1)} \\ =& \frac{b - a\bar{G}(w+1)}{1-a\bar{G}(w+1)}.\end{aligned} \tag{A.5}$$

The inequality follows from the fact that $\bar{G}(w)$ is non-increasing, i.e. $\bar{G}(i) \leq \bar{G}(w+1)$ for all $i \geq w+1$; see (3.1) for its definition. It follows from Corollary 2 that $q_1(w) \geq \bar{q} > 0$ for all $w \geq 1$. Thus, it follows from equation (3.1) that

$$\bar{G}(w) = \prod_{i=1}^{w}(1 - q(i)) \leq (1-\underline{q})^{w}, \ w \geq 1.$$

Substituting this inequality into equation (A.5), we have that for $w \geq 1$,

$$\tilde{\beta}(w) \geq 1 - \frac{1-b}{1-a(1-\underline{q})^{w+1}} = \frac{b - a(1-\underline{q})^{w+1}}{1-a(1-\underline{q})^{w+1}}.$$

This shows the lower-bound of $\tilde{\beta}(w)$ provided in equation (3.5). We end the proof by showing that $\tilde{\beta}(w)$ is non-decreasing in w. Rearranging the terms in equation (3.4), we have the following: For $w \geq 1$,

$$\frac{1}{\tilde{\beta}(w+1)} = \left(\frac{1}{\tilde{\beta}(w)} - 1\right)\left(1 + \frac{b - a\bar{G}(w+1)}{1-b}\right)$$

Substituting this equation into the following, we obtain that for $w \geq 1$,

$$\begin{aligned}\frac{1}{\tilde{\beta}(w+1)} - \frac{1}{\tilde{\beta}(w)} &= \left(\frac{1}{\tilde{\beta}(w)} - 1\right)\left(1 + \frac{b - a\bar{G}(w+1)}{1-b}\right) - \frac{1}{\tilde{\beta}(w)} \\ &= \frac{1}{\tilde{\beta}(w)}\frac{b - a\bar{G}(w+1)}{1-b} - \frac{1 - a\bar{G}(w+1)}{1-b} \\ &\leq \frac{1 - a\bar{G}(w+1)}{b - a\bar{G}(w+1)}\frac{b - a\bar{G}(w+1)}{1-b} - \frac{1 - a\bar{G}(w+1)}{1-b} = 0.\end{aligned}$$

The last inequality follows from (A.5). Thus, we have that $\tilde{\beta}(w) \leq \tilde{\beta}(w+1)$ for all $w \geq 1$, i.e. $\tilde{\beta}(w)$ is non-decreasing in w.

□

Proof of Lemma 2. We first show that $\mathcal{B}$ is a compact set. Define a sequence x_w as follows:

$$x_w = b - \frac{b - a(1-\underline{q})^{w+1}}{1 - a(1-\underline{q})^{w+1}},\ w \geq 1.$$

Thus, it is equivalent to write $\mathcal{B}$ as

$$\mathcal{B} = \{\beta \in l^\infty : b - x_w \leq \beta(w) \leq b, w \geq 1\}.$$

Note that $x_w \to 0$ as $w \to \infty$. Thus, for any $\epsilon > 0$, there exits n such that $x_w < \epsilon$ for $w \geq n$. Define a set $\mathcal{B}_n$ as follows:

$$\mathcal{B}_n = \{\beta \in \mathbb{R}^n : b - x_w \leq \beta(w) \leq b, w = 1, \ldots, n\}.$$

Since $\mathcal{B}_n$ is a compact set in $\mathbb{R}^n$, it is totally bounded, i.e. it has a finite cover of open balls of radius ϵ. In other words, there exist l and $\nu_1, \ldots, \nu_l \in \mathbb{R}^n$ such that $\mathcal{B}_n \subseteq \cup_{i=1}^{l} B_n(\nu_i, \epsilon)$ where $B_n(\nu_i, \epsilon)$ is the open ball in $\mathbb{R}^n$ centered at ν_i and with radius ϵ. Let $w_i = (\nu_i, 0, \ldots)$, $i = 1, \ldots, l$. It is immediate that $\mathcal{B}$ is covered by $B(w_i, \epsilon)$, $i = 1, \ldots, l$, where $B(w_i, \epsilon)$ is the open ball in l^∞ that centers at w_i and has a radius ϵ. Since ϵ is arbitrary, $\mathcal{B}$ is

totally bounded. Since l^∞ is a complete metric space, the totally bounded subset $\mathcal{B}$ of l^∞ is compact; see Theorem 3.28 in Aliprantis and Border [5].

Next we show that $\Phi(\Gamma(\cdot))$ is continuous. Note that $\Phi(\Gamma(\cdot))$ is characterized by equations (2.8), (2.10), (3.1) and (3.3). Let $\beta_n, \beta \in \mathcal{B}$ be sequences such that $\beta_n \to \beta$ (under the sup norm). Let J_n, q_n, G_n and $\tilde{\beta}_n$ and J, q, G and $\tilde{\beta}$ be the left-hand sides of equations (2.8), (2.10), (3.1) and (3.3) by substituting β_n and β into $\Phi(\Gamma(\cdot))$, respectively. Thus, we have that $\tilde{\beta}_n = \Phi(\Gamma(\beta_n))$ and $\tilde{\beta} = \Phi(\Gamma(\beta))$. We need to show that $\tilde{\beta}_n \to \tilde{\beta}$ under the sup-norm.

It follows from $\beta_n \to \beta$ that for any $\epsilon > 0$, there exists n_1 such that $|\beta_n(w) - \beta(w)| < \epsilon$ for all $n \geq n_1$ and $w \geq 1$. It follows from (2.8) that for all $w \geq 1$ and $n \geq n_1$,

$$
\begin{aligned}
&|J_n(w) - J(w)| \\
=&|\mathbb{E}_\varepsilon \left[\max\{\varepsilon(1), -c + \alpha[\beta_n(w)r + (1-\beta_n(w))J_n(w+1)] + \varepsilon(0)\}\right] \\
&- \mathbb{E}_\varepsilon \left[\max\{\varepsilon(1), -c + \alpha[\beta(w)r + (1-\beta(w))J(w+1)] + \varepsilon(0)\}\right]| \\
=&|\mathbb{E}_\varepsilon \left[-c + \alpha[\beta_n(w)r + (1-\beta_n(w))J_n(w+1)] + \varepsilon(0) - \varepsilon(1)\right]^+ \\
&- \mathbb{E}_\varepsilon \left[-c + \alpha[\beta_n(w)r + (1-\beta(w))J(w+1)] + \varepsilon(0) - \varepsilon(1)\right]^+| \\
\leq&|\mathbb{E}_\varepsilon \left[\alpha[\beta_n(w)r + (1-\beta_n(w))J_n(w+1)] - \alpha[\beta(w)r + (1-\beta(w))J(w+1)]\right]| \\
=&\alpha|(r - J_n(w+1)(\beta_n(w) - \beta(w)) - (1-\beta(w))(J_n(w+1) - J(w+1))| \\
\leq&\alpha r|\beta_n(w) - \beta(w)| + \alpha|J_n(w+1) - J(w+1)| \\
\leq&\alpha r\epsilon + \alpha|J_n(w+1) - J(w+1)|.
\end{aligned}
$$

The equality in the third line follows from the fact that $\mathbb{E}[\varepsilon(1)] = 0$. The inequality in the fourth line follows from $|x_1^+ - x_2^+| \leq |x_1 - x_2|$ for all $x_1, x_2 \in \mathbb{R}$. It follows from Corollary 2 that $J_n(w), J(w) \in [0, r]$ for all $w \geq 1$. Thus $|J_n(w+1) - J(w+1)|$ is bounded for all $w \geq 1$. Thus, by applying this inequality recursively, we obtain that for all $w \geq 1$ and $n \geq n_1$,

$$
|J_n(w) - J(w)| \leq \alpha r\epsilon \sum_{i=0}^{\infty} \alpha^i = \frac{\alpha r\epsilon}{1-\alpha}. \tag{A.6}
$$

Let $C_0 = \sup_{x\in[-c,r]} f(x)$. It follows from Assumption 1 that $f(\cdot)$ is continuous. Thus, $C_0 < \infty$. Since $\bar{F}'(x) = -f(x)$, it holds that for any $x_1, x_2 \in [-c, r]$,

$$|\bar{F}(x_1) - \bar{F}(x_2)| \leq C_0|x_1 - x_2|. \tag{A.7}$$

Since $J_n(w) \leq r$ and $J(w) \leq r$ for all $w \geq 1$, we have that (for $w \geq 1$,)

$$\begin{aligned}
-c \leq -c + \alpha(\beta_n(w)r + (1-\beta_n(w))J_n(w+1)) \leq -c + \alpha r \leq r,\\
-c \leq -c + \alpha(\beta(w)r + (1-\beta(w))J(w+1)) \leq -c + \alpha r \leq r.
\end{aligned}$$

Thus, it follows from (2.10) and (A.6) that for all $n \geq n_1$ and $w \geq 1$,

$$\begin{aligned}
|q_n(w) - q(w)| \leq& C_0|\alpha[\beta_n(w)r + (1-\beta_n(w))J_n(w+1)] - \alpha[\beta(w) + (1-\beta(w))J(w+1)]|\\
=& C_0\alpha|(r - J_n(w+1))(\beta_n(w) - \beta(w)) - (1-\beta(w))(J_n(w+1) - J(w+1))|\\
\leq& C_0\alpha r|\beta_n(w) - \beta(w)| + c\alpha|J_n(w+1) - J(w+1)|\\
\leq& C_0\alpha r\epsilon + c\alpha\frac{\alpha r\epsilon}{1-\alpha}\\
=& \frac{C_0\alpha r\epsilon}{1-\alpha} = C_1\epsilon,
\end{aligned}$$

where $C_1 = C_0\alpha r/(1-\alpha)$. The first inequality follows from (A.7) and the second one follows from $J_n(w+1) \in [0, r]$ and $\beta(w) \in [0, 1]$. The last inequality follows from the assumption that $|\beta_n(w) - \beta(w)| < \epsilon$ for $n \geq n_1$ and (A.6). Substituting this inequality into (3.1), we have that for all $n \geq n_1$ and $w \geq 1$,

$$\begin{aligned}
|G_n(w+1) - G(w+1)| =& |\bar{G}_n(w+1) - \bar{G}(w+1)|\\
=& |\bar{G}_n(w)(1 - q_n(w+1)) - \bar{G}(w)(1 - q(w+1)|\\
\leq& \bar{G}_n(w)|q_n(w+1) - q(w+1)| + (1 - q(w+1))|\bar{G}_n(w) - \bar{G}(w)|\\
\leq& (1-\underline{q})^w C_1\epsilon + (1-\underline{q})|\bar{G}_n(w) - \bar{G}(w)|,
\end{aligned} \tag{A.8}$$

where the last inequality follows from Corollary 2 and equation (3.1) that

$$|1-q(w+1)| \le 1-\underline{q},\ 0 \le \bar{G}_n(w) \le (1-\underline{q})^w \ \text{ and } \ 0 \le \bar{G}(w) \le (1-\underline{q})^w,\ w \ge 1. \quad \text{(A.9)}$$

Applying (A.8) recursively, we obtain that: For all $w \ge 1$ and $n \ge n_1$,

$$\begin{aligned}
|G_n(w) - G(w)| \le & \sum_{i=1}^{w-1} (1-\underline{q})^i C_1\epsilon + (1-\underline{q})^{w-1} |\bar{G}_n(1) - \bar{G}(1)| \\
\le & \sum_{i=1}^{\infty} (1-\underline{q})^i C_1\epsilon + (1-\underline{q})^{w-1} |q_n(1) - q(1)| \\
\le & \frac{C_1\epsilon}{\underline{q}} + C_1\epsilon = \left(1 + \frac{1}{\underline{q}}\right) C_1\epsilon.
\end{aligned}$$

By letting $C_2 = (1 + 1/\underline{q})C_1$, we that $|G_n(w) - G(w)| \le C_2\epsilon$ for all $w \ge 1$ and $n \ge n_1$. It follows from (3.4) that for $w \ge 1$ and $n \ge n_1$,

$$\begin{aligned}
& |\tilde{\beta}_n(w) - \tilde{\beta}(w)| \\
=& \tilde{\beta}_n(w)\tilde{\beta}(w) \left| \frac{1}{\tilde{\beta}_n(w)} - \frac{1}{\tilde{\beta}(w)} \right| \\
\le& b^2 \left| \frac{1-b}{(1-a\bar{G}_n(w+1))\tilde{\beta}_n(w+1)} - \frac{1-b}{(1-a\bar{G}(w+1))\tilde{\beta}(w+1)} \right|
\end{aligned}$$

$$= b^2(1-b) \frac{|(1-a\bar{G}_n(w+1))(\tilde{\beta}_n(w+1) - \tilde{\beta}(w+1)) + a\tilde{\beta}(w+1)(\bar{G}_n(w+1) - \bar{G}(w+1))|}{(1-a\bar{G}_n(w+1))\tilde{\beta}_n(w+1)(1-a\bar{G}(w+1))\tilde{\beta}(w+1)}$$

$$\le b^2(1-b) \frac{(1-a\bar{G}_n(w+1))|\tilde{\beta}_n(w+1) - \tilde{\beta}(w+1)| + a\tilde{\beta}(w+1)|\bar{G}_n(w+1) - \bar{G}(w+1)|}{(1-a\bar{G}_n(w+1))\tilde{\beta}_n(w+1)(1-a\bar{G}(w+1))\tilde{\beta}(w+1)}$$

$$= \frac{b^2(1-b)|\tilde{\beta}_n(w+1) - \tilde{\beta}(w+1)|}{\tilde{\beta}_n(w+1)(1-a\bar{G}(w+1))\tilde{\beta}(w+1)} + \frac{ab^2(1-b)|\bar{G}_n(w+1) - \bar{G}(w+1)|}{(1-a\bar{G}_n(w+1))\tilde{\beta}_n(w+1)(1-a\bar{G}(w+1))}, \quad \text{(A.10)}$$

where the first inequality follows from Lemma 1 that $\tilde{\beta}_n(w) \leq b$, $\tilde{\beta}(w) \leq b$ and (3.4). Note that $(b - a(1-\underline{q})^w)/(1 - a(1-\underline{q})^w) \to b$ as $w \to \infty$. In addition, $(1-\underline{q})^w \to 0$ as $w \to \infty$. Thus, there exits w_1 such that for $w \geq w_1$,

$$\left(\frac{b - a(1-\underline{q})^w}{1 - a(1-\underline{q})^w}\right)^2 (1 - a(1 - (1-\underline{q})^w) \geq b^2\sqrt{1-b},$$
$$\frac{b - a(1-\underline{q})^w}{1 - a(1-\underline{q})^w}(1 - a(1 - (1-\underline{q})^w)^2 \geq b\sqrt{1-b}.$$

Substituting these two inequalities into (3.5) and (A.9), we have that for $w \geq w_1$ and $n \geq n_1$,

$$\tilde{\beta}_n(w+1)(1 - a\bar{G}(w+1))\tilde{\beta}_n(w+1) \geq b^2\sqrt{1-b},$$
$$(1 - a\bar{G}_n(w+1))\tilde{\beta}_n(w+1)(1 - a\bar{G}(w+1)) \geq b\sqrt{1-b}.$$

Substituting these two inequalities into (A.10) yields that for $w \geq w_1$ and $n \geq n_1$

$$\begin{aligned}|\tilde{\beta}_n(w) - \tilde{\beta}(w)| \leq& \sqrt{1-b}|\beta_n(w+1) - \tilde{\beta}(w+1)| + ab\sqrt{1-b}|\bar{G}_n(w+1) - \bar{G}(w+1)| \\ \leq& \sqrt{1-b}|\tilde{\beta}_n(w+1) - \tilde{\beta}(w+1)| + ab\sqrt{1-b}C_2\epsilon.\end{aligned}$$

Applying this inequality recursively, we have that for all $w \geq w_1$ and $n \geq n_1$,

$$|\tilde{\beta}_n(w) - \tilde{\beta}(w)| \leq ab\sqrt{1-b}C_2 e \sum_{i=1}^{\infty}\left(\sqrt{1-b}\right)^{i-1} = \frac{ab\sqrt{1-b}C_2\epsilon}{1-\sqrt{1-b}} = C_3\epsilon, \tag{A.11}$$

where $C_3 = ab\sqrt{1-b}C_2/(1-\sqrt{1-b})$. It follows from (A.10) that for $w < w_1$,

$$\begin{aligned}&|\tilde{\beta}_n(w) - \tilde{\beta}(w)| \\ \leq& \frac{b^2(1-b)|\beta_n(w+1) - \beta'(w+1)|}{\tilde{\beta}_n(w+1)(1 - a\bar{G}(w+1))\tilde{\beta}(w+1)} + \frac{ab^2(1-b)|\bar{G}_n(w+1) - \bar{G}(w+1)|}{(1 - a\bar{G}_n(w+1))\tilde{\beta}_n(w+1)(1 - a\bar{G}(w+1))} \\ \leq& \frac{b^2(1-b)|\tilde{\beta}_n(w+1) - \tilde{\beta}(w+1)|}{(b-a)^2(1-a)} + \frac{ab^2(1-b)C_2\epsilon}{(1-a)^2(b-a)}.\end{aligned} \tag{A.12}$$

The last inequality follows from $\bar{G}_n(w+1) \leq 1$ and $\bar{G}(w+1) \leq 1$ and from Lemma 1 that

$$\tilde{\beta}_n(w) \geq \frac{b - a(1-\underline{q})^w}{1 - a(1-\underline{q})^w} \geq b - a \quad \text{and} \quad \tilde{\beta}(w) \geq b - a.$$

By applying (A.12) recursively, we have that for all $w < w_1$ and $n \geq n_1$

$$\begin{aligned} |\tilde{\beta}_n(w) - \tilde{\beta}(w)| &\leq \left(\frac{b^2(1-b)}{(b-a)^2(1-a)}\right)^{w_1-w} |\tilde{\beta}_n(w_1) - \tilde{\beta}(w_1)| \\ &\quad + \frac{ab^2(1-b)C_2\epsilon}{(1-a)^2(b-a)} \sum_{i=0}^{w_1-w-1} \left(\frac{b^2(1-b)}{(b-a)^2(1-a)}\right)^i \\ &\leq \left(\frac{b^2(1-b)}{(b-a)^2(1-a)}\right)^{w_1-w} C_3\epsilon \\ &\quad + \frac{ab^2(1-b)C_2\epsilon}{(1-a)^2(b-a)} \sum_{i=0}^{w_1-w-1} \left(\frac{b^2(1-b)}{(b-a)^2(1-a)}\right)^i \leq c_1\epsilon, \end{aligned} \tag{A.13}$$

where

$$c_1 = \sup_{1 \leq w \leq w_1} \left(\frac{b^2(1-b)}{(b-a)^2(1-a)}\right)^{w_1-w} C_3 + \frac{ab^2(1-b)C_2}{(1-a)^2(b-a)} \sum_{i=0}^{w_1-w-1} \left(\frac{b^2(1-b)}{(b-a)^2(1-a)}\right)^i.$$

Note that w_1 is independent of ϵ. Thus, the constant c_1 is independent of ϵ as well. Combining equations (A.11) and (A.13), we have that

$$|\tilde{\beta}_n(w) - \tilde{\beta}(w)| \leq c_2\epsilon, \ w \geq 1 \ \text{ and } n \geq n_1,$$

where $c_2 = \max\{C_3, c_1\}$. By letting $\epsilon \to 0$, we have that $\tilde{\beta}(w) \to \tilde{\beta}$ uniformly. This gives the continuity of $\Phi(\Gamma(\cdot))$.

□

Proof of Corollary 4. Since β^* is the solution to the fixed point problem $\beta^* = \Phi(\Gamma(\beta^*))$, it is immediate from Lemma 1 that $\beta^*(w)$ is increasing in w and satisfies inequality (3.5) for all $w \geq 1$. Note that the left-hand side of equation (3.5) converges to b as w goes to infinity.

Therefore, $\lim_{w\to\infty}\beta^*(w) = b$.

□

Proof of Lemma 3. Let $e^* = (\beta^*, q^*)$ be an equilibrium. Let J^* be the expected utility associated with q^*. It follows from Proposition 1 that $J^* = T_{\beta^*} J^*$, where T_{β^*} is given by (A.1). Let $\mathcal{V} = \{J \in l^\infty : J(w_1) \le J(w_2) \le r, 1 \le w_1 \le w_2\}$. We first show that for any $J_0 \in \mathcal{V}$, $J = T_{\beta^*} J^0 \in \mathcal{V}$. It follows from (A.1) that for $w \ge 1$,

$$\begin{aligned} J(w) &= \mathbb{E}_\varepsilon \max\{\varepsilon(1), -c + \alpha[\beta^*(w)r + (1-\beta^*(w))J^0(w+1)] + \varepsilon(0)\} \\ &\le \mathbb{E}_\varepsilon \max\{\varepsilon(1), -c + \alpha[\beta^*(w+1)r + (1-\beta^*(w+1))J^0(w+1)] + \varepsilon(0)\} \\ &\le \mathbb{E}_\varepsilon \max\{\varepsilon(1), -c + \alpha[\beta^*(w+1)r + (1-\beta^*(w+1))J^0(w+2)] + \varepsilon(0)\} \\ &= J(w+1). \end{aligned} \tag{A.14}$$

The first inequality follows from Corollary 4 and the assumption that $J_0 \in \mathcal{V}$. In particular, $\beta^*(w) \le \beta^*(w+1)$ and $J_0(w+1) \le r$. The second inequality follows from the assumption that $J^0(w+1) \le J^0(w+2)$. In addition, the following holds: For $w \ge 1$,

$$\begin{aligned} J(w) &= \mathbb{E}_\varepsilon \max\{\varepsilon(1), -c + \alpha[\beta^*(w)r + (1-\beta^*(w))J^0(w+1)] + \varepsilon(0)\} \\ &\le \mathbb{E}_\varepsilon \max\{\varepsilon(1), -c + \alpha r + \varepsilon(0)\} \le r, \end{aligned} \tag{A.15}$$

where the first inequality follows from $J_0(w+1) \le r$ and the second inequality follows from Assumption 2. Therefore, it follows from (A.14)-(A.15) that $J \in \mathcal{V}$. We have shown in the proof of Proposition 1 that T is a contraction mapping. Since $\mathcal{V}$ is a closed set, we have that $J^* = \lim_{n\to\infty} T^n_{\beta^*} J^0 \in \mathcal{V}$. In particular, $J^*(w)$ is increasing in w and bounded above by r.

In addition, it follows from (2.10) that for $w \ge 1$,

$$\begin{aligned} q^*(w) =& \bar{F}(-c + \alpha[\beta^*(w)r + (1-\beta^*(w))J^*(w+1)]) \\ \ge& \bar{F}(-c + \alpha[\beta^*(w+1)r + (1-\beta^*(w+1))J^*(w+2)]) \\ =& q^*(w+1). \end{aligned}$$

The inequality follows from Corollary 4 that $\beta^*(w) \leq \beta^*(w+1)$ and the monotonicity of J^*, i.e. $J^*(w+1) \leq J^*(w+2) \leq r$. Thus, $q^*(w)$ is decreasing in w.

□

Proof of Corollary 5. It follows from Lemma 3 and Corollary 2 that $J^*(w)$ is increasing in w and bounded above by r. Thus, there exists $J'_\infty \leq r$ such that $\lim_{w\to\infty} J^*(w) = J'_\infty$.

Note that the right-hand side of (3.7) equals to $\kappa(b, x)$ where $\kappa(\cdot)$ is defined in (A.55). It follows from Lemma 35 that the fixed point of (3.7) is unique. Let $J_\infty = j(b)$ be the fixed point of (3.7) where $j(\cdot)$ is given in Lemma 35.

Next, we show that $J'_\infty = J_\infty$. We first show that $J'_\infty \leq J_\infty$. Let $\beta_1(w) = b$ and $J_1(w) = J_\infty$ for all $w \geq 1$. It is immediate that J_1 is a fixed point of $J = T_{\beta_1} J$ where the operator T_{β_1} is given by (A.1). As we have shown in the proof of Proposition 1 that T_{β_1} is a contraction mapping, this fixed point is unique. Substituting the inequalities $\beta^*(w) \leq \beta_1(w) = b$ and $J^*(w) \leq r$ into (A.1), we have that

$$J^*(w) = T_{\beta^*} J^*(w) \leq T_{\beta_1} J^*(w), \ w \geq 1,$$

Substituting this inequality recursively into (A.1), we have that $J^*(w) \leq T^n_{\beta_1} J^*(w)$ for all n, w. Thus, the following holds:

$$J^*(w) \leq \lim_{n\to\infty} T^n_{\beta_1} J^*(w) = J_1(w) = J_\infty, \ w \geq 1. \tag{A.16}$$

Letting w go to infinity, we have that $J'_\infty = \lim_{w\to\infty} J^*(w) \leq J_\infty$.

Next we show that $J'_\infty \geq J_\infty$. It follows from Corollary 4 that $\beta^*(w) \to b$. Thus, fixing $\epsilon > 0$, there exists w_1 such that $\beta^*(w) \geq b - \epsilon$ for all $w \geq w_1$. Let $\beta_2(w) = \beta^*(w + w_1)$ and $\beta_3(w) = b - \epsilon$ for $w \geq 1$. In addition, let $J_2(w) = J^*(w + w_1)$ and $J_3(w) = j(b - \epsilon)$, where $j(\cdot)$ is given in Lemma 35. It is immediate that J_i is the unique fixed point of $J = T_{\beta_i} J$, $i = 2, 3$. Since $\beta_2(w) \geq \beta_3(w)$ for all $w \geq 1$, we can repeat the proof of (C.22) and show

that $J_2(w) \geq J_3(w) = j(b-\epsilon)$ for all $w \geq 1$. Letting w go to infinity, we obtain that

$$J'_\infty = \lim_{w\to\infty} J^*(w) = \lim_{w\to\infty} J_2(w) \geq j(b-\epsilon).$$

By letting $\epsilon \to 0$, it follows from the contintuity of $j(\cdot)$ (cf. Lemma 36) that $J'_\infty \geq j(b) = J_\infty$. Thus, we conclude that $\lim_{w\to\infty} J^*(w) = J'_\infty = J_\infty$.

It follows from Lemma 3 and Corollary 2 that $q^*(w)$ is decreasing in w and bounded above from $\underline{q}$. Thus, there exists a constant q_∞ such that $\lim_{w\to\infty} q^*(w) = q_\infty$. In addition, it follows from (2.10) that

$$\begin{aligned} q_\infty = \lim_{w\to\infty} q^*(w) = \lim_{w\to\infty} \bar{F}(-c + \alpha[\beta^*(w)r + (1-\beta^*(w))J^*(w+1)]) \\ = \bar{F}(-c + \alpha(br + (1-b)J_\infty)). \end{aligned}$$

The last inequality follows from the continuity of $\bar{F}(\cdot)$ and that $\beta^*(w) \to b$ and $J^*(w) \to J_\infty$ as $w \to \infty$.

□

A.1.4 Proofs of the proposition and the lemma in Section 3.3

Proof of Lemma 8. Fixing N and comparing (3.30)-(3.33) and (A.30)-(A.32), we have that

$$(e_N(w), \bar{G}_N(w)) = h(e_N(w+1), \bar{G}_N(w+1)),\ w < N. \tag{A.17}$$

Note that the truncation in (A.32) is immaterial in this case because $\bar{G}_N(w) \leq 1$ for $w \geq 1$. Fixing $w = N$ and substituting $z(N) = (\beta_N(N), q_N(N), \bar{G}_N(N))$ into equation (A.43), we have that the resulting $z(1) = (\beta_N(1), q_N(1), \bar{G}_N(1))$ satisfies (A.44). In particular, $\bar{G}_N(1) = 1 - q_N(1)$. Thus, it follows from the definition of function $f_N(\cdot)$ that

$$\bar{G}_N(w) = f_w(e_N(w)), w \geq N. \tag{A.18}$$

In particular, $\bar{G}_N(N) = f_N(e_N(N))$. In other words, the value of $\bar{G}_N(N)$ is uniquely determined. Since the truncated equilibrium is fully characterized by $\bar{G}_N(N)$, we conclude that the truncated equilibrium is unique.

□

Proof of Proposition 5. To facilitate the analysis to follows, we define a function $\tilde{h} = (\tilde{h}_1, \tilde{h}_2)$ as follows: For $w \geq 1$ and $(\beta, q) \in \mathcal{Z}_1(w) \times \mathcal{Z}_2(w) \subseteq (0, b] \times [q_\infty, 1)$,

$$\tilde{h}_i(\beta, q; w) = h(\beta, q, f_w(\beta, q)), \ i = 1, 2,$$

where the functions $h(\cdot)$ and $f_w(\cdot)$ are defined in (A.30)-(A.32) and (A.42)-(A.44) and $\mathcal{Z}_1(w) \times \mathcal{Z}_2(w)$ is given in (A.61). Define a matrix $D\tilde{h}(\beta_1, q_1, \beta_2, q_2; w)$ as follows: For $w \geq 1$ and $(\beta_1, q_1), (\beta_2, q_2) \in \mathcal{Z}_1(w) \times \mathcal{Z}_2(w)$,

$$D\tilde{h}(\beta_1, q_1, \beta_2, q_2; w) = \begin{bmatrix} Dh_{11}(\beta_1, q_1; w) & Dh_{12}(\beta_1, q_1; w) \\ Dh_{21}(\beta_2, q_2; w) & Dh_{22}(\beta_2, q_2; w) \end{bmatrix},$$

where

$$\begin{aligned}
D\tilde{h}_{11}(\beta, q; w) &= \frac{\partial h_1(\beta, q, f_w(\beta, q))}{\partial z_1} + \frac{\partial h_1(\beta, q, f_w(\beta, q))}{\partial z_3} \frac{\partial f_w(\beta, q)}{\partial \beta}, \\
D\tilde{h}_{12}(\beta, q; w) &= \frac{\partial h_1(\beta, q, f_w(\beta, q))}{\partial z_2} + \frac{\partial h_1(\beta, q, f_w(\beta, q))}{\partial z_3} \frac{\partial f_w(\beta, q)}{\partial q}, \\
D\tilde{h}_{21}(\beta, q; w) &= \frac{\partial h_2(\beta, q, f_w(\beta, q))}{\partial z_1} + \frac{\partial h_2(\beta, q, f_w(\beta, q))}{\partial z_3} \frac{\partial f_w(\beta, q)}{\partial \beta}, \\
D\tilde{h}_{22}(\beta, q; w) &= \frac{\partial h_2(\beta, q, f_w(\beta, q))}{\partial z_2} + \frac{\partial h_2(\beta, q, f_w(\beta, q))}{\partial z_3} \frac{\partial f_w(\beta, q)}{\partial q}.
\end{aligned}$$

It is immediate that $D\tilde{h}(\beta, q, \beta, q; w)$ is the Jacobian matrix of $\tilde{h}(\beta, q)$. In addition, define a constant matrix Dh_0 as follows:

$$D\tilde{h}_0 = \begin{bmatrix} \dfrac{\partial h_1(z_0)}{\partial z_1} & \dfrac{\partial h_1(z_0)}{\partial z_2} \\ \dfrac{\partial h_2(z_0)}{\partial z_1} & \dfrac{\partial h_2(z_0)}{\partial z_2} \end{bmatrix} = \begin{bmatrix} 1-b & 0 \\ -f(\bar{F}^{-1}(q_\infty))\alpha(r - J_\infty)(1-b) & \alpha(1-q_\infty)(1-b) \end{bmatrix},$$

where $z_0 = (\beta, q_\infty, 0)$. It is immediate that the eigenvalues of $D\tilde{h}_0$ are $1-b$ and $\alpha(1-q_\infty)(1-b)$. Thus, there exists a invertible matrix S such that the following holds:

$$\begin{bmatrix} 1-b & 0 \\ 0 & \alpha(1-q_\infty)(1-b) \end{bmatrix} = S(D\tilde{h}_0)S^{-1}.$$

Define a vector norm$||\cdot||_S$ and a matrix norm $|||\cdot|||_S$ as follows: For $x \in \mathbb{R}^2$ and $M \in \mathcal{M}_2$,

$$||x||_S = ||Sx||_\infty \quad \text{and} \quad |||M|||_S = |||SMS^{-1}|||_\infty.$$

It is immediate that $|||D\tilde{h}_0|||_S = 1-b$. Define a sequnce a_w as follows:

$$a_w = \sup\left\{|||D\tilde{h}(\beta_1, q_1, \beta_2, q_2; w)|||_S : (\beta_1, q_1), (\beta_2, q_2) \in \mathcal{Z}_1(w) \times \mathcal{Z}_2(w)\right\}, \ w \geq 1. \quad \text{(A.19)}$$

We then show that $a_w \to |||D\tilde{h}_0|||_S = 1-b$ as $w \to \infty$. It follows from Lemma 32 that

$$\sup\{|f_w(\beta, q)| : (\beta, q) \in \mathcal{Z}_1(w) \times \mathcal{Z}_2(w) \subseteq (0, b] \times [q_\infty, 1)\} \to 0 \quad \text{as} \quad w \to \infty.$$

It follows from Lemma 37 and equation (A.61) that

$$\sup\{|\beta - b| + |q - q_\infty| : (\beta, q) \in \mathcal{Z}_1(w) \times \mathcal{Z}_2(w)\} \to 0 \quad \text{as} \quad w \to \infty.$$

Thus, it follows from the continuity of the partial derivatives of $h(\cdot)$ (see (A.33)-(A.41)) that for $i = 1, 2$,

$$\sup_{(\beta,q)\in\mathcal{Z}_1(w)\times\mathcal{Z}_2(w)} \left| \frac{\partial h_i(\beta, q, f_w(\beta, q))}{\partial z_j} - \frac{\partial h_i(b, q_\infty, 0)}{\partial z_j} \right| \to 0 \quad \text{as} \quad w \to \infty. \tag{A.20}$$

In addition, it follows from Lemma 40 that as $w \to \infty$,

$$\sup_{(\beta,q)\in\mathcal{Z}_1(w)\times\mathcal{Z}_2(w)} \left| \frac{\partial f_w(\beta, q)}{\partial \beta} \right| \to 0 \quad \text{and} \quad \sup_{(\beta,q)\in\mathcal{Z}_1(w)\times\mathcal{Z}_2(w)} \left| \frac{\partial f_w(\beta, q)}{\partial q} \right| \to 0. \tag{A.21}$$

Substituting (A.20)-(A.21) into $D\tilde{h}(\cdot)$, we have that

$$\sup_{(\beta_1,q_1),(\beta_2,q_2)\in\mathcal{Z}_1(w)\times\mathcal{Z}_2(w)} \left| D\tilde{h}(\beta_1, q_1, \beta_2, q_2; w) - D\tilde{h}_0 \right| \to 0 \quad \text{as} \quad w \to \infty.$$

By the continuity of the norm $|||\cdot|||_S$, we have that

$$\begin{aligned} \lim_{w\to\infty} a_w = &\lim_{w\to\infty} \sup \{|||Dh(\beta_1, q_1, \beta_2, q_2; w)|||_S : (\beta_1, q_1), (\beta_2, q_2) \in \mathcal{Z}_1(w) \times \mathcal{Z}_2(w)\} \\ =&|||D\tilde{h}_0|||_S = 1 - b. \end{aligned}$$

Thus, there exists $w_1 \geq 1$ such that

$$a_w \leq 1 - b/2 < 1, \ w \geq w_1. \tag{A.22}$$

Next we show that $e^N \to e^*$ uniformly. Define the difference of the truncated equilibrium and the equilibrium as follows:

$$\delta_\beta^N(w) = \beta_N(w) - \beta^*(w) \quad \text{and} \quad \delta_q^N(w) = q_N(w) - q^*(w), \ N, w \geq 1.$$

To show that $e^N \to e^*$ uniformly, we need to show that

$$\sup_{w\geq 1} \delta_\beta^N(w) \to 0 \ \text{ and } \ \sup_{w\geq 1} \delta_q^N(w) \to 0 \ \text{ as } N \to \infty.$$

It is equivalent to show that

$$\sup_{w\geq 1} ||\delta^N(w)||_S \to 0 \ \text{ as } N \to \infty, \tag{A.23}$$

where $\delta^N(w) = [\delta_\beta^N(w), \delta_q^N(w)]^T$ for all $N, w \geq 1$. The rest of this proof shows that (A.23) holds. It follows from Corollaries 4-5 that $\beta^*(w) \to b$ and $q^*(w) \to q_\infty$ as $w \to \infty$. Note that $\beta_N(w) = b$ and $q_N(w) = q_\infty$ for $w \geq N$. Thus, for any $\epsilon > 0$, there exists $N_1 \geq w_1$ such that

$$||\delta^N(w)||_S < \epsilon, \ w \geq N \geq N_1. \tag{A.24}$$

It follows from (A.17)-(A.18), Lemmas 41 and Corollary 13 that for $N > w \geq 1$,

$$(\beta_N(w), q_N(w)) = \tilde{h}(\beta_N(w+1), q_N(w+1)) \ \text{ and } \ (\beta^*(w), q^*(w)) = \tilde{h}(\beta^*(w+1), q^*(w+1)).$$

Thus, it follows from the mean vale theorem that for $N > w \geq 1$,

$$\begin{aligned} \delta^N(w) =& \tilde{h}(\beta_N(w+1), q_N(w+1)) - \tilde{h}(\beta_N(w+1), q_N(w+1)) \\ =& D\tilde{h}(w+1; \beta_1^N(w+1), q_1^N(w+1), \beta_2^N(w+1), q_2^N(w+1))\delta^N(w+1), \end{aligned} \tag{A.25}$$

where

$$\begin{aligned} &(\beta_i^N(w+1), q_i^N(w+1)) \\ =& c_i^N(w+1)(\beta_N(w+1), q_N(w+1)) + (1 - c_i^N(w+1))(\beta^*(w+1), q^*(w+1)) \end{aligned}$$

for some $c_i^N(w+1) \in (0,1)$, $i = 1, 2$. Note that

$$(\beta_N(N), q_N(N)) = (b, q_\infty) \in \mathcal{Z}_1(N) \times \mathcal{Z}_2(N).$$

In addition, it follows from Lemma 42 that $(\beta^*(N), q^*(N)) \in \mathcal{Z}_1(N) \times \mathcal{Z}_2(N)$. It follows from Lemma 38 and the convexity of $\mathcal{Z}_1(w) \times \mathcal{Z}_2(w)$ that

$$(\beta_i^N(w+1), q_i^N(w+1)) \in \mathcal{Z}_1(w+1) \times \mathcal{Z}_2(w+1), \ w = 1, \ldots, N-1, \ \text{ and } \ i = 1, 2.$$

Thus, it follows from (A.19) that for $w = 1, \ldots, N-1$,

$$|||D\tilde{h}(\beta_1^N(w+1), q_1^N(w+1), \beta_2^N(w+1), q_2^N(w+1); w+1)|||_S \leq a_{w+1}.$$

By taking the norm of the both sides of (A.25), we obtain that for $w = 1, \ldots, N-1$.

$$\begin{aligned} ||\delta^N(w)||_S \leq & |||D\tilde{h}(w+1; \beta_1^N(w+1), q_1^N(w+1), \beta_2^N(w+1), q_2^N(w+1))|||_S ||\delta^N(w+1)||_S \\ \leq & a_{w+1} ||\delta^N(w+1)||_S. \end{aligned} \tag{A.26}$$

Substituting (A.22) and (A.24) into (A.26) yields that for $N \geq N_1$ and $w = w_1, \ldots, N$,

$$||\delta^N(w)||_S \leq \left(1 - \frac{1}{2}b\right) ||\delta^N(w+1)||_S \leq \cdots \leq \left(1 - \frac{1}{2}b\right)^{N-w} ||\delta^N(N)||_S < \epsilon. \tag{A.27}$$

In addition, it follows from (A.26) that for $N \geq N_1$ and $w = 1, \ldots, w_1 - 1$,

$$||\delta^N(w)||_S \leq \left(\prod_{i=w+1}^{w_1} a_i\right) ||\delta^N(w_1)||_S < \bar{a}\epsilon, w < w_1, \tag{A.28}$$

where

$$\bar{a} = \max\left\{ \sup_{i\in\{1,\ldots,w_1-1\}} \prod_{j=i+1}^{w_1} a_j, 1 \right\}.$$

Thus, we conclude from (A.24), (A.27)-(A.28) that $||\delta^N(w)||_S < \bar{a}\epsilon$ for all $N \geq N_1$ and $w \geq 1$. By letting $\epsilon \to 0$, equation (A.23) holds.

□

Proof of Lemma 17. We show by induction that for $w = 1, \ldots, N$,

$$\beta_N^1(w) \leq \beta_N^2(w), \quad q_N^1(w) \geq q_N^2(w) \text{ and } \bar{G}_N^1(w) > \bar{G}_N^2(w). \tag{A.29}$$

This is true for $w = N$ by assumption. As the inductive assumption, suppose (A.29) is true for w, then we argue that it is also true for $w-1$. It follows from equation (3.30) and the inductive assumption that $\beta_N^1(w-1) \leq \beta_N^2(w-1)$. Similarly, it follows from (3.31)-(3.33) that

$$q_N^1(w-1) \geq q_N^2(w-1) \text{ and } \bar{G}_N^1(w-1) > \bar{G}_N^2(w-1).$$

In particular, both of the following must be true:

$$q_N^1(1) \geq q_N^2(1) \text{ and } \bar{G}_N^1(1) > \bar{G}_N^2(1).$$

Thus, the following holds:

$$\bar{G}_N^1(0) = \frac{\bar{G}_N^1(1)}{1 - q_N^1(1)} > \bar{G}_N^2(0) = \frac{\bar{G}_N^2(1)}{1 - q_N^2(1)}.$$

□

A.2 Technical Lemmas Characterizing the Equilibrium Quantities in Discrete Time

This section proves Lemma 4 that facilitates the proof of uniqueness of the equilibrium. To prove this result, we define an auxiliary function $f_w(\cdot)$ implicitly and study its properties(especially the monotonicity and convergence of its partial derivatives as w gets large). The function helps characterize $\bar{G}$ in terms of β and q. We then apply the mean value theorem to $f_w(\cdot)$ to establish the result in Lemma 4.

A.2.1 Definition of the auxiliary function $f_w(\cdot)$

The function $f_w(\cdot)$ is constructed such that for an equilibrium, the following holds:

$$\bar{G}^*(w) = f_w(\beta^*(w), q^*(w)),\ w \geq 1.$$

We establish this relationship in Section A.2.4. As a preliminary, we first define a function $h(\cdot)$. The function $f_w(\cdot)$ is then defined implicitly as the value satisfying a set of equations characterized by $h(\cdot)$ recursively.

To facilitate the analysis to follow, we define a function $h = (h_1, h_2, h_3) : Z \to \mathbb{R}^3$ as follows[2]:

$$h_1(z) = \left(1 + \frac{1-b}{1-az_3}\frac{1}{z_1}\right)^{-1}, \tag{A.30}$$

$$h_2(z) = \bar{F}\left(-c + \alpha\left[h_1(z)r + (1-h_1(z))\int_{-\infty}^{\bar{F}^{-1}(z_2)} F(x)\,\mathrm{d}x\right]\right), \tag{A.31}$$

$$h_3(z) = \min\left(\frac{z_3}{1-z_2}, 1\right), \tag{A.32}$$

where $z = (z_1, z_2, z_3)$ and $Z = (0, b] \times [q_\infty, 1) \times [0, 1] \subseteq \mathbb{R}^3$, where q_∞ is the constant defined

2. We truncate the value of $h_3(z)$ by one to ensure that $h_3(z) \in [0, 1]$ for all $z \in Z$. In the following analysis, we are only interested in $z \in Z$ that satisfies certain conditions. For those z of interest, the truncation is immaterial; see Lemma 31.

in Corollary 5. The following lemma shows that $h(\cdot)$ maps Z to Z.

Lemma 27. *We have that $h(z) \in Z$ for all $z \in \mathcal{Z}$. In addition, the following inequality holds for all $z \in Z$:*

$$\int_{-\infty}^{\bar{F}^{-1}(z_2)} F(x)\,\mathrm{d}x = \mathbb{E}[\bar{F}^{-1}(z_2) - (\varepsilon(1) - \varepsilon(0))]^+ \leq J_\infty \leq r,$$

where J_∞ is the constant defined in Corollary 5.

Proof. For any $z \in Z$, it is straightforward that $h_1(z) > 0$. Since $h_1(\cdot)$ is increasing in z_1 and decreasing in z_3, $h_1(z) \leq (1 + (1-b)/b)^{-1} = b$. Thus $h_1(z) \in (0, b]$.

Recall that the cdf $F(\cdot)$ is the distribution function of the difference of the idiosyncratic shocks $\varepsilon(1) - \varepsilon(0)$. It follows from integration by parts that

$$\begin{aligned}
\mathbb{E}[\bar{F}^{-1}(z_2) - (\varepsilon(1) - \varepsilon(0))]^+ &= \int_{-\infty}^{\bar{F}^{-1}(z_2)} \left(\bar{F}^{-1}(z_2) - x\right) \mathrm{d}F(x) \\
&= (\bar{F}^{-1}(z_2) - x)F(x)|_{-\infty}^{\bar{F}^{-1}(z_2)} - \int_{-\infty}^{\bar{F}^{-1}(z_2)} F(x)\,\mathrm{d}\left(\bar{F}^{-1}(z_2) - x\right) \\
&= \int_{-\infty}^{\bar{F}^{-1}(z_2)} F(x)\,\mathrm{d}x.
\end{aligned}$$

By substituting the definitions of J_∞ and q_∞ (in Corollary 5) into the right-hand side, the following inequality holds: For any $z_2 \in [q_\infty, 1)$,

$$\begin{aligned}
\int_{-\infty}^{\bar{F}^{-1}(z_2)} F(x)\,\mathrm{d}x &\leq \int_{-\infty}^{\bar{F}^{-1}(q_\infty)} F(x)\,\mathrm{d}x \\
&= \mathbb{E}[\bar{F}^{-1}(q_\infty) - (\varepsilon(1) - \varepsilon(0))]^+ \\
&= \mathbb{E}[-c + \alpha[br + (1-b)J_\infty] - (\varepsilon(1) - \varepsilon(0))]^+ \\
&= J_\infty \leq r.
\end{aligned}$$

The second inequality holds because by definition of q_∞, i.e. $q_\infty = \bar{F}(-c + \alpha[br + (1-b)J_\infty)$.

The third equality follows from equation (3.7). This proves that for any $z \in Z$,

$$\int_{-\infty}^{\bar{F}^{-1}(z_2)} F(x) \leq J_\infty.$$

Substituting this inequality into (A.31), we obtain that for $z \in \mathcal{Z}$,

$$\begin{aligned} h_2(z) &= \bar{F}\left(-c + \alpha\left[h_1(z)r + (1 - h_1(z))\int_{-\infty}^{\bar{F}^{-1}(z_2)} F(x)\,\mathrm{d}x\right]\right) \\ &\geq \bar{F}\left(-c + \alpha(h_1(z)r + (1 - h_1(z))J_\infty\right) \\ &\geq \bar{F}\left(-c + \alpha(br + (1 - b)J_\infty\right) \\ &= q_\infty, \end{aligned}$$

where the last inequality follows from $h_1(z) \in (0, b]$. In addition, $h_2(z) \leq \bar{F}(-c) < 1$. Thus, $h_2(z) \in [q_\infty, 1)$ for any $z \in Z$. Since $z_3 \geq 0$ and $z_2 < 1$, it follows from (A.32) that $h_3(z) \in [0, 1]$. Thus, $h(z) \in Z$. □

The following lemma shows the elements of the Jacobian matrix of $h(\cdot)$ and the sign of each element.

Lemma 28. *The partial derivatives of $h(\cdot)$ are given as follows:*

$$\frac{\partial h_1}{\partial z_1} = \frac{h_1^2(z)}{z_1^2}\frac{1-b}{1-az_3} > 0, \tag{A.33}$$

$$\frac{\partial h_1}{\partial z_2} = 0, \tag{A.34}$$

$$\frac{\partial h_1}{\partial z_3} = -\frac{ah_1^2(z)}{(1-az_3)^2}\frac{1-b}{z_1} < 0, \tag{A.35}$$

$$\frac{\partial h_2}{\partial z_1} = -f(\bar{F}^{-1}(h_2(z)))\alpha\left(r - \int_{-\infty}^{\bar{F}^{-1}(z_2)} F(x)\,\mathrm{d}x\right)\frac{h_1^2(z)}{z_1^2}\frac{1-b}{1-az_3} < 0, \tag{A.36}$$

$$\frac{\partial h_2}{\partial z_2} = \alpha(1-h_1(z))(1-z_2)\frac{f(\bar{F}^{-1}(h_2(z)))}{f(\bar{F}^{-1}(z_2))} > 0, \tag{A.37}$$

$$\frac{\partial h_2}{\partial z_3} = f(\bar{F}^{-1}(h_2(z))\alpha\left(r - \int_{-\infty}^{\bar{F}^{-1}(z_2)} F(x)\,\mathrm{d}x\right)\frac{ah_1^2(z)}{(1-az_3)^2}\frac{1-b}{z_1} > 0, \tag{A.38}$$

$$\frac{\partial h_3}{\partial z_1} = 0, \tag{A.39}$$

$$\frac{\partial h_3}{\partial z_2} = \begin{cases} \dfrac{z_3}{(1-z_2)^2} \geq 0, & \text{if } \ z_3 < 1-z_2, \\ 0 & \text{if } \ z_3 > 1-z_2, \end{cases} \tag{A.40}$$

$$\frac{\partial h_3}{\partial z_3} = \begin{cases} \dfrac{1}{1-z_2} > 1, & \text{if } \ z_3 < 1-z_2, \\ 0 & \text{if } \ z_3 > 1-z_2. \end{cases} \tag{A.41}$$

Proof. Since $h(\cdot)$ is given explicitly by equations (A.30)-(A.32), the partial derivatives of $h(\cdot)$ is immediate. The signs of equations (A.36) and (A.38) follow from Lemma 27. To be specific,

$$r - \int_{-\infty}^{\bar{F}^{-1}(z_2)} F(x)\,\mathrm{d}x \geq 0, \ z_2 \in [q_\infty, 1).$$

The signs of other equations are immediate. □

In addition, the following lemma will be useful in the analysis to follows.

Lemma 29. *We have that $h_3(z) \geq z_3$ for all $z \in Z$.*

Proof. For all $z \in Z$, the following inequality holds:

$$h_3(z) = \min\left(\frac{z_3}{1-z_2}, 1\right) \geq \min(z_3, 1) = z_3.$$

□

For every $w \geq 1$, we define an implicit function $f_w(\beta, q) : (0, b] \times [q_\infty, 1) \to [0, 1]$ through the set of equations immediately below. Namely, for $(\beta, q) \in (0, b] \times [q_\infty, 1)$, $f_w(\beta, q)$ and $z(k)$ for $k = w, \ldots, 1$ are defined implicitly by equations (A.42)-(A.44).

$$z(w) = (\beta, q, f_w(\beta, q)), \tag{A.42}$$

$$z(k-1) = h(z(k)) \quad \text{for} \quad k = w, \ldots, 2, \tag{A.43}$$

$$z_2(1) = 1 - z_3(1), \tag{A.44}$$

Intuitively, given (β, q) and an initial guess of $f_w(\beta, q)$, $z(w)$ is defined by (A.42) and $z(k)$ (for $k = 1, \ldots, w-1$) are well-defined by (A.43) and Lemma 27. The essence of what the next lemma shows is that there is a unique value of $f_w(\cdot)$ such that the boundary condition (A.44) is satisfied.

Lemma 30. *The function $f_w(\cdot)$ is well defined for all $w \geq 1$. In addition, $f_w(\beta, q) \in (0, 1)$ for all $(\beta, q) \in (0, b] \times [q_\infty, 1)$.*

Proof. Note that given $z(w) = (\beta, q, f_w(\beta, q))$, equation (A.43) defines $z(k)$ for $k = 1, \ldots, w-1$. However, the resulting z values must also satisfy the boundary condition (A.44). In other words, we need to show that for any $(\beta, q) \in (0, b] \times [q_\infty, 1)$, there exists a unique value $f_w(\beta, q)$ such that the resulting $z(k)$, $k = 1, \ldots, w$, satisfy (A.42)-(A.44).

We first show that there exists a value of $f_w(\beta, q) = \eta$ such that (A.42)-(A.44) are satisfied. To this end, we view $z(k)$ for $k = 1, \ldots, w-1$ as functions of η, denoted by $z(k; \eta)$

for $\eta \in [0,1]$. In addition, we define a function $\phi(\eta)$ as follows:

$$\phi(\eta) = z_2(1;\eta) - (1 - z_3(1;\eta)).$$

Note that $z_3(1;0) = 0$ which follows from equation (A.32) inductively. It follows from Lemma 27 by induction that $z_2(1;\eta) \in [q_\infty, 1)$ for all η. In particular, $z_2(1;0) < 1$, so $\phi(0) = z_2(1;0) - 1 < 0$. Next, we argue that $\phi(1) > 0$. To see this, note that for any $\eta \in [0,1]$,

$$1 \geq z_3(1;\eta) \geq z_3(2;\eta) \geq \cdots \geq z_3(w;\eta) = \eta \geq 0, \tag{A.45}$$

which follows from Lemma 29 inductively. Note from equation (A.45) that $z_3(1;1) = 1$. Thus, $\phi(1) = z_2(1;1) \geq q_\infty > 0$.

Since $h(\cdot)$ is continuous, $\phi(\cdot)$ is continuous as well. Combining the facts that $\phi(0) < 0$ and $\phi(1) > 0$, we conclude that there exists $\eta \in (0,1)$ such that $\phi(\eta) = 0$, i.e. condition (A.44) is satisfied. For such η, the resulting vectors $z(k;\eta)$, $k = 1, \ldots, w$ satisfy conditions (A.42)-(A.44).

For η such that $\phi(\eta) = 0$, $z_3(1;\eta) = 1 - z_2(1;\eta) < 1$. It follows from equation (A.45) that $z_3(k;\eta) < 1$. Substituting the definition of $h_3(\cdot)$ into the inequality $z_3(k;\eta) < 1$, we argue that for such η, the following holds:

$$z_3(k;\eta) = \frac{z_3(k+1;\eta)}{1 - z_2(k+1;\eta)},\ k = 1, \ldots, w, \tag{A.46}$$

because $z_3(k;\eta)$ cannot take the value 1. In other words, the truncation by 1 on the right-hand side of (A.32) is immaterial for solutions of $f_w(\beta, q)$ and $z(k)$ for $k = 1, \ldots, w-1$ defined through (A.42)-(A.44).

We conclude the proof by showing that there exists a unique η satisfying conditions (A.42)-(A.44). Suppose there are multiple values of $f_w(\beta, q)$, say $\eta \neq \tilde{\eta}$, satisfying the conditions (A.42)-(A.44). Without loss of generality, assume $\eta > \tilde{\eta}$. Next, we show by

induction that for $k = 1, \ldots, w$,

$$z_1(k;\eta) \leq z_1(k;\tilde{\eta}), \quad z_2(k;\eta) \geq z_2(k;\tilde{\eta}) \text{ and } z_3(k;\eta) > z_3(k;\tilde{\eta}). \tag{A.47}$$

This is true for $k = w$ by assumption. Suppose it is true for k, then we argue that it is also true for $k-1$. Note that $z_1(k-1;\eta) = h_1(z(k;\eta))$ and $z_1(k-1;\tilde{\eta}) = h_1(z(k;\tilde{\eta}))$. Since $h_1(\cdot)$ is decreasing in its first argument whereas increasing in its last two arguments, we conclude that $z_1(k-1;\eta) \leq z_1(k-1;\tilde{\eta})$. Similarly, because $h_2(\cdot)$ is decreasing in the first argument whereas increasing in its last two arguments, we conclude that

$$z_2(k-1;\eta) = h_2(z(k;\eta)) \geq h_2(z(k;\tilde{\eta})) = z_2(k-1;\tilde{\eta}).$$

Also, it follows from equation (A.46) that

$$z_3(k-1;\eta) = \frac{z_3(k;\eta)}{1 - z_2(k;\eta)} > \frac{z_3(k;\tilde{\eta})}{1 - z_2(k;\tilde{\eta})} = z_3(k-1;\tilde{\eta}).$$

In particular, both of the following must be true:

$$z_3(1;\eta) > z_3(1;\tilde{\eta}) \text{ and } z_2(1;\eta) \geq z_2(1;\tilde{\eta}).$$

However, by equation (A.44), we conclude that

$$z_2(1;\eta) = 1 - z_3(1;\eta) < 1 - z_3(1;\tilde{\eta}) = \tilde{z}_2(1;\tilde{\eta}),$$

which is a contraction. Therefore, there exists at most one value of $f_w(\beta, q)$ satisfying conditions (A.42)-(A.44).

□

A.2.2 Partial derivatives of the auxiliary function $f_w(\cdot)$

This subsection characterizes the partial derivatives of $f_w(\cdot)$ and establishes the monotonicity of $f_w(\cdot)$.

To facilitate the analysis to follow, fix $w \geq 1$ and denote by $z(k; w, \beta, q)$ (for $k = 1, \ldots, w$) the $z(k)$s defined by substituting $z(w) = (\beta, q, f_w(\beta, q))$ into equation (A.43). The following lemma shows that in this construction, the truncation by 1 in defining $h_3(\cdot)$ is immaterial, cf. equation (A.32).

Lemma 31. *For $w \geq 1$ and $(\beta, q) \in (0, b] \times [q_\infty, 1)$, the following holds:*

$$z_3(k; w, \beta, q) = h_3(z(k+1; w, \beta, q)) = \frac{z_3(k+1; w, \beta, q)}{1 - z_2(k+1; w, \beta, q)},\ k = 1, \ldots, w-1.$$

Proof. It follows from Lemma 29 inductively that

$$z_3(1; w, \beta, q) \geq z_3(2; w, \beta, q) \geq \cdots \geq z_3(w; w, \beta, q) = f_w(\beta, q) > 0.$$

In addition, condition (A.44) ensures that $z_3(1; w, \beta, q) = 1 - z_2(1; w, \beta, q) \leq 1 - q_\infty < 1$. Combining these inequalities with the equation (A.32) yields the following: For $k = 1, \ldots, w-1$,

$$1 > z_3(k; w, \beta, q) = h_3(z(k+1; w, \beta, q)) = \min\left(1, \frac{z_3(k+1; w, \beta, q)}{1 - z_2(k+1; w, \beta, q)}\right).$$

Since $z_3(k; w, \beta, q)$ cannot take the value 1, the result follows. □

The following lemma provides an upper bound of $z_3(k; w, \beta, q)$ for $w \geq 1$.

Lemma 32. *For $w \geq 1$ and $(\beta, q) \in (0, b] \times [q_\infty, 1)$, we have the following inequality:*

$$z_3(k; w, \beta, q) \leq (1 - q_\infty)^k,\ k = 1, \ldots, w.$$

In particular, $f_w(\beta, q) \leq (1 - q_\infty)^w$ *for* $w \geq 1$.

Proof. We proceed by induction. As the induction basis, it follows from equation (A.44) that for $k = 1$,

$$z_3(1; w, \beta, q) = (1 - z_2(1; w, \beta, q)) \leq 1 - q_\infty,$$

where the inequality follows because $z_2(1; w, \beta, q) \in [q_\infty, 1)$ by construction.

As the induction hypothesis, suppose that the statement is true for k. Then note from Lemma 31 that

$$z_3(k + 1; w, \beta, q) = z_3(k; w, \beta, q)(1 - z_2(k + 1; w, \beta, q)) \leq (1 - q_\infty)^{k+1},$$

where the inequality follows from the induction hypothesis and that $z_2(k + 1; w, \beta, q) \in [q_\infty, 1)$.

In addition, it follows from (A.42) that

$$f_w(\beta, q) = z_3(w; w, \beta, q) \leq (1 - q_\infty)^w.$$

□

The following lemma characterizes the partial derivatives of $f_w(\cdot)$ recursively.

Lemma 33. *The partial derivatives of* $f_w(\cdot)$ *with respect to* β *and* q *for* $w \geq 1$ *and* $(\beta, q) \in (0, b] \times [q_\infty, 1)$ *are given as follows. For* $w = 1$*, we have that*

$$\frac{\partial f_1(\beta,)}{\partial \beta} = 0 \quad \textit{and} \quad \frac{\partial f_1(\beta, q)}{\partial q} = -1. \tag{A.48}$$

In addition, we have the following recursive characterization of $w \geq 1$*:*

$$\frac{\partial f_{w+1}(\beta,q)}{\partial \beta} = -\frac{\dfrac{\partial f_w(e(w))}{\partial \beta}\dfrac{\partial h_1(z(w+1))}{\partial z_1} + \dfrac{\partial f_w(e(w))}{\partial q}\dfrac{\partial h_2(z(w+1))}{\partial z_1}}{\dfrac{\partial f_w(e(w))}{\partial \beta}\dfrac{\partial h_1(z(w+1))}{\partial z_3} + \dfrac{\partial f_w(e(w))}{\partial q}\dfrac{\partial h_2(z(w+1))}{\partial z_3} - \dfrac{\partial h_3(z(w+1))}{\partial z_3}}, \tag{A.49}$$

$$\frac{\partial f_{w+1}(\beta,q)}{\partial q} = -\frac{\dfrac{\partial f_w(e(w))}{\partial q}\dfrac{\partial h_2(z(w+1))}{\partial z_2} - \dfrac{\partial h_3(z(w+1))}{\partial z_2}}{\dfrac{\partial f_w(e(w))}{\partial \beta}\dfrac{\partial h_1(z(w+1))}{\partial z_3} + \dfrac{\partial f_w(e(w))}{\partial q}\dfrac{\partial h_2(z(w+1))}{\partial z_3} - \dfrac{\partial h_3(z(w+1))}{\partial z_3}}, \tag{A.50}$$

where $z(k) = z(k; w+1, \beta, q)$ *for* $k = w, w+1$*, and* $e(w) = (z_1(w), z_2(w))$*.*

Proof. Note that $f_1(\beta, q) = 1 - q$. Hence equation (A.48) is immediate.

Fixing $w \geq 1$ and $(\beta, q) \in (0, b] \times [q_\infty, 1)$, we want to characterize the partial derivatives of $f_{w+1}(\cdot)$ with respect to β and q. Recall that for $w+1$, $z(w; w+1, \beta, q)$ is computed by substituting $z(w+1) = (\beta, q, f_{w+1}(\beta, q))$ into equation (A.43). In particular, we rewrite equation (A.43) that derives $z(w; w+1, \beta, q)$ as follows:

$$z_i(w; w+1, \beta, q) = h_i(\beta, q, f_{w+1}(\beta, q)). \tag{A.51}$$

Moreover, by substituting $z(w) = z(w; w+1, \beta, q)$ into equation (A.43) for w, we find that condition (A.44) (for w) is satisfied. In other words, solutions of (A.42)-(A.44) for different w's are consistent provided that β, q's are chosen consistently for each w. In particular, the following holds:

$$f_w(z_1(w; w+1, \beta, q), z_2(w; w+1, \beta, q)) = z_3(w; w+1, \beta, q). \tag{A.52}$$

Substituting equation (A.51) into (A.52), we obtain the following identity:

$$f_w(h_1(\beta, q, f_{w+1}(\beta, q)), h_2(\beta, q, f_{w+1}(\beta, q))) = h_3(\beta, q, f_{w+1}(\beta, q)). \tag{A.53}$$

Both the left-hand side and the right-hand side of equation (A.53) are functions of (β, q). Since we focus our analysis on the derivation for $w+1$ with fixed initial values (β, q), we write $z(k) = z(k; w+1, \beta, q)$ in short.

First, we take the partial derivative of both sides of equation (A.53) with respect to β by the chain rule and evaluate the function at point (β, q). It follows from equation (A.52) that the partial derivatives of $f_w(\cdot)$ are evaluated at $(z_1(w), z_2(w))$. Since $z(w+1) = (\beta, q, f_{w+1}(\beta, q))$, the partial derivatives of $h_i(\cdot)$ are evaluated at $z(w+1)$ for $i = 1, 2, 3$. Thus, we obtain the following equation:

$$\begin{aligned}
&\frac{\partial f_w(e(w))}{\partial \beta}\frac{\partial h_1(z(w+1))}{\partial z_1} + \frac{\partial f_w(e(w))}{\partial \beta}\frac{\partial h_1(z(w+1))}{\partial z_3}\frac{\partial f_{w+1}(\beta, q)}{\partial \beta} \\
+&\frac{\partial f_w(e(w))}{\partial q}\frac{\partial h_2(z(w+1))}{\partial z_1} + \frac{\partial f_w(e(w))}{\partial q}\frac{\partial h_2(z(w+1))}{\partial z_3}\frac{\partial f_{w+1}(\beta, q)}{\partial \beta} \\
=&\frac{\partial h_3(z(w+1))}{\partial z_1} + \frac{\partial h_3(z(w+1))}{\partial z_3}\frac{\partial f_{w+1}(\beta, q)}{\partial \beta},
\end{aligned}$$

where $e(w) = (z_1(w), z_2(w))$. Note that $\partial h_3/\partial z_1 = 0$ by (A.34). Thus, we can drop the first term on the right-hand side. Rearranging the terms yields equation (A.49).

Taking the partial derivative of both sides of equation (A.53) with respect to q and evaluating the function at value (β, q), we obtain the following equation:

$$\begin{aligned}
&\frac{\partial f_w(e(w))}{\partial \beta}\frac{\partial h_1(z(w+1))}{\partial z_2} + \frac{\partial f_w(e(w))}{\partial \beta}\frac{\partial h_1(z(w+1))}{\partial z_3}\frac{\partial f_{w+1}(\beta, q)}{\partial q} \\
+&\frac{\partial f_w(e(w))}{\partial q}\frac{\partial h_2(z(w+1))}{\partial z_2} + \frac{\partial f_w(e(w))}{\partial q}\frac{\partial h_2(z(w+1))}{\partial z_3}\frac{\partial f_{w+1}(\beta, q)}{\partial q} \\
=&\frac{\partial h_3(z(w+1))}{\partial z_2} + \frac{\partial h_3(z(w+1))}{\partial z_3}\frac{\partial f_{w+1}(\beta, q)}{\partial q}.
\end{aligned}$$

It follows from (A.39) that $\partial h_1/\partial z_2 = 0$. Thus, we can drop the first term on the left-hand

side. Rearranging the terms yields equation (A.50). □

The following lemma shows the monotonicity of $f_w(\cdot)$.

Lemma 34. *f_w is non-decreasing in β and non-increasing in q. That is, for all $w \geq 1$ and $(\beta, q) \in (0, b] \times [q_\infty, 1)$,*

$$\frac{\partial f_w(\beta, q)}{\partial \beta} \geq 0 \quad \textit{and} \quad \frac{\partial f_w(\beta, q)}{\partial q} \leq 0. \tag{A.54}$$

Proof. Recall that $f_1(\beta, q) = 1 - q$. Thus, (A.54) is immediate for $w = 1$.

We proceed by induction: Suppose (A.54) holds for $k = 1, \ldots, w$, we next show that it hold for $k = w + 1$. Note from equations (A.33)-(A.41) that

$$\frac{\partial h_1}{\partial z_1}, \ \frac{\partial h_2}{\partial z_3}, \ \frac{\partial h_3}{\partial z_3} > 0 \ \text{ and } \ \frac{\partial h_2}{\partial z_1}, \ \frac{\partial h_1}{\partial z_3} < 0.$$

Consider the formula for $\partial f_{w+1}/\partial \beta$ given in equation (A.49). Every term in the numerator is positive whereas every term in the denominator is negative so that

$$\frac{\partial f_{w+1}(\beta, q)}{\partial \beta} \geq 0.$$

Next, consider $\partial f_{w+1}/\partial q$. It follows from equations (A.33)-(A.41) that

$$\frac{\partial h_2}{\partial z_2} > 0 \ \text{ and } \ \frac{\partial h_3}{\partial z_2} \geq 0.$$

Every term in both the numerator and the denominator of equation (A.50) is negative. Thus, we conclude that

$$\frac{\partial f_{w+1}(\beta, q)}{\partial q} \leq 0.$$

□

A.2.3 Properties of $f_w(\cdot)$ on a restricted set

This subsection studies the partial derivatives of $f_w(\beta, q)$ as w gets large. To facilitate this analysis, we define subsets $\mathcal{Z}_1(w) \times \mathcal{Z}_2(w)$ for $w \geq 1$ such that for any potential equilibrium $(\beta^*(w), q^*(w)) \in \mathcal{Z}_1(w) \times \mathcal{Z}_2(w)$ for all $w \geq 1$. Restricting our analysis to the case where $(\beta, q) \in \mathcal{Z}_1(w) \times \mathcal{Z}_2(w) \subseteq (0, b] \times [q_\infty, 1)$, we establish the desired convergence results for the partial derivatives of $f_w(\beta, q)$. (Note from Corollaries 4 and 5 that $(\beta^*(w), q^*(w)) \to (b, q_\infty)$ as $w \to \infty$ for any potential equilibrium. We define $\mathcal{Z}_1(w) \times \mathcal{Z}_2(w)$ such that they shrink to the point (b, q_∞) as $w \to \infty$.)

To facilitate the analysis to follow, we first define a function $\kappa(x, y) : [0, 1] \times [0, r] \to \mathbb{R}$ as follows:

$$\kappa(x, y) = \mathbb{E}[-c + \alpha[xr + (1 - x)y] - (\varepsilon(1) - \varepsilon(0))]^+. \tag{A.55}$$

The following lemma shows the properties of κ.

Lemma 35. *The function $\kappa(x, y)$ has the following properties:*

(i) For any $(x, y) \in [0, 1] \times [0, r]$, $\kappa(x, y) \in [0, r)$.

(ii) For any fixed $x \in [0, 1]$, $\kappa(x, y)$ is a contraction mapping. In particular, $|\partial\kappa(x, y)/\partial y| \leq \alpha < 1$ for all $x \in [0, 1]$.

(iii) For any fixed $x \in [0, 1]$, there exists a unique $j(x) \in [0, r)$ satisfying $j(x) = \kappa(x, j(x))$.

Proof. The following inequality shows that (i) holds: For any $(x, y) \in [0, 1] \times [0, r]$,

$$\begin{aligned}
\kappa(x, y) &= \mathbb{E}[-c + \alpha[xr + (1 - x)y] - (\varepsilon(1) - \varepsilon(0))]^+ \\
&\leq \mathbb{E}[-c + \alpha[xr + (1 - x)r] - (\varepsilon(1) - \varepsilon(0))]^+ \\
&= \mathbb{E}[-c + \alpha r - (\varepsilon(1) - \varepsilon(0))]^+ \\
&= \mathbb{E}[\max\{\varepsilon(1), -c + \alpha r + \varepsilon(0)\}] < r.
\end{aligned} \tag{A.56}$$

The first inequality follows from $y \leq r$. The equality in the fourth line holds because

$\mathbb{E}[\varepsilon(1)] = 0$. The last equality follows from Assumption 2. In addition, it is immediate that $\kappa(x, y) \geq 0$ for all $(x, y) \in [0, 1] \times [0, r]$. Thus, we have that $\kappa(x, y) \in [0, r)$.

We can write $\kappa(x, y)$ in the integral form and use integration by parts to arrive at the following.

$$\begin{aligned}\kappa(x, y) &= \mathbb{E}[-c + \alpha[xr + (1 - x)y] - (\varepsilon(1) - \varepsilon(0))]^+ \\ &= \int_{-\infty}^{-c+\alpha[xr+(1-x)y]} (-c + \alpha[xr + (1 - x)y] - u) \, \mathrm{d}F(u) \\ &= \int_{-\infty}^{-c+\alpha[xr+(1-x)y]} F(u) \, \mathrm{d}u,\end{aligned} \tag{A.57}$$

where the last inequality follows from integration by parts. Thus, the partial derivative of $\kappa(x, y)$ with respect to y is given as follows:

$$\frac{\partial \kappa(x, y)}{\partial y} = \frac{\partial(\int_{-\infty}^{-c+\alpha[xr+(1-x)y]} F(u) \, \mathrm{d}u)}{\partial y} = \alpha(1 - x)F(-c + \alpha[xr + (1 - x)y]) \in [0, \alpha].$$

Therefore, for any fixed $x \in [0, 1]$, the following inequality holds:

$$\left| \frac{\partial \kappa(x, y)}{\partial y} \right| \leq \alpha < 1, \ y \in [0, r].$$

Thus, (ii) holds, i.e. $\kappa(x, y)$ is a contraction mapping for any fixed $x \in [0, 1]$.

It follows from properties (i)-(ii) that for any fixed $x \in [0, 1]$, $\kappa(x, y)$ is a contraction mapping from $[0, r]$ to $[0, r)$. By Banach fixed point theorem, there exists a unique fixed point $j(x) \in [0, r]$ such that $j(x) = \kappa(x, j(x))$. It follows from (A.56) that $\kappa(x, r) < r$. Thus, $j(x) \neq r$, which leads to $j(x) \in [0, r)$. In other words, property (iii) holds. □

The following lemma shows useful properties of the function $j(\cdot)$.

Lemma 36. *The function $j(\cdot)$ is increasing and differentiable.*

Proof. The function $j(x)$ is defined implicitly as follows:

$$\kappa(x, j(x)) - j(x) = 0, \ x \in [0, 1]. \tag{A.58}$$

Note that κ is continuously differentiable by equation (A.57). By the implicit function theorem, $j(x)$ is differentiable; see Theorem 9.28 of Rudin [96]. It follows from (A.57) that

$$\frac{\partial \kappa(x, y)}{\partial x} = \alpha(r - y)F(-c + \alpha[xr + (1 - x)y]) \geq \alpha(r - y)F(-c) > 0, \ (x, y) \in [0, 1] \times [0, r),$$

where the last inequality holds because $y < r$ and $-c$ is in the interior of the support of $F(\cdot)$ by Assumption 1. For any fixed $x \in [0, 1]$, taking the derivative of both sides of the equation $j(x) = \kappa(x, j(x))$ yields the following equation:

$$j'(x) = \frac{\partial \kappa(x, j(x))}{\partial x} + \frac{\partial \kappa(x, j(x))}{\partial y} j'(x).$$

Rearranging the terms, we have that

$$j'(x) = \frac{\partial \kappa(x, j(x))}{\partial x} \Big/ \left(1 - \frac{\partial \kappa(x, j(x))}{\partial y}\right) > 0, \ x \in [0, 1],$$

where the inequality follows from that $\partial \kappa(x, y)/\partial x > 0$ and property (ii) in Lemma 35. Therefore, $j(x)$ is increasing. □

To facilitate the definition of $\mathcal{Z}_1$ and $\mathcal{Z}_2$, the sequence $\underline{\beta}(w)$ is defined as follows:

$$\underline{\beta}(w) = \frac{b - a(1 - q_\infty)^w}{1 - a(1 - q_\infty)^w}, \ w \geq 1.$$

Since $b > a$, we have that $\underline{\beta}(w) > 0$ for all $w \geq 1$. Then we define $\underline{J}(w) = j(\underline{\beta}(w))$. By substituting equations (A.55) and (A.58) into the definition of $\underline{J}(w)$, we have that

$$\underline{J}(w) = \mathbb{E}\left[-c + \alpha[\underline{\beta}(w)r + (1 - \underline{\beta}(w))\underline{J}(w)] - (\varepsilon(1) - \varepsilon(0))\right]^+, \ w \geq 1. \tag{A.59}$$

In addition define[3]

$$\bar{q}(w) = \bar{F}(-c + \alpha[\underline{\beta}(w)r + (1 - \underline{\beta}(w))\underline{J}(w)]),\ w \geq 1. \tag{A.60}$$

The following lemma shows the properties of the sequences $\underline{\beta}(w)$, $\underline{J}(w)$ and $\bar{q}(w)$ for $w \geq 1$.

Lemma 37. *The sequences $\underline{\beta}(w)$, $\underline{J}(w)$ and $\bar{q}(w)$ for $w \geq 1$ have the following properties:*

(i) *$\underline{\beta}(w)$ and $\underline{J}(w)$ are increasing whereas $\bar{q}(w)$ is decreasing in w.*

(ii) *$\lim_{w\to\infty} \underline{\beta}(w) = b$, $\lim_{w\to\infty} \underline{J}(w) = J_\infty$ and $\lim_{w\to\infty} \bar{q}(w) = q_\infty$, where J_∞ and q_∞ are constants defined in Corollary 5.*

(iii) *$\underline{J}(w) = \int_{-\infty}^{\bar{F}^{-1}(\bar{q}(w))} F(x)\,\mathrm{d}x$.*

Proof. We first show (i). It is immediate to show that $\underline{\beta}(w)$ is increasing. Thus, $\underline{J}(w)$ is increasing in w by Lemma 36. It follows from property (iii) of Lemma 35 that $j(x) \in [0, r)$ for all $x \in [0, 1]$. Thus, $\underline{J}(w) < r$ for $w \geq 1$. It follows from equation (A.60) that $\bar{q}(w)$ is decreasing in w because $\underline{\beta}(w)$ and $\underline{J}(w)$ are non-increasing in w and $\underline{J}(w) < r$.

Next we show that (ii) holds. Clearly, $\lim_{w\to\infty} \underline{\beta}(w) = b$. It followsl from Lemma 35 that $j(x)$ is a differentiable function and thus is continuous. Therefore, the following equation holds:

$$\lim_{w\to\infty} \underline{J}(w) = \lim_{w\to\infty} j(\underline{\beta}(w)) = j(\lim_{w\to\infty} \underline{\beta}(w)) = j(b) = J_\infty,$$

where the last ineuqality follows from (3.7) that $J_\infty = j(b)$. It follows from the continuity of $\bar{F}$ that

$$\begin{aligned}\lim_{w\to\infty} \bar{q}(w) = &\lim_{w\to\infty} \bar{F}(-c + \alpha[\underline{\beta}(w)r + (1 - \underline{\beta}(w))\underline{J}(w)]) \\ =&\bar{F}(-c + \alpha[br + (1 - b)J_\infty]) = q_\infty.\end{aligned}$$

The last equality follows from the definition of q_∞ in Corollary 5.

3. The sequence $\underline{\beta}(w)$ provides a lower bound of $\beta^*(w)$ while $\bar{q}(w)$ is an upper bound of $q^*(w)$ in any potential equilibrium.

Lastly, we show that (iii) holds. It follows from equation (A.59)-(A.60) that for $w \geq 1$,

$$\begin{aligned}\underline{J}(w) =& \mathbb{E}\left[-c + \alpha[\underline{\beta}(w)r + (1 - \underline{\beta}(w))\underline{J}(w)] - (\varepsilon(1) - \varepsilon(0))\right]^+ \\ =& \mathbb{E}\left[\bar{F}^{-1}(\bar{q}(w)) - (\varepsilon(1) - \varepsilon(0))\right]^+ \\ =& \int_{-\infty}^{\bar{F}^{-1}(\bar{q}(w))} F(x)\,\mathrm{d}x\end{aligned}$$

□

To facilitate the analysis, define

$$\mathcal{Z}_1(w) = [\underline{\beta}(w), b] \text{ and } \mathcal{Z}_2(w) = [q_\infty, \bar{q}(w)],\ w \geq 1. \tag{A.61}$$

It follows from properties (i)-(ii) of Lemma 37 that $\underline{\beta}(w) < b$ and $\bar{q}(w) \geq q_\infty$. Thus, both $\mathcal{Z}_1(w)$ and $\mathcal{Z}_2(w)$ are nonempty for all $w \geq 1$. Since we only consider the underloaded case, i.e. $b > a$, we have that $\underline{\beta}(w) > 0$ for all $w \geq 1$. In addition, it follows from (A.60) that $\bar{q}(w) < \bar{F}(r) \leq 1$, which gives that

$$\mathcal{Z}_1(w) \times \mathcal{Z}_2(w) \subseteq (0, b] \times [q_\infty, 1),\ w \geq 1.$$

Next, we study the properties of $f_w(\beta, q)$ when $(\beta, q) \in \mathcal{Z}_1(w) \times \mathcal{Z}_w(w)$.

Lemma 38. *For any $w \geq 1$, if $(\beta, q) \in \mathcal{Z}_1(w) \times \mathcal{Z}_2(w)$, then*

$$(z_1(k; w, \beta, q), z_2(k; w, \beta, q)) \in \mathcal{Z}_1(k) \times \mathcal{Z}_2(k),\ k = 1, \ldots, w. \tag{A.62}$$

Proof. Fix $w \geq 1$. We proceed by induction. For $k = w$, (A.62) holds by assumption.

As the induction hypothesis, suppose (A.62) holds for $l = k+1, \ldots, w-1, w$. That is,

$$\underline{\beta}(l) \leq z_1(l; w, \beta, q) \leq b \text{ and } q_\infty \leq z_2(l; w, \beta, q) \leq \bar{q}(l),\ l = k+1, \ldots, w.$$

Next, we show that (A.62) holds for $l = k$. The following holds:

$$\begin{aligned}
z_1(k; w, \beta, q) &= h_1(z(k+1; w, \beta, q) \\
&= \left(1 + \frac{1-b}{1 - az_3(k+1; w, \beta, q)} \frac{1}{z_2(k+1; w, \beta, q)}\right)^{-1} \\
&> \left(1 + \frac{1-b}{1-a(1-q_\infty)^{k+1}} \frac{1-a(1-q_\infty)^{k+1}}{b-a(1-q_\infty)^{k+1}}\right)^{-1} \\
&= \left(1 + \frac{1-b}{b-a(1-q_\infty)^{k+1}}\right)^{-1} \\
&> \left(1 + \frac{1-b}{b-a(1-q_\infty)^{k}}\right)^{-1} \\
&= \frac{b-a(1-q_\infty)^k}{1-a(1-q_\infty)^k} = \underline{\beta}(k).
\end{aligned}$$

The first inequality follows from Lemma 32 and that $z_2(k+1; w, \beta, q) > \beta(k+1)$. Thus, $z_1(k; w, \beta, q) \in \mathcal{Z}_1(k)$. Combining the two cases, $z_1(k; w, \beta, q) \in \mathcal{Z}_1(k)$.

Moreover, it follows from equation (A.31) that

$$\begin{aligned}
z_2(k; w, \beta, q) =& h_2(z(k+1; w, \beta, q)) \\
=& \bar{F}\left[-c + \alpha\left(z_1(k; w, \beta, q)r + (1 - z_1(k; w, \beta, q)) \int_{-\infty}^{\bar{F}^{-1}(z_2(k+1; w, \beta, q))} F(x)\,\mathrm{d}x\right)\right] \\
\leq& \bar{F}\left[-c + \alpha\left(\underline{\beta}(k)r + (1 - \underline{\beta}(k)) \int_{-\infty}^{\bar{F}^{-1}(\bar{q}(k+1))} F(x)\,\mathrm{d}x\right)\right] \\
\leq& \bar{F}\left[-c + \alpha\left(\underline{\beta}(k)r + (1 - \underline{\beta}(k)) \int_{-\infty}^{\bar{F}^{-1}(\bar{q}(k)} F(x)\,\mathrm{d}x\right)\right] \\
=& \bar{F}\left[-c + \alpha\left(\underline{\beta}(k)r + (1 - \underline{\beta}(k))\underline{J}(k)\right)\right] = \bar{q}(k).
\end{aligned}$$

The first inequality follows from $z_1(k; w, \beta, q) \geq \underline{\beta}(k)$ and the assumption that $z_2(k+1; w, \beta, q) \leq \bar{q}(k+1)$ and Lemma 27 that the integral is less than r. The second inequality follows from Lemma 37 that $\bar{q}(k+1) < \bar{q}(k)$ and the last two equalities follow from property (iii) of Lemma 37 and (A.60). Since $z_2(k; w, \beta, q) \geq q_\infty$ by construction, $z_2(k; w, \beta, q) \in \mathcal{Z}_2(k)$. □

The following lemma shows the properties of the partial derivatives of $h(\cdot)$ for values in the set $\mathcal{Z}_1(w+1) \times \mathcal{Z}_2(w+1)$.

Lemma 39. *For any $\epsilon > 0$, there exist $w_0, M \geq 0$ such that the following holds for $w \geq w_0$ and $(\beta, q) \in \mathcal{Z}_1(w+1) \times \mathcal{Z}_2(w+1)$:*

$$\frac{\partial h_2(z)}{\partial z_2} \leq 1, \ \frac{\partial h_3(z)}{\partial z_2} \leq \frac{\epsilon}{2} q_\infty, \ \frac{\partial h_1(z)}{\partial z_1} \leq 1 \quad \textit{and} \quad -\frac{\partial h_2(z)}{\partial z_1} \leq M, \tag{A.63}$$

where $z = z(w+1; w+1, \beta, q) = (\beta, q, f_{w+1}(\beta, q))$.

Proof. We show that the four inequalities hold for w large enough one-by-one.

We first show that $\partial h_2 / \partial z_2 \leq 1$ for w large. Recall from equation (A.37) that for $z \in \mathcal{Z}$,

$$\frac{\partial h_2(z)}{\partial z_2} = \alpha(1-\beta)(1-q)\frac{f[\bar{F}^{-1}(h_2(z))]}{f(\bar{F}^{-1}(q))} \leq \alpha(1-q_\infty)\frac{f[\bar{F}^{-1}(h_2(z))]}{f(\bar{F}^{-1}(q))}, \tag{A.64}$$

where the inequality follows because $q \geq q_\infty$ and $\beta \in \mathcal{Z}_1(w+1) \subseteq (0, 1]$.

By continuity of $f(\cdot)$ and $\bar{F}^{-1}(\cdot)$ at q_∞, there exists $\delta_1 > 0$ such that for all x such that $|x - q_\infty| < \delta_1$, the following holds:

$$f(\bar{F}^{-1}(q_\infty))\sqrt{1-q_\infty} \leq f(\bar{F}^{-1}(x)) \leq \frac{1}{\sqrt{1-q_\infty}} f(\bar{F}^{-1}(q_\infty)). \tag{A.65}$$

It follows from Lemma 37 that $\bar{q}(w) \to q_\infty$ as $w \to \infty$. Thus, there exists $w_1 \geq 1$ such that $|\bar{q}(w) - q_\infty| < \delta_1$ for $w \geq w_1$. In particular,

$$|\bar{q}(w+1) - q_\infty| < \delta_1 \ \text{ and } \ |\bar{q}(w) - q_\infty| < \delta_1, \ w \geq w_1. \tag{A.66}$$

It follows from (A.43) that $h_2(z) = z_2(w; w+1, \beta, q)$. We have that $h_2(z) \in \mathcal{Z}_2(w) = [q_\infty, \bar{q}(w)]$ by Lemma 38. In addition, by assumption, $q \in \mathcal{Z}_2(w+1) = [q_\infty, \bar{q}(w+1)]$. Thus,

it follows from (A.66) that for $w \geq w_1$ and $(\beta, q) \in \mathcal{Z}_1(w+1) \times \mathcal{Z}_2(w+1)$,

$$|q - q_\infty| < \delta_1 \text{ and } |h_2(z) - q_\infty| < \delta_1.$$

By equation (A.65), we have that for $w \geq w_1$ and $(\beta, q) \in \mathcal{Z}_1(w+1) \times \mathcal{Z}_2(w+1)$,

$$\frac{f[\bar{F}^{-1}(h_2(z))]}{f(\bar{F}^{-1}(q))} \leq \frac{1}{1-q_\infty}.$$

Substituting this inequality into equation (A.64), we obtain the following:

$$\frac{\partial h_2(z)}{\partial z_2} \leq \alpha \leq 1 \text{ for } w \geq w_1 \text{ and } (\beta, q) \in \mathcal{Z}_1(w+1) \times \mathcal{Z}_2(w+1),$$

where $z = (\beta, q, f_{w+1}(\beta, q)) = z(w+1; w+1, \beta, q)$.

Next we show that $\partial h_3/\partial z_2 \leq \epsilon q_\infty/2$ for w large. It follows from equation (A.40) that

$$\frac{\partial h_3(z)}{\partial z_2} = \frac{z_3(w+1; w+1, \beta, q)}{(1-q)^2} \leq \frac{(1-q_\infty)^{w+1}}{(1-q)^2} \leq \frac{(1-q_\infty)^w}{(1-\bar{q}(w_1))^2}, \tag{A.67}$$

where the first inequality follows from Lemma 32. Recall from Lemma 37 that $\bar{q}(w)$ is decreasing. Thus, the second inequality holds because

$$q \leq \bar{q}(w+1) \leq \bar{q}(w_1), \ w \geq w_1.$$

Since $1 - q_\infty < 1$, there exists a constant $w_2 \geq w_1$ such that

$$(1-q_\infty)^w \leq \frac{\epsilon}{2} q_\infty (1-\bar{q}(w_1))^2, \ w \geq w_2.$$

Substituting this inequality into (A.67), we have the following:

$$\frac{\partial h_3(z)}{\partial z_2} \leq \frac{\epsilon}{2} q_\infty \text{ for } w \geq w_2 \text{ and } (\beta, q) \in \mathcal{Z}_1(w+1) \times \mathcal{Z}_2(w+1).$$

We then show that $\partial h_1/z_1 \leq 1$ for w large enough. Recall from equation (A.33) that for $w \geq w_2$,

$$0 \leq \frac{\partial h_1(z)}{\partial z_1} = \frac{h_1^2(z)}{\beta^2} \frac{1-b}{1-az_3(w+1;w+1,\beta,q))} \leq \frac{h_1^2(z)}{\beta^2} \frac{1-b}{1-a(1-q_\infty)^{w+1}}, \tag{A.68}$$

where the inequality follows from Lemma 32. Note by assumption that $\beta \in \mathcal{Z}_1(w+1) = [\underline{\beta}(w+1), b]$ and Lemma 38 that $h_1(z) = z_1(w; w+1, \beta, q) \in \mathcal{Z}_1(w) = [\underline{\beta}(w), b]$. In other words, the following holds:

$$\underline{\beta}(w+1) \leq \beta \leq b \text{ and } \underline{\beta}(w) \leq h_1(z) \leq b.$$

It follows from Lemma 37 that $\underline{\beta}(w) \to b$ as $w \to \infty$. In addition, $1-a(1-q_\infty)^{w+1} \to 1$ as $w \to \infty$. Thus, there exists $w_3 \geq w_2$ such that

$$\frac{b}{\underline{\beta}(w+1)} \leq \frac{1}{\sqrt[4]{1-b}} \text{ and } 1-a(1-q_\infty)^{w+1} > \sqrt[4]{1-b},\ w \geq w_3.$$

Substituting these two inequalities into (A.68), we have that for $w \geq w_3$ and $(\beta, q) \in \mathcal{Z}_1(w+1) \times \mathcal{Z}_2(w+1)$,

$$0 \leq \frac{\partial h_1(z)}{\partial z_1} \leq \frac{h_1^2(z)}{\beta^2} \frac{1-b}{1-a(1-q_\infty)^{w+1}} \leq \frac{b^2}{\underline{\beta}^2(w+1)} \frac{1-b}{1-a(1-q_\infty)^{w+1}} \leq \sqrt[4]{1-b} < 1.$$

Lastly, we show that there exists $M \geq 0$ such that $\partial h_2/\partial z_1 \leq M$ for w large enough. It follows from equation (A.36) that for all $w \geq w_3$ and $(\beta, q) \in \mathcal{Z}_1(w+1) \times \mathcal{Z}_2(w+1)$,

$$\begin{aligned} -\frac{\partial h_2(z)}{\partial z_1} =& f(\bar{F}^{-1}(h_2(z)))\alpha \left(r - \int_{-\infty}^{\bar{F}^{-1}(q)} F(x)\,\mathrm{d}x \right) \frac{\partial h_1(z)}{\partial z_1} \\ \leq& f(\bar{F}^{-1}(h_2(z))\alpha r \frac{\partial h_1(z)}{\partial z_1} \leq f(\bar{F}^{-1}(h_2(z))\alpha r. \end{aligned} \tag{A.69}$$

The first inequality follows from Lemma 27. The second inequality follows from the first

inequality in (A.63). It follows from the continuity of $f(\cdot)$ and $\bar{F}^{-1}(\cdot)$ that $f(\bar{F}^{-1}(x))$ is bounded on $[q_\infty, \bar{q}(w_3)]$. Recall that $h_2(z) \in \mathcal{Z}_2(w) = [q_\infty, \bar{q}(w)]$. Since $\bar{q}(w)$ is decreasing in w by Lemma 37, it follows that

$$q_\infty \leq h_2(z) \leq \bar{q}(w) \leq \bar{q}(w_3), \ w \geq w_3 \ \text{ and } \ (\beta, q) \in \mathcal{Z}_1(w+1) \times \mathcal{Z}_2(w+1).$$

Thus, the right-hand side of the third line in equation (A.69) is bounded. Letting M denote one of such bounds completes the proof.

In sum, letting $w_0 = w_3$, the four inequalities in (A.63) hold for all $w \geq w_0$ and $(\beta, q) \in \mathcal{Z}_1(w+1) \times \mathcal{Z}_2(w+1)$ □

The next lemma is key to proving Lemma 4.

Lemma 40. *The following holds:*

$$\lim_{w\to\infty} \sup \left\{ \left| \frac{\partial f_w(\beta, q)}{\partial \beta} \right| : (\beta, q) \in \mathcal{Z}_1(w) \times \mathcal{Z}_2(w) \right\} = 0, \tag{A.70}$$

$$\lim_{w\to\infty} \sup \left\{ \left| \frac{\partial f_w(\beta, q)}{\partial q} \right| : (\beta, q) \in \mathcal{Z}_1(w) \times \mathcal{Z}_2(w) \right\} = 0. \tag{A.71}$$

Proof. We first show (A.70). To facilitate the analysis to follow, define a sequence y_w for $w \geq 1$ as follows:

$$y_w = \sup \left\{ -\frac{\partial f_w(\beta, q)}{\partial q} : (\beta, q) \in \mathcal{Z}_1(w) \times \mathcal{Z}_2(w) \right\}. \tag{A.72}$$

It follows from equation (A.54) that $y_w \geq 0$ for all $w \geq 1$. In addition, it follows from

equation (A.50) that for any $(\beta, q) \in \mathcal{Z}_1(w+1) \times \mathcal{Z}_2(w+1)$ and $w \geq 1$,

$$\begin{aligned}
-\frac{\partial f_{w+1}(\beta,q)}{\partial q} =& \frac{\frac{\partial f_w(e(w))}{\partial q}\frac{\partial h_2(z(w+1))}{\partial z_2} - \frac{\partial h_3(z(w+1))}{\partial z_2}}{\frac{\partial f_w(e(w))}{\partial \beta}\frac{\partial h_1(z(w+1))}{\partial z_3} + \frac{\partial f_w(e(w))}{\partial q}\frac{\partial h_2(z(w+1))}{\partial z_3} - \frac{\partial h_3(z(w+1))}{\partial z_3}} \\
=& \frac{-\frac{\partial f_w(e(w))}{\partial q}\frac{\partial h_2(z(w+1))}{\partial z_2} + \frac{\partial h_3(z(w+1))}{\partial z_2}}{-\frac{\partial f_w(e(w))}{\partial \beta}\frac{\partial h_1(z(w+1))}{\partial z_3} - \frac{\partial f_w(e(w))}{\partial q}\frac{\partial h_2(z(w+1))}{\partial z_3} + \frac{\partial h_3(z(w+1))}{\partial z_3}} \\
\leq& \left(-\frac{\partial f_w(e(w))}{\partial q}\frac{\partial h_2(z(w+1))}{\partial z_2} + \frac{\partial h_3(z(w+1))}{\partial z_2}\right) / \frac{\partial h_3(z(w+1))}{\partial z_3} \\
\leq& \left(y_w \frac{\partial h_2(z(w+1))}{\partial z_2} + \frac{\partial h_3(z(w+1))}{\partial z_2}\right) / \frac{\partial h_3(z(w+1))}{\partial z_3},
\end{aligned} \tag{A.73}$$

where $z(k) = z(k; w+1, \beta, q)$ for $k = w, w+1$ and $e(w) = (z_1(w), z_2(w))$. We flip the signs of the terms in the second line of the right-hand side. Thus, it follows from Lemma 28 that every term in both the numerator and the denominator (of the right-hand side of the third line) is positive. This leads to the inequality in the fourth line. The last inequality follows from equation (A.72) because $(z_1(w), z_2(w)) \in \mathcal{Z}_1(w) \times \mathcal{Z}_2(w)$, which in turn follows from Lemma 38.

Rewriting equation (A.73) gives

$$-\frac{\partial f_{w+1}(\beta,q)}{\partial q} \leq \left(y_w \frac{\partial h_2(z(w+1))}{\partial z_2} + \frac{\partial h_3(z(w+1))}{\partial z_2}\right) / \frac{\partial h_3(z(w+1))}{\partial z_3}.$$

Taking the supremum of both sides over $(\beta, q) \in \mathcal{Z}_1(w+1) \times \mathcal{Z}_2(w+1)$ gives the following:

$$y_{w+1} \leq \sup_{(\beta,q)\in\mathcal{Z}_1(w+1)\times\mathcal{Z}_2(w+1)} \left[\left(y_w \frac{\partial h_2(z(w+1))}{\partial z_2} + \frac{\partial h_3(z(w+1))}{\partial z_2} \right) / \frac{\partial h_3(z(w+1))}{\partial z_3} \right]$$

$$\leq \sup_{(\beta,q)\in\mathcal{Z}_1(w+1)\times\mathcal{Z}_2(w+1)} (1-q_\infty) \left(y_w \frac{\partial h_2(z(w+1))}{\partial z_2} + \frac{\partial h_3(z(w+1))}{\partial z_2} \right), \tag{A.74}$$

where the last inequality follows from equation (A.41) and that $q \in [q_\infty, \bar{q}(w+1)]$. In particular,

$$\frac{\partial h_3(z(w+1))}{\partial z_3} = \frac{1}{1-z_2(w+1)} = \frac{1}{1-q} \geq \frac{1}{1-q_\infty}. \tag{A.75}$$

Substituting the first two inequalities in equation (A.63) into (A.74) yields the following:

$$y_{w+1} \leq (1-q_\infty)\left(y_w + \frac{\epsilon}{2} q_\infty \right) \quad \text{for} \quad w \geq w_0,$$

where w_0 is as in Lemma 39. By induction, we obtain the following inequality: For all $w \geq w_0$ and $n \geq 1$,

$$\begin{aligned} y_{w+n} \leq & (1-q_\infty) y_{w+n-1} + \frac{\epsilon}{2} q_\infty (1-q_\infty) \\ \leq & (1-q_\infty)^2 y_{w+n-2} + \frac{\epsilon}{2} q_\infty (1-q_\infty)^2 + \frac{\epsilon}{2} q_\infty (1-q_\infty) \\ \leq & \cdots \leq (1-q_\infty)^n y_w + \frac{\epsilon}{2} q_\infty \sum_{i=1}^{n} (1-q_\infty)^i \\ = & (1-q_\infty)^n y_w + \frac{\epsilon}{2}. \end{aligned}$$

Now fix $w = w_0$. There exists n_1 such that for all $n \geq n_1$, $(1-q_\infty)^n y_{w_0} < \epsilon/2$. That is, for any $w \geq w_0 + n_1$, $y_w < \epsilon$. Therefore, $y_w \to 0$, as $w \to \infty$. Since $y_w \geq 1$ for all $w \geq 1$, we deduce equation (A.70).

Next, we prove (A.71) in a similar fashion. Define a sequence x_w as follows:

$$x_w = \sup \left\{ \frac{\partial f_w(\beta, q)}{\partial \beta} : (\beta, q) \in \mathcal{Z}_1(w) \times \mathcal{Z}_2(w) \right\}, \ w \geq 1.$$

It follows from equation (A.54) that $x_w \geq 0$ for all $w \geq 1$. In addition, it follows from equation (A.49) that the following holds: For any $w \geq 1$ and $(\beta, q) \in \mathcal{Z}_1(w+1) \times \mathcal{Z}_2(w+1)$

$$\begin{aligned}
\frac{\partial f_{w+1}(\beta, q)}{\partial \beta} =& - \frac{\dfrac{\partial f_w(e(w))}{\partial \beta} \dfrac{\partial h_1(z(w+1))}{\partial z_1} + \dfrac{\partial f_w(e(w))}{\partial q} \dfrac{\partial h_2(z(w+1))}{\partial z_1}}{\dfrac{\partial f_w(e(w))}{\partial \beta} \dfrac{\partial h_1(z(w+1))}{\partial z_3} + \dfrac{\partial f_w(e(w))}{\partial q} \dfrac{\partial h_2(z(w+1))}{\partial z_3} - \dfrac{\partial h_3(z(w+1))}{\partial z_3}} \\
=& \frac{\dfrac{\partial f_w(e(w))}{\partial \beta} \dfrac{\partial h_1(z(w+1))}{\partial z_1} + \dfrac{\partial f_w(e(w))}{\partial q} \dfrac{\partial h_2(z(w+1))}{\partial z_1}}{-\dfrac{\partial f_w(e(w))}{\partial \beta} \dfrac{\partial h_1(z(w+1))}{\partial z_3} - \dfrac{\partial f_w(e(w))}{\partial q} \dfrac{\partial h_2(z(w+1))}{\partial z_3} + \dfrac{\partial h_3(z(w+1))}{\partial z_3}} \\
\leq& \left(\frac{\partial f_w(e(w))}{\partial \beta} \frac{\partial h_1(z(w+1))}{\partial z_1} + \frac{\partial f_w(e(w))}{\partial q} \frac{\partial h_2(z(w+1))}{\partial z_1} \right) / \frac{\partial h_3(z(w+1))}{\partial z_3} \\
\leq& \left(x_w \frac{\partial h_1(z(w+1))}{\partial z_1} + y_w \left(- \frac{\partial h_2(z(w+1))}{\partial z_1} \right) \right) / \frac{\partial h_3(z(w+1))}{\partial z_3}.
\end{aligned}$$

The first inequality holds because every term in both the numerator and denominator of the right-hand side of the second line is positive. The last inequality holds because $(z_1(w), z_2(w)) \in \mathcal{Z}_1(w) \times \mathcal{Z}_2(w)$. Taking the supremum of both sides over $(\beta, q) \in \mathcal{Z}_1(w+1) \times \mathcal{Z}_2(w+1)$ yields the following: For $w \geq 1$,

$$x_{w+1} \leq \sup_{(\beta,q) \in \mathcal{Z}_1(w+1) \times \mathcal{Z}_2(w+1)} \left(x_w \frac{\partial h_1(z(w+1))}{\partial z_1} - y_w \frac{\partial h_2(z(w+1))}{\partial z_1} \right) / \frac{\partial h_3(z(w+1))}{\partial z_3}.$$

Substituting the last two inequalities in equation (A.63) and equation (A.75) into this in-

equality yields the following:

$$x_{w+1} \leq (1 - q_\infty)(x_w + My_w) \quad \text{for} \quad w \geq w_0.$$

We have just shown that $y_w \to 0$ as $w \to \infty$. Therefore, for $\epsilon > 0$, there exists $w_1 \geq w_0$ such that $y_w < q_\infty \epsilon / 2M$ for $w \geq w_1$. Thus, the following holds:

$$x_{w+1} \leq (1 - q_\infty) \left(x_w + \frac{\epsilon}{2} q_\infty \right) \quad \text{for} \quad w \geq w_1.$$

Fixing $w = w_1$, we have that

$$\begin{aligned}
x_{w_1+n} \leq & (1 - q_\infty) x_{w_1+n-1} + \frac{\epsilon}{2} q_\infty (1 - q_\infty) \\
\leq & (1 - q_\infty)^2 x_{w_1+n-2} + \frac{\epsilon}{2} q_\infty (1 - q_\infty)^2 + \frac{\epsilon}{2} q_\infty (1 - q_\infty) \\
\leq & \cdots \leq (1 - q_\infty)^n x_{w_1} + \frac{\epsilon}{2} q_\infty \sum_{i=1}^{n} (1 - q_\infty)^i \\
\leq & (1 - q_\infty)^n x_{w_1} + \frac{\epsilon}{2}.
\end{aligned}$$

There exists n_2 such that for $n \geq n_2$, $(1 - q_\infty)^n x_{w_1} < \epsilon/2$. Thus, for $w \geq w_1 + n_2$, $x_w < \epsilon$. Since $x_w \geq 1$, we have that $\lim_{w \to \infty} x_w = 0$, which gives (A.71). □

A.2.4 Characterizing the equilibrium quantities with $f_w(\cdot)$

This subsection relates the equilibrium quantities with $f_w(\cdot)$. The following lemma shows that the equilibrium e^* can be characterized by $h(\cdot)$.

Lemma 41. *For any equilibrium e^*, the following holds:*

$$(\beta^*(w), q^*(w), \bar{G}^*(w)) = h(\beta^*(w+1), q^*(w+1), \bar{G}^*(w+1)), \ w \geq 1,$$

Proof. It follows from (3.4) and (A.30) that $\beta^*(w) = h_1(\beta^*(w+1), q^*(w+1), \bar{G}^*(w+1))$

for $w \geq 1$.

It follows from equation (2.11) that

$$J^*(w) = \mathbb{E}_\varepsilon[\bar{F}^{-1}(q^*(w)) - (\varepsilon(1) - \varepsilon(0)]^+ = \int_{-\infty}^{\bar{F}^{-1}(q^*(w))} F(x)\,\mathrm{d}x.$$

Substituting this equation and $\beta^*(w) = h_1(\beta^*(w+1), q^*(w+1), \bar{G}^*(w+1))$ into (2.10) and comparing it with (A.31), we have that

$$\begin{aligned} q^*(w) =& \bar{F}\left(-c + \alpha\left\{\beta^*(w)r + (1-\beta^*(w))\int_{-\infty}^{\bar{F}^{-1}(q^*(w))} F(x)\,\mathrm{d}x\right\}\right) \\ =& h_2(\beta^*(w+1), q^*(w+1), \bar{G}^*(w+1)). \end{aligned}$$

In addition, it follows from equations (3.1) and (A.32) that

$$\bar{G}^*(w) = \frac{\bar{G}^*(w+1)}{1-q^*(w+1)} = \min\left(1, \frac{\bar{G}^*(w+1)}{1-q^*(w+1)}\right) = h_3(\beta^*(w+1), q^*(w+1), \bar{G}^*(w+1)),$$

where the second equality holds because $\bar{G}^*(w) = \prod_{i=1}^{w}(1-q^*(i)) \leq 1$. □

Thus, it is immediate that $f_w(\cdot)$ characterizes $\bar{G}^*(w)$ in terms of $\beta^*(w)$ and $q^*(w)$, which is formalized in the following corollary.

Corollary 13. *We have that $\bar{G}^*(w) = f_w(e^*(w))$ for all $w \geq 1$.*

Proof. Fixing a $w \geq 1$ and substituting $z(w) = (\beta^*(w), q^*(w), \bar{G}^*(w))$ into equation (A.43). By applying Lemma 41 inductively, we have that $z(1) = (\beta^*(1), q^*(1), \bar{G}^*(1))$. Note that the resulting $z(1) = (\beta^*(1), q^*(1), \bar{G}^*(1))$ satisfies condition (A.44). In particular, $\bar{G}^*(1) = 1 - q^*(1)$. Thus, it follows from the definition of function $f_w(\cdot)$ that $\bar{G}^*(w) = f_w(e^*(w))$ for all $w \geq 1$. □

The following lemma shows that the equilibrium quantities live in the set $\mathcal{Z}_1(w) \times \mathcal{Z}_2(w)$.

Lemma 42. *For any equilibrium $e^* = (\beta^*, q^*)$, we have that*

$$e^*(w) = (\beta^*(w), q^*(w)) \in \mathcal{Z}_1(w) \times \mathcal{Z}_2(w),\ w \geq 1.$$

Proof. By Lemma 3 and Corollaries 4-5, $\beta^*(w) \leq b$ and $q^*(w) \geq q_\infty$. Thus, it suffices to show that $\beta^*(w) \geq \underline{\beta}(w)$ and $q^*(w) \leq \bar{q}(w)$ for all $w \geq 1$.

We first show that $\beta^*(w) \geq \underline{\beta}(w)$. It follows from Proposition 2 that

$$\begin{aligned}
\beta^*(w) &= \left(1 + \sum_{t=w}^{\infty} \prod_{i=w}^{t} \frac{1-b}{1 - a\bar{G}^*(i+1)}\right)^{-1} \\
&\geq \left(1 + \sum_{t=w}^{\infty} \prod_{i=w}^{t} \frac{1-b}{1 - a\bar{G}^*(w)}\right)^{-1} \\
&= \left(1 + \sum_{i=1}^{\infty} \left(\frac{1-b}{1 - a\bar{G}^*(w)}\right)^{i}\right)^{-1} \\
&= 1 - \frac{1-b}{1 - a\bar{G}^*(w)} \\
&\geq 1 - \frac{1-b}{1 - a(1-q_\infty)^w} = \frac{b - a(1-q_\infty)^w}{1 - a(1-q_\infty)^w} = \underline{\beta}(w),
\end{aligned}$$

where the first inequality follows from

$$\bar{G}^*(i+1) = \bar{G}^*(w) \prod_{j=w+1}^{i+1} (1 - q^*(j)) \leq \bar{G}^*(w),\ i \geq w,$$

and the last inequality follows from Lemma 32. In particular, it follows from

$$\bar{G}^*(w) = z_3(w; w, \beta^*(w), q^*(w)) \leq (1 - q_\infty)^w.$$

Therefore, we have that $\beta^*(w) \in \mathcal{Z}_1(w)$.

We then prove that $q^*(w) \in \mathcal{Z}_2(w)$. We first show that $J^*(w) \geq \underline{J}(w)$ for all $w \geq 1$ by contradiction, where $J^*(w)$ is the expected discounted utility of waiting. Suppose this is not

true and there exists w_0 such that $J^*(w_0) < \underline{J}(w_0)$.

We first show by induction that $J^*(w) < \underline{J}(w_0)$ for all $w \geq w_0$. It is true for w_0 by assumption. As the induction hypothesis, suppose it is true for $k = w$. In particular, $J^*(w) < \underline{J}(w_0)$. We then show that $J^*(w+1) < \underline{J}(w_0)$. Substituting equation (2.8) and $\underline{J}(w_0) = \kappa(\underline{\beta}(w_0), \underline{J}(w_0))$ into $J^*(w) < \underline{J}(w_0)$, we obtain that

$$\begin{aligned}
&\mathbb{E}[-c + \alpha[\beta^*(w)r + (1-\beta^*(w))J^*(w+1)] - (\varepsilon(1) - \varepsilon(0))]^+ \\
=&\mathbb{E}[\max\{\varepsilon(1), -c + \alpha[\beta^*(w)r + (1-\beta^*(w))J^*(w+1)] + \varepsilon(0)\}] \\
=&J^*(w) < \underline{J}(w_0) \\
=&\mathbb{E}[-c + \alpha[\underline{\beta}(w_0)r + (1-\underline{\beta}(w_0))\underline{J}(w_0)] - (\varepsilon(1) - \varepsilon(0))]^+,
\end{aligned}$$

where the first equality holds because $\mathbb{E}[\varepsilon(1)] = 0$. Comparing the first line and right-hand side of the last line, we conclude that the following inequality holds:

$$\begin{aligned}
&\beta^*(w)r + (1-\beta^*(w))J^*(w+1) \\
<&\underline{\beta}(w_0)r + (1-\underline{\beta}(w_0))\underline{J}(w_0) \\
=&\beta^*(w)r + (1-\beta^*(w))\underline{J}(w_0) + (\underline{\beta}(w_0) - \beta^*(w))(r - \underline{J}(w_0)).
\end{aligned}$$

Rearranging the terms, we have that

$$J^*(w+1) < \underline{J}(w_0) + \frac{(\underline{\beta}(w_0) - \beta^*(w))(r - \underline{J}(w_0))}{1 - \beta^*(w)}. \tag{A.76}$$

Note that the last term in the right-hand side of equation (A.76) is nonpositive. Too see this, recall that we have shown $\beta^*(w) \geq \underline{\beta}(w)$ at the beginning of this proof. It follows from property (i) of Lemma 37 that $\underline{\beta}(w) \geq \underline{\beta}(w_0)$ for $w \geq w_0$. Combining the two inequalities, we have that $\underline{\beta}(w_0) - \beta^*(w) \leq 0$. In addition, it follows from Lemma 37 that $r - \underline{J}(w_0) \geq 0$. Since $1 - \hat{\beta}^*(w) > 0$, we have that the last term in the right-hand side of equation (A.76) is nonpositive.

By dropping the nonpositive term in the right-hand side of (A.76), we have that $J^*(w+1) < \underline{J}(w_0)$, which completes the induction argument. In sum, $J^*(w) < \underline{J}(w_0)$ for all $w \geq w_0$.

On one hand, by the induction argument, we prove that $J^*(w) < \underline{J}(w_0)$ for all $w \geq w_0$. Thus, it follows from Corollary 5 that $J_\infty = \lim_{w\to\infty} J^*(w) \leq \underline{J}(w_0)$. On the other hand, it follows from properties (i)-(ii) in Lemma 37 that $\underline{J}(w_0) < J_\infty$. This leads to a contradiction. Therefore, there exists no w_0 such that $J^*(w_0) < \underline{J}(w_0)$. In other words, $J^*(w) \geq \underline{J}(w)$ for all $w \geq 1$.

We complete the proof by showing $q^*(w) \leq \bar{q}(w)$ for $w \geq 1$. It follows from equation (2.11) that

$$J^*(w) = \mathbb{E}_\varepsilon[\bar{F}^{-1}(q^*(w)) - (\varepsilon(1) - \varepsilon(0))^+ = \int_{-\infty}^{\bar{F}^{-1}(q^*(w))} F(x)\,\mathrm{d}x.$$

In addition, recall from Lemma 37 that

$$\underline{J}(w) = \int_{-\infty}^{\bar{F}^{-1}(\bar{q}(w))} F(x)\,\mathrm{d}x.$$

By comparing these two equations, we can conclude that $q^*(w) \leq \bar{q}(w)$ because $J^*(w) \geq \underline{J}(w)$ for $w \geq 1$.

□

A.2.5 Proof of Lemma 4

It follows from Corollary 13 that

$$\bar{G}_i^*(w) = f_w(\beta_i^*(w), q_i^*(w)),\ i = 1, 2 \ \text{ and } \ w \geq 1.$$

Applying the mean value theorem for multivariable functions to $f_w(\cdot)$, the following holds: For $w \geq 1$,

$$\begin{aligned}\delta_{\bar{G}}(w) =& f_w((\beta_1^*(w), q_1^*(w))) - f_w((\beta_2^*(w), q_2^*(w))) \\ =& \frac{\partial f_w(\tilde{e}(w))}{\partial q}\delta_q(w) + \frac{\partial f_w(\tilde{e}(w))}{\partial \beta}\delta_\beta(w),\end{aligned} \tag{A.77}$$

where $\tilde{e}(w) = C(w)e_1^*(w) + (1 - C(w))e_2^*(w)$ for some $C(w) \in (0, 1)$. It follows from Lemma 42 that $e_1^*(w), e_2^*(w) \in \mathcal{Z}_1(w) \times \mathcal{Z}_2(w)$. Since $\mathcal{Z}_1(w) \times \mathcal{Z}_2(w)$ is convex, $\tilde{e}(w) \in \mathcal{Z}_1(w) \times \mathcal{Z}_2(w)$ for all $w \geq 1$. It follows from Lemma 40 that

$$\lim_{w\to\infty} \frac{\partial f_w(\tilde{e}(w))}{\partial q} = 0 \quad \text{and} \quad \lim_{w\to\infty} \frac{\partial f_w(\tilde{e}(w))}{\partial \beta} = 0.$$

Thus, we conclude that for any $\epsilon > 0$, there exists a nonnegative constant w_1 such that the following inequalities are satisfied:

$$\left|\frac{\partial f_w(\tilde{e}(w))}{\partial q}\right| \leq \epsilon \quad \text{and} \quad \left|\frac{\partial f_w(\tilde{e}(w))}{\partial \beta}\right| \leq \epsilon, \ w \geq w_1.$$

Substituting the two inequalities into equation (A.77), we obtain that

$$|\delta_{\bar{G}}(w)| \leq \epsilon(|\delta_\beta(w)| + |\delta_q(w)|), \ w \geq w_1.$$

□

A.3 The Roadmap for the Proof of Uniqueness

This appendix provides a detailed roadmap of the uniqueness proof (Proposition 9). The proof is done by contradiction. In what follows, we first provide an overview of the key steps that lead to the contradiction using various auxiliary lemmas (see Figure B.1). We then summarize the key steps to proving Lemma 4 provided in Appendix A.2, which is an

important technical lemma for the uniqueness proof. Two auxiliary functions, denoted by $h(\cdot)$ and $f_w(\cdot)$, and several lemmas in Appendix A.2 facilitate the proof of Lemma 4. Figure B.2 provides a diagram to show how the lemmas in Appendix A.2 are used to prove Lemma 4.

The proof (of uniqueness) by contradiction proceeds as follows: Suppose that there are two different equilibria and define their difference as (δ_β, δ_q). The contradiction is built on the limiting properties of the difference $(\delta_\beta(w), \delta_q(w))$ as w tends to infinity. Figure B.1 shows how the the contradiction is constructed.

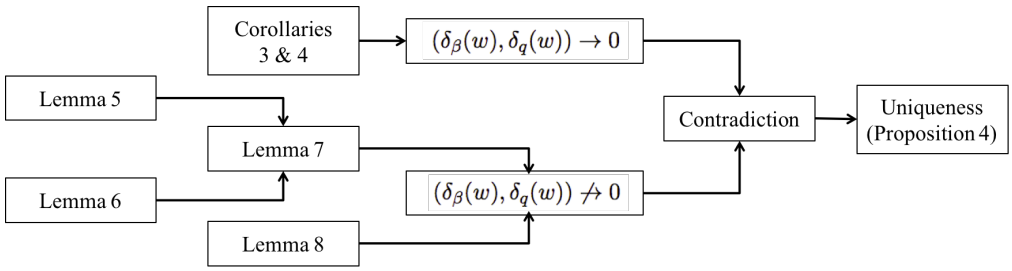


Figure A.1: The logic flow for proving uniqueness of the equilibrium

On the one hand, Corollaries 4-5 provide the limits of equilibrium quantities (in any potential equilibrium). It is immediate from these two corollaries that the difference of the equilibrium quantities (of two different equilibria) vanishes as w goes to infinity, i.e.

$$(\delta_\beta(w), \delta_q(w)) \to 0 \quad \text{as} \quad w \to \infty.$$

On the other hand, Lemmas 6 and 7 show that this convergence cannot hold. Lemma 6 shows that the difference of the equilibrium quantities $(\delta_\beta(w), \delta_q(w)))$, $w \geq 0$ are characterized by a dynamical system. To be more specific, the function characterizing the evolution of this dynamical system has two parts: A constant matrix A with a special structure and a matrix function $B(\cdot)$. In addition, the perturbation function $B(w)$ vanishes as w goes to infinity, i.e. $|||B(w)|||_\infty \to 0$ as $w \to \infty$. This property is proved with the help of the technical Lemma 4. Then Lemma 7 shows that such the dynamical system given in Lemma

6 cannot converge to zero. Combining Lemmas 6 and 7, we conclude that the difference $(\delta_\beta(w), \delta_q(w)))$ cannot converge to zero, which leads to the contradiction.

The rest of this section summarizes the critical steps in Appendix A.2 to prove the technical Lemma 4, which characterizes $\delta_{\bar{G}}(w)$ in terms of $\delta_\beta(w)$ and $\delta_q(w)$ for $w \geq 0$ using functions $g_1(\cdot)$ and $g_2(\cdot)$. Figure B.2 illustrates how various lemmas are used (and relate to one another) to prove Lemma 4. To be specific, Appendices A.2.1-A.2.3 construct two

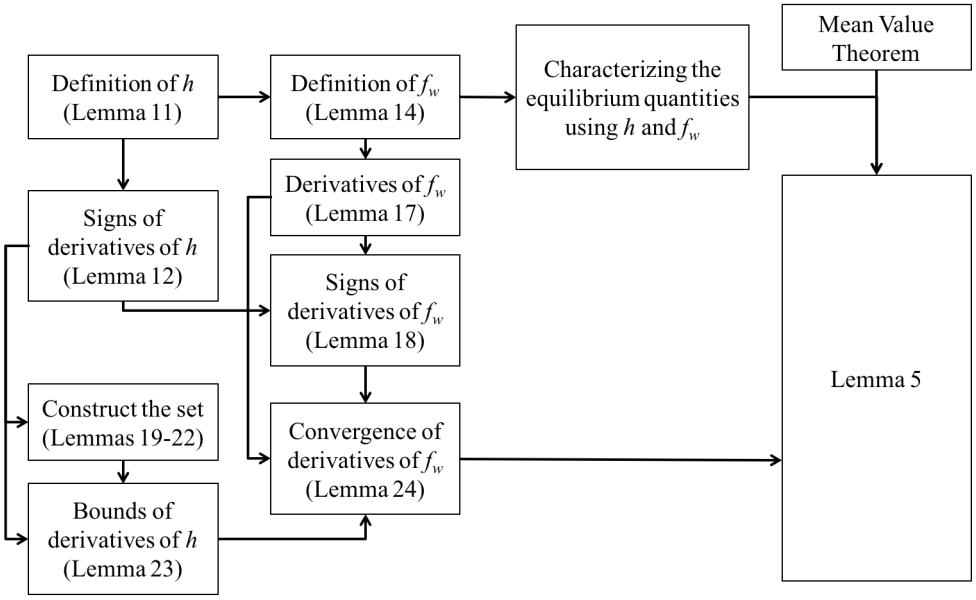


Figure A.2: The logic flow for proving Lemma 4

auxiliary functions $h(\cdot)$ and $f_w(\cdot)$ and provide various properties of these two functions. Appendix A.2.4 shows the characterization of the equilibrium quantities using the auxiliary functions $h(\cdot)$ and $f_w(\cdot)$. Thus, the properties of the auxiliary functions provided in Appendices A.2.1-A.2.3 are applicable to the equilibrium quantities. Appendix A.2.5 proves Lemma 4.

The proof of Lemma 4 in Appendix A.2.5 includes two parts. The first part constructs the functions $g_1(\cdot)$ and $g_2(\cdot)$ in two steps. In the first step, we use the auxiliary function $f_w(\cdot)$ defined in Appendix A.2.1 to characterize $\bar{G}(w)$ in terms of $\beta(w)$ and $q(w)$, i.e. $\bar{G} = f_w(\beta(w), q(w))$ for $w \geq 0$; see Corollary 13 in Appendix A.2.4. In the second step, we apply the mean-value theorem and construct the functions g_1 and g_2 using the partial derivatives

of the function $f_w(\cdot)$; see equation (A.77).

The second part of the proof of Lemma 4 shows that the two functions $g_1(w)$ and $g_2(w)$ converge to zero as $w \to \infty$. To show this, it is sufficient to show that the supremum norm of the partial derivatives of the function $f_w(\cdot)$ converge to zero as $w \to \infty$; see equation (A.77). However, this statement is not true in general, but is valid if we restrict the arguments of the function $f_w(\cdot)$ to be in the set $\mathcal{L}_1(w) \times \mathcal{L}_2(w)$ defined in Appendix A.2.3. The convergence result of the partial derivatives of the function $f_w(\cdot)$ restricted in the set $\mathcal{L}_1(w) \times \mathcal{L}_2(w)$ is given by Lemma 40. In addition, Lemma 42 ensures that the equilibrium quantities lie in the $\mathcal{L}_1(w) \times \mathcal{L}_2(w)$. Thus, applying Lemma 40 completes the second part of the proof of Lemma 4.

Appendices A.2.1-A.2.3 are dedicated to proving Lemma 40 using the auxiliary functions $h(\cdot)$ and $f_w(\cdot)$. To be specific, Appendix A.2.1 defines the auxiliary functions $h(\cdot)$ and $f_w(\cdot)$. Appendix A.2.2 provides the recursive equations to characterize the partial derivatives of the function $f_w(\cdot)$ and the signs of the partial derivatives. Appendix A.2.3 constructs the restricted set $\mathcal{L}_1(w) \times \mathcal{L}_2(w)$ and proves the convergence of the partial derivatives of the function $f_w(\cdot)$ in the restricted set.

The auxiliary function $f_w(\beta, q)$ is defined implicitly though the following equations (A.42)-(A.44), which are rewritten for convenience as follows:

$$
\begin{aligned}
&z(w) = (\beta, q, f_w(\beta, q)),\\
&z(k-1) = h(z(k)) \quad \text{for} \quad k = w, \ldots, 2,\\
&z_2(1) = 1 - z_3(1).
\end{aligned}
$$

Lemma 30 ensures that the function $f_w(\beta, q)$ is well-defined. In order to make sense of this definition of the implicit function $f_w(\cdot)$, the auxiliary function $h(\cdot)$ needs to be introduced. The function $h(\cdot)$ is constructed such that it characterizes the time-reversed evolution of the equilibrium quantities; see Lemma 41 in Appendix A.2.4. This immediately leads to the

observation that if we substitute the equilibrium quantities at time w into $z(w)$ in equation (A.42), i.e. $z(w) = (\beta^*(w), q^*(w), \bar{G}^*(w))$, then the values of $z(k)$ (for $k = w-1, \ldots, 1$) in equation (A.43) equal to the equilibrium quantities as well, i.e.

$$z(k) = (\beta^*(k), q^*(k), \bar{G}^*(k)) \quad \text{for} \quad k = w-1, \ldots, 1.$$

In addition, the condition in equation (A.44) is automatically satisfied by the definition of $\bar{G}$ in equation (3.1). Therefore, the function $f_w(\cdot)$ is the implicit function that characterizes the equilibrium quantity $\bar{G}^*(w)$ in terms of $\beta^*(w), q^*(w)$, i.e. $\bar{G}^*(w) = f_w(\beta^*(w), q^*(w))$, $w \geq 0$; see Corollary 13 in Appendix A.2.4.

Lemma 40 provides the convergence property of the partial derivates of the implicit function $f_w(\cdot)$. In order to prove this lemma, we first provide a recursive characterization of the partial derivates of the implicit function $f_w(\cdot)$. Since the function $f_w(\cdot)$ is defined implicitly by using the function $h(\cdot)$ recursively, the partial derivates of the implicit function $f_w(\cdot)$ are characterized using the partial derivatives of the function $h(\cdot)$ (provided in Lemma 28) recursively; see Lemma 33. By analyzing the the partial derivatives of the functions $f_w(\cdot)$ and $h(\cdot)$ (provided in Lemmas 28 and 33), we provide useful properties of the partial derivatives. These properties eventually leads to the convergence property in Lemma 40; see Figure B.2.

We end this section by providing a comment on Lemma 39. Lemma 39 provides critical bounds of the partial derivatives of the function $h(\cdot)$ to prove Lemma 40. However, these bounds only hold after we restrict the arguments of the function $h(\cdot)$ to the set $\mathcal{L}_1(w) \times \mathcal{L}_2(w)$. The set $\mathcal{L}_1(w) \times \mathcal{L}_2(w)$ is carefully constructed to satisfied two conditions. First, the set is narrow enough such that the bounds in Lemma 39 hold. Second, the set is wide enough to ensure that the equilibrium quantities lie in the set; see Lemma 42.

APPENDIX B

APPENDIX OF CHAPTER 4

B.1 Proofs of Lemmas, Propositions, Corollary and Theorems in the paper

B.1.1 Formal Derivations of Various Approximations in the Heavy Traffic Limit

Approximation of the abandonment process $\hat{R}_k^n(\cdot)$. In our model, a customer in the n^{th} system abandons only at times $t \in \{j/2^n : j = 1, \ldots\}$, while a customer in Kim and Ward [69]'s model can abandon at any time in $[0, \infty)$. Equation (4.20) is an analogue of equation (10) in Kim and Ward [69] for discrete-time abandonments. We follow Kim and Ward [69] to derive equation (4.20).

It follows from Reed and Ward [93] that the scaled virtual offered waiting time $\hat{V}_k^n(t)$ is approximately $\hat{Q}_k^n(t)/\lambda_k$. Thus, the virtual offered waiting time $V_k^n(t)$ is approximately $\hat{Q}_k^n(t)/(\sqrt{2^n}\lambda_k)$. Let $G_k^n(w)$ denote the probability that a class k customer abandons within waiting for w units of time. Thus, $G_k^n(w)$ is given as follows: For $w \in \{j/2^n : j = 1, 2, \ldots\}$,

$$G_k^n(w) = 1 - \prod_{i=1}^{w} \left(1 - q_k^n\left(\frac{i}{2^n}\right)\right). \tag{B.1}$$

A class-k customer arriving at time t sees VOWT $\hat{Q}_k^n(t)/\sqrt{2^n}\lambda_k$ and abandons with probability $G_k^n(\hat{Q}_k^n(t)/\sqrt{2^n}\lambda_k)$. Therefore, the abandonment rate at time t can be approximated by

$$G_k^n\left(\frac{\hat{Q}_k^n(t)}{\sqrt{2^n}\lambda_k}\right) \mathrm{d}A_k^n(t).$$

Thus, the abandonment process is approximated by

$$R_k^n(t) \approx \int_0^t G_k^n\left(\frac{\hat{Q}_k^n(s)}{\sqrt{2^n}\lambda_k}\right) \mathrm{d}A_k^n(s).$$

By substituting $\hat{R}_k^n(t) = R_k^n(t)/\sqrt{2^n}$ into this equation, we obtain that (for $t \geq 0$)

$$\begin{aligned}\hat{R}_k^n(t) &\approx \int_0^t \sqrt{2^n}\lambda_k G_k^n\left(\frac{\hat{Q}_k^n(s)}{\sqrt{2^n}\lambda_k}\right) \mathrm{d}\left(\frac{A_k^n(s)}{\lambda_k 2^n}\right) \\ &\approx \int_0^t \sqrt{2^n}\lambda_k G_k^n\left(\frac{\hat{Q}_k^n(s)}{\sqrt{2^n}\lambda_k}\right) \mathrm{d}s,\end{aligned} \tag{B.2}$$

where the approximation in the second line follows because $A_k^n(t)/(2^n\lambda_k) \Rightarrow t$ as $n \to \infty$. Moreover, it follows from (B.1) that

$$\begin{aligned}\sqrt{2^n}G_k^n\left(\frac{\hat{Q}_k^n(s)}{\sqrt{2^n}\lambda_k}\right) &= \sqrt{2^n}\left[1 - \prod_{i=1}^{\lfloor 2^n\hat{Q}_k^n(s)/(\sqrt{2^n}\lambda_k)\rfloor}\left(1 - q_k^n\left(\frac{i}{2^n}\right)\right)\right] \\ &\approx \sqrt{2^n}\left[1 - \prod_{i=1}^{\lfloor \sqrt{2^n}\hat{Q}_k^n(s)/\lambda_k\rfloor}\exp\left(-q_k^n\left(\frac{i}{2^n}\right)\right)\right] \\ &= \sqrt{2^n}\left[1 - \exp\left(-\sum_{i=1}^{\lfloor \sqrt{2^n}\hat{Q}_k^n(s)/\lambda_k\rfloor} q_k^n\left(\frac{i}{2^n}\right)\right)\right].\end{aligned} \tag{B.3}$$

The approximation in the second line of (B.3) follows from $\exp(-x) \approx 1 - x$ for x close to zero. Note also that

$$\begin{aligned}q_k^n\left(\frac{i}{2^n}\right) &= \int_i^{i+1} q_k^n\left(\frac{\lfloor x\rfloor}{2^n}\right) \mathrm{d}x \\ &= \int_i^{i+1} \frac{1}{\sqrt{2^n}}\hat{q}_k^n\left(\frac{x}{\sqrt{2^n}}\right) \mathrm{d}\left(\frac{x}{\sqrt{2^n}}\right),\end{aligned}$$

where the equality in the second line follows from (4.19) that for $x \geq 0$,

$$\hat{q}_k^n\left(\frac{x}{\sqrt{2^n}}\right) = 2^n q_k^n\left(\frac{\lfloor x \rfloor}{2^n}\right).$$

Substituting this into (B.3), we have that

$$\begin{aligned}
\sqrt{2^n} G_k^n\left(\frac{\hat{Q}_k^n(s)}{\sqrt{2^n}\lambda_k}\right) &\approx \sqrt{2^n}\left[1 - \exp\left(-\sum_{i=1}^{\lfloor\sqrt{2^n}\hat{Q}_k^n(s)/\lambda_k\rfloor} \int_i^{i+1} \frac{1}{\sqrt{2^n}} \hat{q}_k^n\left(\frac{x}{\sqrt{2^n}}\right) \mathrm{d}\left(\frac{x}{\sqrt{2^n}}\right)\right)\right] \\
&= \sqrt{2^n}\left[1 - \exp\left(-\int_1^{\lfloor\sqrt{2^n}\hat{Q}_k^n(s)+1)/\lambda_k\rfloor} \frac{1}{\sqrt{2^n}} \hat{q}_k^n\left(\frac{x}{\sqrt{2^n}}\right) \mathrm{d}\left(\frac{x}{\sqrt{2^n}}\right)\right)\right] \\
&\approx \sqrt{2^n}\left[1 - \exp\left(-\int_0^{\lfloor\sqrt{2^n}\hat{Q}_k^n(s))/\lambda_k\rfloor} \frac{1}{\sqrt{2^n}} \hat{q}_k^n\left(\frac{x}{\sqrt{2^n}}\right) \mathrm{d}\left(\frac{x}{\sqrt{2^n}}\right)\right)\right] \\
&\approx \sqrt{2^n}\left[1 - \exp\left(-\int_0^{\hat{Q}_k^n(s))/\lambda_k} \frac{1}{\sqrt{2^n}} \hat{q}_k^n(x)\, \mathrm{d}x\right)\right].
\end{aligned}$$

Substituting the approximation $1 - \exp(-x) \approx x$ into this equation, we obtain that

$$\sqrt{2^n} G_k^n\left(\frac{\hat{Q}_k^n(s)}{\sqrt{2^n}\lambda_k}\right) \approx \sqrt{2^n} \int_0^{\hat{Q}_k^n(s))/\lambda_k} \frac{1}{\sqrt{2^n}} \hat{q}_k^n(x)\, \mathrm{d}x = \int_0^{\hat{Q}_k^n(s))/\lambda_k} \hat{q}_k^n(x)\, \mathrm{d}x.$$

Substituting the approximation into (B.2), we conclude that

$$\hat{R}_k^n(t) \approx \lambda_k \int_0^t \int_0^{\hat{Q}_k^n(s))/\lambda_k} \hat{q}_k^n(x)\, \mathrm{d}x\, \mathrm{d}s,\ t \geq 0.$$

□

Characterization of $\hat{\beta}_k^n(\cdot)$. We derive equation (4.24) that characterizes $\hat{\beta}_k^n(\cdot)$, which is the hazard rate (per time unit) of (the steady-state distribution) of $\hat{V}_k^n(\cdot)$.

Since the VOWT V_k^n lives on $\{j/2^n : j \geq 1\}$, its scaled version $\hat{V}_k^n = \sqrt{2^n} V_k^n$ lives on $\{j/\sqrt{2^n} : j \geq 1\}$. We first define the hazard rate $\hat{\beta}_k^n(w)$ for $w \in \{j/\sqrt{2^n} : j \geq 1\}$. Note $\beta_k^n(\cdot)$ is the per period hazard rate of (the steady state distribution of) the VOWT $V_k^n(\cdot)$.

Dividing $\beta_k^n(\cdot)$ by the period length $1/2^n$, we obtain the hazard rate (per time unit) of $V_k^n(\cdot)$, which is $2^n\beta_k^n(\cdot)$. Thus, the hazard rate of the scaled VOWT $\hat{V}_k^n(\cdot)$ is given as follows: For $w \in \left\{j/\sqrt{2^n} : j \geq 1\right\}$.

$$\begin{aligned}
\hat{\beta}_k^n(w) &= \lim_{h\to 0} \frac{\mathbb{P}(w \leq \hat{V}_k^n(t) \leq w+h)}{\mathbb{P}(\hat{V}_k^n(t) \geq w)h} \\
&= \lim_{h\to 0} \frac{\mathbb{P}(w/\sqrt{2^n} \leq V_k^n(t) \leq (w+h)/\sqrt{2^n})}{\mathbb{P}(V_k^n(t) \geq w/\sqrt{2^n})h} \\
&= \frac{1}{\sqrt{2^n}} \lim_{h\to 0} \frac{\mathbb{P}(w/\sqrt{2^n} \leq V_k^n(t) \leq (w+h)/\sqrt{2^n})}{\mathbb{P}(V_k^n(t) \geq w/\sqrt{2^n})(h/\sqrt{2^n})} \\
&= \frac{1}{\sqrt{2^n}} \left(2^n \beta_k^n \left(\frac{w}{\sqrt{2^n}}\right)\right) = \sqrt{2^n}\beta_k^n \left(\frac{w}{\sqrt{2^n}}\right).
\end{aligned}$$

Extending the definition of $\hat{\beta}_k^n(w)$ for $w \geq 0$, we obtain that

$$\hat{\beta}_k^n(w) = \sqrt{2^n}\beta_k^n \left(\frac{\lfloor \sqrt{2^n} w \rfloor}{2^n}\right), \ w \geq 0.$$

□

The scaled abandonment probability $\hat{q}_k(\cdot)$ under Assumption 6. Fix a class k, for $k = 1, \ldots, K$. It follows from (2.10) that the abandonment probability of class k customers in the n^{th} system satisfies the following equation: For $w \in \{j/2^n : j = 1, 2, \ldots\}$,

$$q_k^n(w) = \bar{F}_k^n \left(-c_k^n + \alpha^n \left[\beta_k^n(w) r_k^n + (1 - \beta_k^n(w)) J_k^n \left(w + \frac{1}{2^n}\right)\right]\right). \tag{B.4}$$

where $\alpha^n = (\alpha)^{1/2^n}$ is the discount factor for one period in the n^{th} system. Substituting (B.4) into (4.19), we obtain the characterization of the scaled abandonment rate $\hat{q}_k^n(w)$ (for

$w \geq 0$) as follows:

$$\begin{aligned}
&\hat{q}_k^n(w) \\
&= 2^n \bar{F}_k^n \left(-c_k^n + \alpha^n \left[\beta_k^n \left(\frac{\lfloor \sqrt{2^n} w \rfloor}{2^n} \right) r_k^n + \left(1 - \beta_k^n \left(\frac{\lfloor \sqrt{2^n} w \rfloor}{2^n} \right) \right) J_k^n \left(\frac{\lfloor \sqrt{2^n} w \rfloor}{2^n} + \frac{1}{2^n} \right) \right] \right) \\
&= 2^n \bar{F}_k^n \left(-\frac{c_k}{\sqrt{2^n}} + \alpha^n \left[\frac{\hat{\beta}_k^n(w)}{\sqrt{2^n}} r_k + \left(1 - \frac{\hat{\beta}_k^n(w)}{\sqrt{2^n}} \right) J_k^n \left(\frac{\lfloor \sqrt{2^n}(w + 1/\sqrt{2^n}) \rfloor}{2^n} \right) \right] \right).
\end{aligned} \tag{B.5}$$

The equality in the third line follows from (4.24) and the scaling $c_k^n = c_k/\sqrt{2^n}$ and $r_k^n = r_k$. To facilitate the analysis to follow, recall that we scale the time of the expected discounted utility of waiting as follows:

$$\hat{J}_k^n(w) = J_k^n \left(\frac{\lfloor \sqrt{2^n} w \rfloor}{2^n} \right), \ w \geq 0. \tag{B.6}$$

Substituting $\hat{J}_k^n(w)$ into equation (B.5), we have the following: For $w \geq 0$,

$$\hat{q}_k^n(w) = 2^n \bar{F}_k^n \left(\hat{v}_k^n(w) \right), \tag{B.7}$$

where

$$\begin{aligned}
&\hat{v}_k^n(w) \\
&= -\frac{c_k}{\sqrt{2^n}} + \alpha^n \left[\frac{\hat{\beta}_k^n(w)}{\sqrt{2^n}} r_k + \left(1 - \frac{\hat{\beta}_k^n(w)}{\sqrt{2^n}} \right) \hat{J}_k^n \left(w + \frac{1}{\sqrt{2^n}} \right) \right] \\
&= \hat{J}_k^n \left(w + \frac{1}{\sqrt{2^n}} \right) + \frac{1}{\sqrt{2^n}} \left[c_k + \hat{\beta}_k^n(w) \left(r_k - \hat{J}_k^n \left(w + \frac{1}{\sqrt{2^n}} \right) \right) \right] + (\alpha^n - 1) \hat{J}_k^n \left(w + \frac{1}{\sqrt{2^n}} \right).
\end{aligned} \tag{B.8}$$

It follows from (B.7)-(B.8) that the abandonment rate $\hat{q}_k^n(w)$ is fully characterized by the expected discounted utility $\hat{J}_k^n(w)$ for $w \geq 0$. Thus, we first derive the differential equation that characterizes the limit of $\hat{J}_k^n(\cdot)$ and then derive the limit of scaled abandonment

probability $\hat{q}_k^n(\cdot)$.

Recall from equation (2.8) that the expected discounted utility $J_k^n(\cdot)$ satisfies the following Bellman equation: For $w \in \{j/2^n : j = 1, 2, \ldots\}$,

$$\begin{aligned} J_k^n(w) &= \mathbb{E}_{\varepsilon_k^n}\left[\max\{\varepsilon_k^n(1), -c_k^n + \alpha^n\left[\beta_k^n(w) r_k^n + (1-\beta_k^n(w)) J_k^n\left(w + \frac{1}{2^n}\right)\right] + \varepsilon_k^n(0)\}\right] \\ &= \mathbb{E}_{\varepsilon_k^n}\left[-c_k^n + \alpha^n\left[\beta_k^n(w) r_k^n + (1-\beta_k^n(w)) J_k^n\left(w + \frac{1}{2^n}\right)\right] - (\varepsilon_k^n(1) - \varepsilon_k^n(0))\right]^+. \end{aligned}$$

Combining this with (B.6) and (B.8) yields the following: For $w \geq 0$,

$$\begin{aligned} \hat{J}_k^n(w) &= J_k^n\left(\frac{\lfloor\sqrt{2^n} w\rfloor}{2^n}\right) \\ &= \mathbb{E}_{\varepsilon_k^n}\left[\hat{v}_k^n(w) - (\varepsilon_k^n(1) - \varepsilon_k^n(0))\right]^+ \\ &= \int_{-\infty}^{\hat{v}_k^n(w)} (\hat{v}_k^n(w) - x)\, \mathrm{d}F_k^n(x) \\ &= \int_{-\infty}^{\hat{v}_k^n(w)} F_k^n(x)\, \mathrm{d}x. \end{aligned} \tag{B.9}$$

The equality in the last line follows from integration by parts. Recall from Assumption 6 that the cumulative distribution function F_k^n of $\varepsilon_k^n(1) - \varepsilon_k^n(0)$ is given as follows:

$$F_k^n(x) = \begin{cases} \frac{x_n}{2(-x)^\delta}, & x \leq -(x_n)^{1/\delta}, \\ 1/2, & -(x_n)^{1/\delta} < x < (x_n)^{1/\delta}, \\ 1 - \frac{x_n}{2x^\delta}, & x \geq (x_n)^{1/\delta}, \end{cases}$$

where $x_n = 1/2^n$ and $\delta > 1$. Substituting this cdf into (B.9), we have the following three

cases:

$$\hat{J}_k^n(w) = \begin{cases} \dfrac{x_n}{2(\delta-1)(-\hat{v}_k^n(w))^{\delta-1}}, & \text{if } \hat{v}_k^n(w) < -(x_n)^{1/\delta}, \\ \dfrac{1}{2}\hat{v}_k^n(w) + \dfrac{\delta(x_n)^{1/\delta}}{2(\delta-1)}, & \text{if } -(x_n)^{1/\delta} \le \hat{v}_k^n(w) \le (x_n)^{1/\delta}, \\ \hat{v}_k^n(w) + \dfrac{x_n}{2(\delta-1)(\hat{v}_k^n(w))^{\delta-1}}, & \text{if } \hat{v}_k^n(w) > (x_n)^{1/\delta}. \end{cases} \tag{B.10}$$

Note that $\bar{F}_k^n(\hat{v}^n(w)) \geq 1/2$ if $\hat{v}_k^n(w) \leq (x_n)^{1/\delta}$. In addition, it follows from equation (B.7) that

$$\hat{q}_k^n(w) = 2^n \bar{F}_k^n(\hat{v}_k^n(w)).$$

Under the first two cases of (B.10) where $\hat{v}_k^n(w) \leq (x_n)^{1/\delta}$, $\hat{q}_k^n(w) \to \infty$ as $n \to \infty$. Since we focus on the case when $\hat{q}_k^n(\cdot)$ converges to a bounded function, we only consider the third case in equation (B.10) (for n large). To be specific, we first assume that $\hat{v}_n(w) > (x_n)^{1/\delta}$ for $w \geq 0$ and n large enough and derive the limit of $\hat{J}_k^n(\cdot)$ as $n \to \infty$. Then we show that the solutions obtained satisfy this assumption. Substituting (B.8) in the third case of equation (B.10), we obtain the following equation: For $w \geq 0$,

$$\hat{J}_k^n(w) = -\frac{c_k}{\sqrt{2^n}} + \alpha^n \left[\frac{\hat{\beta}_k^n(w)}{\sqrt{2^n}} r_k + \left(1 - \frac{\hat{\beta}_k^n(w)}{\sqrt{2^n}}\right) \hat{J}_k^n\left(w + \frac{1}{\sqrt{2^n}}\right) \right] + \frac{x_n}{2(\delta-1)} (v_k^n(w))^{-\delta+1}.$$

Multiplying both sides by $\sqrt{2^n}$ and rearranging the terms, we obtain the following equation:

$$\begin{aligned} &\sqrt{2^n}\left(\hat{J}_k^n(w) - \hat{J}_k^n\left(w + \frac{1}{\sqrt{2^n}}\right)\right) \\ =& -c_k + \alpha^n \hat{\beta}_k^n(w)\left(r_k - \hat{J}_k^n\left(w + \frac{1}{\sqrt{2^n}}\right)\right) \\ &+ \frac{\sqrt{2^n} x_n (v_k^n(w))^{-\delta+1}}{2(\delta-1)} + \sqrt{2^n}(\alpha^n - 1)\hat{J}_k^n\left(w + \frac{1}{\sqrt{2^n}}\right). \end{aligned} \tag{B.11}$$

Since the last two terms in the right-hand side of the last line in (B.8) go to zero as $n \to \infty$, the following holds:

$$\hat{v}_k^n(w) - \hat{J}_k^n\left(w + \frac{1}{\sqrt{2^n}}\right) \to 0 \text{ as } n \to \infty. \tag{B.12}$$

Thus, $\hat{v}_k^n(w)$ has the same order of magnitude as $\hat{J}_k^n(w + 1/\sqrt{2^n})$, which is of order 1. Since $x_n = 1/2^n$, the third term in the last line of (B.11) is of order $1/\sqrt{2^n}$ and goes to zero as $n \to \infty$. In addition, the last term in the last line goes to zero as $n \to \infty$ as well because

$$\sqrt{2^n}(\alpha^n - 1) = \sqrt{2^n}[(\alpha)^{1/2^n} - 1] \to 0 \text{ as } n \to \infty.$$

Since we assume that $\hat{\beta}_k^n(w) \to \hat{\beta}_k(w)$ for $w \geq 0$ and $\alpha^n \to 1$ as $n \to \infty$, it follows from (B.11) that

$$\sqrt{2^n}\left(\hat{J}_k^n(w) - \hat{J}_k^n\left(w + \frac{1}{\sqrt{2^n}}\right)\right) = -c_k + \hat{\beta}_k(w)\left(r_k - \hat{J}_k^n\left(w + \frac{1}{\sqrt{2^n}}\right)\right) + o(1).$$

This equation is the linear interpolation (with step size $\Delta t = 1/\sqrt{2^n}$) of the following ordinary differential equation:

$$-\hat{J}_k'(w) = -c_k + \hat{\beta}_k(w)(r - \hat{J}_k(w)), \ w \geq 0. \tag{B.13}$$

We assume that the solution to (B.13), $\hat{J}_k(w)$, is the limit of $\hat{J}_k^n(w)$ as $n \to \infty$, i.e. $\hat{J}_k(w) = \lim_{n\to\infty} \hat{J}_k^n(w)$ for $w \geq 0$. Since $\hat{\beta}_k(\cdot)$ is given, equation (B.13) is a linear differential equation. The general form of the solution to equation (B.13) is given as follows: For $w \geq 0$,

$$\hat{J}_k(w) = \exp\left(\int_0^w \hat{\beta}_k(s)\mathrm{d}s\right)\left[\int_0^w -\exp\left(\int_0^s -\hat{\beta}_k(u)\mathrm{d}u\right)(-c_k + \hat{\beta}_k(w)r_k)\mathrm{d}s + C_1\right]; \tag{B.14}$$

see Chapter 2.2 of Sanchez [98].

The constant C_1 gives the initial value of $\hat{J}_k(\cdot)$, i.e. $\hat{J}_k(0) = C_1$. It follows from Lemma 2 that $\hat{J}_k^n(w) \in [0, r_k]$ for $w \geq 0$ and $n \geq 1$. Since $\hat{J}_k(w) = \lim_{n\to\infty} \hat{J}_k^n(w)$, we have that

$\hat{J}_k(w) \in [0, r_k]$ for $w \geq 0$. Moreover, it follows from Corollary 7 that any $\hat{\beta}_k(\cdot)$ in equilibrium is bounded away from zero. Thus, we have that

$$\exp\left(\int_0^w \hat{\beta}_k(s)\,\mathrm{d}s\right) \to \infty \quad \text{as} \quad w \to \infty.$$

Since $\hat{J}_k(w) \leq r_k$ for $w \geq 0$, it follows from (B.14) that it must be that

$$\int_0^w -\exp\left(\int_0^s -\hat{\beta}_k(u)\,\mathrm{d}u\right)(-c_k + \hat{\beta}_k(s) r_k)\,\mathrm{d}s + C_1 \to 0 \quad \text{as} \quad w \to \infty. \tag{B.15}$$

In addition, the left-hand side of (B.15) must be non-negative for $w \geq 0$ because $\hat{J}_k(w) \geq 0$. Therefore, the only constant C_1 satisfying (B.15) is

$$C_1 = \int_0^\infty \exp\left(\int_0^s -\hat{\beta}_k(u)\,\mathrm{d}u\right)(-c_k + \hat{\beta}_k(w) r_k)\,\mathrm{d}s.$$

Substituting the value of C_1 into equation (B.14), we obtain that for $w \geq 0$,

$$\begin{aligned}
\hat{J}_k(w) &= \exp\left(\int_0^w \hat{\beta}_k(s)\,\mathrm{d}s\right)\int_w^\infty \exp\left(\int_0^s -\hat{\beta}_k(u)\,\mathrm{d}u\right)(-c_k + \hat{\beta}_k(s) r_k)\,\mathrm{d}s \\
&= \int_w^\infty \exp\left(\int_w^s -\hat{\beta}_k(u)\,\mathrm{d}u\right)(-c_k + \hat{\beta}_k(s) r_k)\,\mathrm{d}s \\
&= r_k - c_k \int_w^\infty \exp\left(\int_w^s -\hat{\beta}_k(u)\,\mathrm{d}u\right)\mathrm{d}s,
\end{aligned}$$

which proves (4.33).

By substituting the lower bound of $\hat{\beta}_k(\cdot)$ in Corollary 7 into $\hat{J}_k(\cdot)$, we obtain the following inequality: For $w \geq 0$,

$$\begin{aligned}
\hat{J}_k(w) &\geq r_k - c_k \int_w^\infty \exp\left(\int_w^s \frac{2\theta\rho_k}{\sigma^2 \sup_{x\geq 0} \gamma_k'(x)}\,\mathrm{d}u\right)\mathrm{d}s \\
&= r_k + \frac{\sigma^2 \sup_{x\geq 0} \gamma_k'(x)}{2\theta\rho_k} c_k > 0,
\end{aligned} \tag{B.16}$$

where the last inequality follows from Assumption 5. This inequality shows that $\hat{J}_k(\cdot)$ is

bounded away from zero (in equilibrium).

We now show that $v_k^n(w) > (x_n)^{1/\delta}$ in equilibrium for $w \geq 0$ and n large enough[1]. It follows from the assumption $\hat{J}_k^n(w) \to \hat{J}_k(w)$ and (B.12) that $v_k^n(w) \to \hat{J}_k(w)$ as $n \to \infty$ for $w \geq 0$. Since $x_n \to 0$ as $n \to \infty$, the following holds: For $w \geq 0$,

$$v_k^n(w) - (x_n)^{1/\delta} \to \hat{J}_k(w) > 0 \quad \text{as} \quad n \to \infty,$$

where the last inequality follows from (B.16) that $\hat{J}_k(w)$ is bounded away from zero in equilibrium. Thus, the inequality $v_k^n(w) > (x_n)^{1/\delta}$ holds for $w \geq 0$ and n large enough.

Substituting the distribution function $F_k^n(\cdot)$ into equation (B.7) and letting n go to infinity, we obtain the function that characterizes the abandonment rate as follows:

$$\hat{q}_k(w) = \lim_{n\to\infty} \hat{q}_k^n(w) = \lim_{n\to\infty} 2^n \bar{F}_k^n(\hat{v}_k^n(w)) = \lim_{n\to\infty} 2^n \frac{x_n}{2(\hat{v}_k^n(w))^\delta} = \frac{1}{2(\hat{J}_k(w))^\delta}, \ w \geq 0.$$

The last equality holds because $x_n = 1/2^n$ and $\hat{v}_k^n(w) \to \hat{J}_k(w)$ as $n \to \infty$ for $w \geq 0$. Since $\hat{J}_k(\cdot)$ is strictly bounded away from zero, the abandonment rate $\hat{q}_k(\cdot)$ is well-defined.

□

B.1.2 Proof of Results in Sections 4.1.2 and 4.1.3

Proof of Corollary 6. Substituting equations (4.27) and (4.30) into $\hat{V}_k(t) = \hat{Q}_k(t)/\lambda_k$ yields the characterization of the VOWT of class k customers as follows:

$$\hat{V}_k(t) = \frac{\gamma_k(\hat{W}(t))}{\rho_k} \quad \text{for} \quad k = 1, \ldots, K. \tag{B.17}$$

1. This is not true in general, but it holds for the equilibrium quantities which are of our interest.

The cdf of the VOWT $\hat{V}_k(\cdot)$ in steady state is obtained by substituting equation (B.17) into the cdf $\Pi(\cdot)$. Thus, the pdf of $\hat{V}_k(w)$ in steady state is given by

$$\begin{aligned}
&\lim_{h\to 0} \frac{\Pi(w \le \hat{V}_k(\infty) \le w+h)}{h} \\
&= \lim_{h\to 0} \frac{\Pi(\gamma^{-1}(\rho_k w) \le \hat{W}(\infty) \le \gamma^{-1}(\rho_k(w+h)))}{h} \\
&= \lim_{h\to 0} \frac{\Pi(\gamma^{-1}(\rho_k w) \le \hat{W}(\infty) \le \gamma^{-1}(\rho_k(w+h)))}{\gamma^{-1}(\rho_k(w+h)) - \gamma^{-1}(\rho_k w)} \frac{\gamma^{-1}(\rho_k(w+h)) - \gamma^{-1}(\rho_k w)}{h} \\
&= \pi(\gamma_k^{-1}(\rho_k w))(\gamma_k^{-1})'(\rho_k w)\rho_k.
\end{aligned}$$

where $\hat{V}_k(\infty)$ and $\hat{W}(\infty)$ are the scaled VOWT of class k customers and the workload in stead state, and $(\gamma_k^{-1})'$ is the derivative of γ_k^{-1} for $k = 1, \ldots, K$. Thus, the hazard rate $\hat{\beta}_k(\cdot)$ of class k is given as follows: (For $k = 1, \ldots, K$ and $w \ge 0$,)

$$\hat{\beta}_k(w) = \frac{\pi(\gamma_k^{-1}(\rho_k w))(\gamma_k^{-1})'(\rho_k w)}{1 - \Pi(\gamma_k^{-1}(\rho_k w))} = \hat{\beta}_W\left(\gamma_k^{-1}(\rho_k w)\right)(\gamma_k^{-1})'(\rho_k w)\,\rho_k.$$

□

Proof of Corollary 7. It follows from equation (4.37) that $\hat{\beta}_W(w) \ge -2\theta/\sigma^2$ for $w \ge 0$. To be more specific, we have that for $w \ge 0$,

$$\begin{aligned}
\hat{\beta}_W(w) &= \left(\int_w^\infty \exp\left[\frac{2}{\sigma^2}\left(\theta(s-w) - \int_w^s H(u)\,\mathrm{d}u\right)\right]\,\mathrm{d}s\right)^{-1} \\
&\ge \left(\int_w^\infty \exp\left[\frac{2}{\sigma^2}\theta(s-w)\right]\,\mathrm{d}s\right)^{-1} = -\frac{2\theta}{\sigma^2},
\end{aligned}$$

where the inequality follows from the fact that $H(w) \ge 0$ for $w \ge 0$. Thus, it follows from equation (4.38) that for $k = 1, \ldots, K$ and $w \ge 0$,

$$\hat{\beta}_k(w) = \hat{\beta}_W(\gamma_k^{-1}(\rho_k w))\frac{\rho_k}{\gamma_k'(\gamma_k^{-1}(\rho_k w))} \ge -\frac{2\theta}{\sigma^2}\frac{\rho_k}{\sup_{t\ge 0}\gamma_k'(t)}.$$

which establishes (4.39). To establish (4.40), we use the following identity:

$$\mathbb{E}V_k(\infty) = \int_0^\infty \mathbb{P}\left(v_k(\infty) > x\right)\,\mathrm{d}x = \int_0^\infty \exp\left(-\int_0^x \hat{\beta}_k(u)\,\mathrm{d}u\right)\,\mathrm{d}x,\ k = 1, \ldots, K,$$

which holds generally for any non-negative random variable (between its expected value and the hazard rate). Substituting (4.39) into this gives (4.40).

□

Proof of Lemma 10. We first prove property (ii), because property (ii) facilities the proof of property (i). If $x(w) \geq -2\theta/\sigma^2$, we conclude from (4.51) that for $w \geq 0$ and $k = 1, \ldots, K$,

$$\begin{aligned}
\tilde{J}_{x,k}(w) &= r_k - c_k \int_w^\infty \exp\left(\int_w^s -x(u)\,\mathrm{d}u\right) \frac{\gamma_k'(s)}{\rho_k}\,\mathrm{d}s \\
&\geq r_k - c_k \sup_{t\geq 0} \frac{\gamma_k'(t)}{\rho_k} \int_w^\infty \exp\left(\frac{2\theta}{\sigma^2}(s-w)\right)\,\mathrm{d}s \\
&= r_k + \frac{c_k\sigma^2}{2\theta\rho_k} \sup_{t\geq 0} \gamma_k'(t),
\end{aligned} \tag{B.18}$$

where the inequality follows from the assumption that $x(w) \geq -2\theta/\sigma^2$. It is also immediate from (4.43) that $\tilde{J}_{x,k}(w) < r_k$ for k, w because $c_k, r_k > 0$ and $\gamma_k'(s) > 0$. Substituting (B.18) and $\tilde{J}_{x,k}(w) \leq r_k$ into equation (4.52) gives the following:

$$\underline{H}(w) \leq H_x(w) = \int_0^w \sum_{k=1}^K \frac{\gamma_k'(s)}{2(\tilde{J}_k(s))^\delta}\,\mathrm{d}s \leq \bar{H}(w),\ w \geq 0.$$

Substituting this into equation (4.53), we obtain the following inequalities: For $w \geq 0$,

$$\begin{aligned}
\hat{\beta}_x(w) &\leq \left(\int_w^\infty \exp\left[\frac{2}{\sigma^2}\left(\theta(s-w) - \int_w^s \bar{H}(u)\,\mathrm{d}u\right)\right]\,\mathrm{d}s\right)^{-1} = \bar{\beta}(w), \\
\hat{\beta}_x(w) &\geq \left(\int_w^\infty \exp\left[\frac{2}{\sigma^2}\left(\theta(s-w) - \int_w^s \underline{H}(u)\,\mathrm{d}u\right)\right]\,\mathrm{d}s\right)^{-1} = \underline{\beta}(w).
\end{aligned}$$

Thus, $\hat{\beta}_x(w) = \Psi \circ x(w) \in [\underline{\beta}(w), \bar{\beta}(w)]$ for $w \geq 0$.

Next, we prove property (i). Let $x_n \in C[0,\infty)$ for $n \geq 1$ be a sequence of functions such that $x_n(w) \geq -2\theta/\sigma^2$ for al $w \geq 0$. Suppose that $x_n \to x$ uniformly over compact subsets of $[0,\infty)$. It suffices to show that $\Psi \circ x_n \to \Psi \circ x$ uniformly on $[0,M]$ for $M \geq 0$. To simplify the notation, let

$$\tilde{J}_n = \tilde{J}_{x_n},\ H_n = H_{x_n} \ \text{ and } \ \hat{\beta}_n = \hat{\beta}_{x_n}.$$

We prove property (i) by proving the following:

$$\tilde{J}_n \to \tilde{J}_x,\ H_n \to H_x \ \text{ and } \ \hat{\beta}_n \to \hat{\beta}_x \ \text{ uniformly on } \ [0,M].$$

We first show that $\tilde{J}_n \to \tilde{J}_x$ uniformly on $[0,M]$. Recall from Assumption 4 that $\gamma_k'(w) \to \bar{\gamma}_k$ as $w \to \infty$ for $k = 1,\ldots,K$. Thus, $\gamma_k'(\cdot)$ is bounded for $k = 1,\ldots,K$. Let M_1 be one such that

$$\frac{\gamma_k'(w)}{\rho_k} \leq M_1,\ w \geq 0 \ \text{ and } \ k = 1,\ldots,K. \tag{B.19}$$

Also for any $\epsilon > 0$, there exists $w_1 \geq M$ such that

$$-c_k \exp\left(\frac{2\theta}{\sigma^2}(w_1 - M)\right)\frac{M_1\sigma^2}{2\theta} < \frac{\epsilon}{4},\ k = 1,\ldots,K.$$

Therefore, the following holds: For $w \in [0,M]$, $n \geq 1$ and $k = 1,\ldots,K$,

$$\begin{aligned}
&c_k \int_{w_1}^{\infty} \exp\left(\int_w^s -x_n(u)\ \mathrm{d}u\right)\frac{\gamma_k'(s)}{\rho_k}\,\mathrm{d}s \\
&\leq c_k \int_{w_1}^{\infty} \exp\left(\frac{2\theta}{\sigma^2}(s-w)\right) M_1\,\mathrm{d}s \\
&\leq c_k \int_{w_1}^{\infty} \exp\left(\frac{2\theta}{\sigma^2}(s-M)\right) M_1\,\mathrm{d}s \\
&= c_k \exp\left(\frac{2\theta}{\sigma^2}(w_1-M)\right)\int_0^{\infty} \exp\left(\frac{2\theta}{\sigma^2}s\right) M_1\,\mathrm{d}s \\
&= -c_k \exp\left(\frac{2\theta}{\sigma^2}(w_1-M)\right)\frac{M_1\sigma^2}{2\theta} < \frac{\epsilon}{4}.
\end{aligned} \tag{B.20}$$

The first inequality follows from the assumption that $x_n(w) \geq -2\theta/\sigma^2$ and equation (B.19)

for $k = 1, \ldots, K$ and $w \geq 0$. The second inequality follows from that $w \in [0, M]$. Similarly, the following holds because $x(w) \geq -2\theta/\sigma^2$ for $w \geq 0$: For $w \in [0, M]$ and $k = 1, \ldots, K$,

$$c_k \int_{w_1}^{\infty} \exp\left(\int_w^s -x\left(u\right)\, \mathrm{d}u\right) \frac{\gamma_k'(s)}{\rho_k}\, \mathrm{d}s < \frac{\epsilon}{4}. \tag{B.21}$$

Note that for any positive real numbers a, b, we have that $|a - b| < a + b$. Combining this with (B.20)-(B.21), we deduce that for $w \in [0, M]$, $n \geq 1$ and $k = 1, \ldots, K$,

$$c_k \int_{w_1}^{\infty} \left|\exp\left(\int_w^s -x_n\left(u\right)\, \mathrm{d}u\right) - \exp\left(\int_w^s -x\left(u\right)\, \mathrm{d}u\right)\right| \frac{\gamma_k'(s)}{\rho_k}\, \mathrm{d}s < \frac{\epsilon}{2}. \tag{B.22}$$

In addition, define

$$c_{M,n} = \sup_{t_1, t_2 \in [0, w_1]} \left|\exp\left(\int_{t_1}^{t_2} -x_n\left(u\right)\, \mathrm{d}u\right) - \exp\left(\int_{t_1}^{t_2} -x\left(u\right)\, \mathrm{d}u\right)\right|$$

Since w_1 is a constant that depends on M, $c_{M,n}$ depends on M. Since $x_n \to x$ uniformly on $[0, w_1]$, $c_{M,n} \to 0$ as $n \to \infty$. Thus, there exists n_1 such that for $n > n_1$ and $k = 1, \ldots, K$,

$$c_k \int_0^{w_1} c_{M,n} M_1\, \mathrm{d}s < \frac{\epsilon}{2}. \tag{B.23}$$

It follows from equation (4.51) that for $n > n_1$, $w \in [0, M]$ and $k = 1, \ldots, K$

$$\begin{aligned}
&|\tilde{J}_{n,k}(w) - \tilde{J}_{x,k}(w)| \\
&\leq c_k \int_w^{\infty} \left|\exp\left(\int_w^s -x_n\left(u\right)\, \mathrm{d}u\right) - \exp\left(\int_w^s -x\left(u\right)\, \mathrm{d}u\right)\right| \frac{\gamma_k'(s)}{\rho_k}\, \mathrm{d}s \\
&\leq c_k \int_w^{w_1} \left|\exp\left(\int_w^s -x_n\left(u\right)\, \mathrm{d}u\right) - \exp\left(\int_w^s -x\left(u\right)\, \mathrm{d}u\right)\right| M_1\, \mathrm{d}s \\
&\quad + c_k \int_{w_1}^{\infty} \left|\exp\left(\int_w^s -x_n\left(u\right)\, \mathrm{d}u\right) - \exp\left(\int_w^s -x\left(u\right)\, \mathrm{d}u\right)\right| M_1\, \mathrm{d}s \\
&\leq c_k \int_0^{w_1} c_{M,n} M_1\, \mathrm{d}s + \frac{\epsilon}{2} \leq \epsilon.
\end{aligned}$$

The inequality in the third line follows from (B.19). The inequalities in the last line follow

from equations (B.22) and (B.23). Note that n_1 is independent of w. Thus, $\tilde{J}_{n,k} \to \tilde{J}_{x,k}$ uniformly on $[0, M]$ for $k = 1, \ldots, K$.

Next, we show that $H_n \to H$ uniformly on $[0, M]$. Since $x(w) \geq -2\theta/\sigma^2$ and $x_n(w) \geq -2\theta/\sigma^2$ for n, w, it follows from (B.18) that for $k = 1, \ldots, K$, $n \geq 1$ and $w \geq 0$,

$$\tilde{J}_{x,k}(w) \geq r_k + \frac{c_k \sigma^2}{2\theta\rho_k} \sup_{t \geq 0} \gamma_k'(t) \quad \text{and} \quad \tilde{J}_{n,k}(w) \geq r_k + \frac{c_k \sigma^2}{2\theta\rho_k} \sup_{t \geq 0} \gamma_k'(t).$$

Thus, applying the mean value theorem to the function $1/2x^\delta$ yields the following: For $w \geq 0$, $k = 1, \ldots, K$ and $n \geq 1$,

$$\begin{aligned} \left| \frac{1}{2(\tilde{J}_{n,k}(w))^\delta} - \frac{1}{2(\tilde{J}_{x,k}(w))^\delta} \right| &= \frac{\delta}{2(\bar{J}_{n,k}(w))^{\delta+1}} |\tilde{J}_{k,n}(w) - \tilde{J}_{x,k}(w)| \\ &\leq \frac{\delta}{2} \left(r_k + \frac{c_k \sigma^2}{2\theta\rho_k} \sup_{t \geq 0} \gamma_k'(t) \right)^{-(\delta+1)} |\tilde{J}_{n,k}(w) - \tilde{J}_{x,k}(w)|. \end{aligned}$$

where $\bar{J}_{n,k}(w)$ is between $\tilde{J}_{n,k}(w)$ and $\tilde{J}_{x,k}(w)$. Substituting this inequality into equation (4.52), we obtain that for $w \in [0, M]$,

$$\begin{aligned} |H_n(w) - H_x(w)| &\leq \int_0^w \sum_{k=1}^K \gamma_k'(s) \left| \frac{1}{2(\tilde{J}_{n,k}(s))^\delta} - \frac{1}{2(\tilde{J}_{x,k}(s))^\delta} \right| \mathrm{d}s \\ &\leq \int_0^M \sum_{k=1}^K \gamma_k'(s) \left| \frac{1}{2(\tilde{J}_{k,n}(s))^\delta} - \frac{1}{2(\tilde{J}_{x,k}(s))^\delta} \right| \mathrm{d}s \\ &\leq \int_0^M \sum_{k=1}^K \frac{\delta}{2} \left(r_k + \frac{c_k \sigma^2}{2\theta\rho_k} \sup_{t \geq 0} \gamma_k'(t) \right)^{-(\delta+1)} |\tilde{J}_{k,n}(s) - \tilde{J}_{x,k}(s)| \, \mathrm{d}s. \end{aligned}$$

Since $\tilde{J}_{n,k} \to \tilde{J}_{x,k}$ uniformly on $[0, M]$, the term on the right-hand side of the last line converges to zero as $n \to \infty$. Thus, H_n converges to H_x uniformly on $[0, M]$.

We end the proof of property (i) by showing that $\hat{\beta}_n \to \hat{\beta}_x$ uniformly on $[0, M]$. It follows

from (4.53) that for $w \in [0, M]$,

$$
\begin{aligned}
|\hat{\beta}_n(w) - \hat{\beta}_x(w)| &= \hat{\beta}_n(w)\hat{\beta}_x(w) \left| \frac{1}{\hat{\beta}_n(w)} - \frac{1}{\hat{\beta}_x(w)} \right| \\
&= \hat{\beta}_n(w)\hat{\beta}_x(w) \left| \int_w^\infty \exp\left[\frac{2}{\sigma^2} \left(\theta(s-w) - \int_w^s H_n(u)\,\mathrm{d}u \right) \right] \mathrm{d}s \right. \\
&\quad \left. - \int_w^\infty \exp\left[\frac{2}{\sigma^2} \left(\theta(s-w) - \int_w^s H_x(u)\,\mathrm{d}u \right) \right] \mathrm{d}s \right| \\
&= \hat{\beta}_n(w)\hat{\beta}_x(w) \left| \int_w^\infty \exp\left(\frac{2\theta(s-w)}{\sigma^2} \right) \exp\left[-\frac{2}{\sigma^2} \int_w^s H_n(u)\,\mathrm{d}u \right] \right. \\
&\quad \left. - \int_w^\infty \exp\left[-\frac{2}{\sigma^2} \int_w^s H_x(u)\,\mathrm{d}u \right] \mathrm{d}s \right| \\
&\leq \left(\bar{\beta}(M)\right)^2 \int_w^\infty \exp\left(\frac{2\theta(s-w)}{\sigma^2} \right) \left| \exp\left[-\frac{2}{\sigma^2} \int_w^s H_n(u)\,\mathrm{d}u \right] \right. \\
&\quad \left. - \int_w^\infty \exp\left[-\frac{2}{\sigma^2} \int_w^s H_x(u)\,\mathrm{d}u \right] \right| \mathrm{d}s.
\end{aligned}
\tag{B.24}
$$

The inequality follows by applying Jensen's inequality for the absolute value function and property (ii) that

$$
\hat{\beta}_n(w) \leq \bar{\beta}(w) \leq \bar{\beta}(M) \quad \text{and} \quad \hat{\beta}_x(w) \leq \bar{\beta}(w) \leq \bar{\beta}(M).
$$

For any $\epsilon > 0$, there exits $t_1 > M$ such that

$$
\int_{t_1}^\infty \exp\left(\frac{2\theta(s-M)}{\sigma^2} \right) \mathrm{d}s = -\frac{\sigma^2}{2\theta(t_1 - M)} < \epsilon. \tag{B.25}
$$

Substituting this inequality into (B.24), we have the following: For $w \in [0, M]$,

$$
\begin{aligned}
&|\hat{\beta}_n(w) - \hat{\beta}(w)| \\
&\leq \left(\bar{\beta}(M)\right)^2 \int_w^{t_1} \exp\left(\frac{2\theta(s-w)}{\sigma^2}\right) \left| \exp\left[-\frac{2}{\sigma^2}\int_w^s H_n(u)\,\mathrm{d}u\right] \right. \\
&\quad \left. - \int_w^{\infty} \exp\left[-\frac{2}{\sigma^2}\int_w^s H_x(u)\,\mathrm{d}u\right] \right| \mathrm{d}s \\
&\quad + \left(\bar{\beta}(M)\right)^2 \int_{t_1}^{\infty} \exp\left(\frac{2\theta(s-w)}{\sigma^2}\right) \left| \exp\left[-\frac{2}{\sigma^2}\int_w^s H_n(u)\,\mathrm{d}u\right] \right. \\
&\quad \left. - \int_w^{\infty} \exp\left[-\frac{2}{\sigma^2}\int_w^s H_x(u)\,\mathrm{d}u\right] \right| \mathrm{d}s \\
&\leq \left(\bar{\beta}(M)\right)^2 \int_w^{t_1} \exp\left(\frac{2\theta(s-w)}{\sigma^2}\right) \left| \exp\left[-\frac{2}{\sigma^2}\int_w^s H_n(u)\,\mathrm{d}u\right] \right. \\
&\quad \left. - \int_w^{\infty} \exp\left[-\frac{2}{\sigma^2}\int_w^s H_x(u)\,\mathrm{d}u\right] \right| \mathrm{d}s + \left(\bar{\beta}(M)\right)^2 \int_{t_1}^{\infty} 2\exp\left(\frac{2\theta(s-w)}{\sigma^2}\right) \mathrm{d}s \\
&\leq \left(\bar{\beta}(M)\right)^2 \int_w^{t_1} \exp\left(\frac{2\theta(s-w)}{\sigma^2}\right) \left| \exp\left[-\frac{2}{\sigma^2}\int_w^s H_n(u)\,\mathrm{d}u\right] \right. \\
&\quad \left. - \int_w^{\infty} \exp\left[-\frac{2}{\sigma^2}\int_w^s H_x(u)\,\mathrm{d}u\right] \right| \mathrm{d}s + 2\left(\bar{\beta}(M)\right)^2 \epsilon.
\end{aligned}
\tag{B.26}
$$

The second inequality follows from

$$
\exp\left(-\frac{2}{\sigma^2}\int_w^s H_n(u)\,\mathrm{d}u\right) \leq 1,
$$

and that $|a-b| \leq |a| + |b|$. The last inequality follows from (B.25). Note that the following holds: For $w \in [0, M]$,

$$
\begin{aligned}
&\int_w^{t_1} \exp\left(\frac{2\theta(s-w)}{\sigma^2}\right) \left| \exp\left(-\frac{2}{\sigma^2}\int_w^s H_n(u)\,\mathrm{d}u\right) - \exp\left(-\frac{2}{\sigma^2}\int_w^s H_x(u)\,\mathrm{d}u\right) \right| \mathrm{d}s \\
&\leq \int_w^{t_1} \left| \exp\left(-\frac{2}{\sigma^2}\int_w^s H_n(u)\,\mathrm{d}u\right) - \exp\left(-\frac{2}{\sigma^2}\int_w^s H_x(u)\,\mathrm{d}u\right) \right| \mathrm{d}s \\
&= \int_w^{t_1} \exp\left(-\frac{2}{\sigma^2}\int_w^s H_n(u)\,\mathrm{d}u\right) \left| 1 - \exp\left(-\frac{2}{\sigma^2}\int_w^s (H_x(u) - H_n(u))\,\mathrm{d}u\right) \right| \mathrm{d}s \\
&\leq \int_0^{t_1} \left| 1 - \exp\left(-\frac{2}{\sigma^2}\int_w^s (H_x(u) - H_n(u))\,\mathrm{d}u\right) \right| \mathrm{d}s.
\end{aligned}
\tag{B.27}
$$

The first inequality follows from $\theta < 0$. Therefore, we have that $\exp(2\theta(s-w)/\sigma^2) < 1$. The second inequality follows from the non-negativity of $H_n(\cdot)$. Since $H_n \to H_x$ uniformly on $[0, t_1]$, there exists n_2 such that $|H_n(w) - H_x(w)| < \epsilon$ for $w \in [0, t_1]$ and $n > n_2$. Therefore, the following holds: For $w \in [0, M]$ and $s \in [0, t_1]$,

$$\left|1 - \exp\left(-\frac{2}{\sigma^2}\int_w^s (H_x(u) - H_n(u))\,\mathrm{d}u\right)\right| \leq \exp\left(\frac{2\epsilon(s-w)}{\sigma^2}\right) - 1 \leq \exp\left(\frac{2\epsilon t_1}{\sigma^2}\right) - 1.$$

Substituting this inequality and (B.27) into (B.26) yields the following: For $w \in [0, M]$,

$$|\hat{\beta}_n(w) - \hat{\beta}(w)| \leq \left(\bar{\beta}(M)\right)^2 \left(\exp\left(\frac{2\epsilon t_1}{\sigma^2}\right) - 1\right) t_1 + 2\left(\bar{\beta}(M)\right)^2 \epsilon.$$

Note that the right-hand side of the inequality is independent of w. Thus, we have shown that $\hat{\beta}_n \to \hat{\beta}_x$ uniformly on $[0, M]$.

Property (iii) follows from (4.51). To be more specific, we have that for $w \geq 0$ and $k = 1, \ldots, K$,

$$\begin{aligned}
\tilde{J}_{x,k}(w) &= r_k - c_k \int_w^\infty \exp\left(\int_w^s -x(u)\,\mathrm{d}u\right) \frac{\gamma_k'(s)}{\rho_k}\,\mathrm{d}s \\
&\geq r_k - c_k \sup_{t\geq 0} \frac{\gamma_k'(t)}{\rho_k} \int_w^\infty \exp\left(-\underline{\beta}(w)(s-w)\right)\,\mathrm{d}s \\
&= r_k - \frac{c_k}{\underline{\beta}(w)\rho_k} \sup_{t\geq 0} \gamma_k'(t) = \underline{J}_k(w).
\end{aligned}$$

Since $\underline{\beta}(w) \geq -2\theta/\sigma^2$, any $x \in [\underline{\beta}(w), \bar{\beta}(w)]$ satisfies the condition $x \geq -2\theta/\sigma^2$. Thus, it follows from property (ii) that $\tilde{J}_{x,k}(w) < r_k$ for $w \geq 0$.

It follows from (4.52) that $H_x(w)$ is non-decreasing in w. It then follows from (4.37) that

for $w \geq 0$,

$$\begin{aligned}\hat{\beta}_x(w) &\geq \left(\int_w^\infty \exp\left[\frac{2}{\sigma^2}\left(\theta(s-w) - H_x(w)(s-w)\right)\right] \mathrm{d}s\right)^{-1} \\ &= \left(\int_0^\infty \exp\left[\frac{2}{\sigma^2}\left(\theta s - H_x(w)s\right)\right] \mathrm{d}s\right)^{-1} \\ &= -\frac{2}{\sigma^2}(\theta - H_x(w)).\end{aligned}$$

Rearranging the terms gives us that

$$\hat{\beta}_x(w) + \frac{2}{\sigma^2}(\theta - H_x(w)) \geq 0,\ w \geq 0.$$

We end the proof by showing property (iv). Note from equation (4.55) that $\underline{H}(w) \to \infty$ as $w \to \infty$. Thus, it follows from (4.57) that $\underline{\beta}(w) \to \infty$ as $w \to \infty$. Combining this and equation (4.59), we have that $\underline{J}_k(w) \to r_k$ as $w \to \infty$. Thus, we have that $\tilde{J}_{x,k}(w) \to r_k$ as $w \to \infty$.

Since we have shown in property (ii) that $H_x(w) \geq \underline{H}(w)$ and that $\underline{H}(w) \to \infty$, we have that $\lim_{w\to\infty} H_x(w) = \infty$.

We then show that

$$\lim_{w\to\infty} \hat{\beta}_x(w) + \frac{2}{\sigma^2}(\theta - H_x(w)) = 0.$$

It suffices to show that for any $\epsilon > 0$, there exist positive constants w_0 such that

$$\hat{\beta}_x(w) + \frac{2}{\sigma^2}(\theta - H_x(w)) \leq \epsilon,\ w \geq w_0.$$

It follows from Assumption 4 and Lemma 11 that

$$\lim_{w\to\infty} \frac{\gamma_k'(w)}{2(\tilde{J}_k(w))^\delta} = \frac{\bar{\gamma}_k'}{2r_k^\delta},\ k = 1, \ldots, K.$$

Thus, there exists w_1 such that

$$\frac{\gamma_k'(w)}{2(\tilde{J}_{x,k}(w))^\delta} \leq \frac{\bar{\gamma}_k'}{r_k^\delta},\ w \geq w_1.$$

Substituting this inequality into equation (4.52) and letting $\hat{q}_\infty = \sum_{k=1}^K \bar{\gamma}_k'/2r_k^\delta$ yields the following: For $u > 0$ and $w \geq w_1$,

$$0 < H_x(w+u) - H_x(w) \leq \sum_{k=1}^K \frac{\bar{\gamma}_k'}{r_k^\delta} u = 2\hat{q}_\infty u. \tag{B.28}$$

Thus, it follows from equation (4.53) that for $w \geq w_1$,

$$\begin{aligned}
\frac{1}{\hat{\beta}_x(w)} &= \int_w^\infty \exp\left[\frac{2}{\sigma^2}\left(\theta(s-w) - \int_w^s H_x(u)\,\mathrm{d}u\right)\right]\,\mathrm{d}s \\
&= \int_0^\infty \exp\left[\frac{2}{\sigma^2}\left(\theta s - \int_0^s H_x(w+u)\,\mathrm{d}u\right)\right]\,\mathrm{d}s \\
&= \int_0^\infty \exp\left[\frac{2}{\sigma^2}\left(\theta s - H_x(w)s - \int_0^s H_x(w+u) - H_x(w)\,\mathrm{d}u\right)\right]\,\mathrm{d}s \\
&\geq \int_0^\infty \exp\left[\frac{2}{\sigma^2}\left(\theta s - H_x(w)s - \int_0^s 2\hat{q}_\infty u\,\mathrm{d}u\right)\right]\,\mathrm{d}s \\
&= \int_0^\infty \exp\left[\frac{2}{\sigma^2}\left(\theta s - H_x(w)s - \hat{q}_\infty s^2\right)\right]\,\mathrm{d}s.
\end{aligned} \tag{B.29}$$

The inequality follows from equation (B.28). Note that $\hat{q}_\infty s^2 \leq \epsilon s$ for $s < \epsilon/\hat{q}_\infty$. Substituting this inequality into equation (B.29), we have that for $w \geq w_1$,

$$\begin{aligned}
\frac{1}{\hat{\beta}_x(w)} &\geq \int_0^{\epsilon/q_\infty} \exp\left[\frac{2}{\sigma^2}\left(\theta s - H_x(w)s - \epsilon s\right)\right]\,\mathrm{d}s \\
&= \left(-\frac{2}{\sigma^2}(\theta - H_x(w) - \epsilon)\right)^{-1}\left[1 - \exp\left(\frac{2}{\sigma^2}\left(\theta - H_x(w) - \epsilon\right)\frac{\epsilon}{\hat{q}_\infty}\right)\right].
\end{aligned} \tag{B.30}$$

Because $\lim_{t\to\infty} te^{-at} = 0$ for any $a > 0$, there exists t_1 such that

$$t\exp\left(-\frac{\epsilon}{\hat{q}_\infty}t\right) < \epsilon,\ t \geq t_1.$$

Because $H_x(w) \to \infty$ as $w \to \infty$, there exists $w_2 > w_1$ such that

$$-\frac{2}{\sigma^2}\left(\theta - H_x(w) - \epsilon\right) \geq t_1,\ w \geq w_2.$$

Thus, the following holds:

$$-\frac{2}{\sigma^2}\left(\theta - H_x(w) - \epsilon\right)\exp\left(\frac{2}{\sigma^2}\left(\theta - H_x(w) - \epsilon\right)\frac{\epsilon}{\hat{q}_\infty}\right) < \epsilon,\ w \geq w_2.$$

Substituting this inequity into equation (B.30), we have the following inequality:

$$\frac{1}{\hat{\beta}_x(w)} \geq \left(-\frac{2}{\sigma^2}(\theta - H_x(w) - \epsilon)\right)^{-1}\left[1 - \frac{\epsilon}{-\frac{2}{\sigma^2}\left(\theta - H_x(w) - \epsilon\right)}\right],\ w \geq w_2.$$

Therefore, we conclude that for $w \geq w_2$,

$$\hat{\beta}_x(w) \leq -\frac{2}{\sigma^2}(\theta - H_x(w) - \epsilon)\left[1 - \frac{\epsilon}{-\frac{2}{\sigma^2}\left(\theta - H_x(w) - \epsilon\right)}\right]^{-1}. \tag{B.31}$$

Taylor's expansion of $(1-t)^{-1}$ at $t = 0$ gives

$$\frac{1}{1-t} = 1 + \frac{1}{(1-\tilde{t})^2}t, \quad \text{where } \tilde{t} \in (0, t).$$

Because $(1-\tilde{t})^2 \to 1$ as $t \to 0$. There exists t_2 such that

$$(1-t)^{-1} < 1 + 2t \text{ for } 0 < t < t_2.$$

Moreover, since $H_x(w) \to \infty$, there exists $w_3 > w_2$ such that

$$0 < \frac{\epsilon}{-\frac{2}{\sigma^2}\left(\theta - H_x(w) - \epsilon\right)} < t_2,\ w \geq w_3.$$

Thus, it follows that

$$\left[1 - \frac{\epsilon}{-\frac{2}{\sigma^2}\left(\theta - H_x(w) - \epsilon\right)}\right]^{-1} < 1 + 2\frac{\epsilon}{-\frac{2}{\sigma^2}\left(\theta - H_x(w) - \epsilon\right)},\ w \geq w_3.$$

Substituting this inequality into equation (B.31), we obtain the following

$$\begin{aligned}\hat{\beta}_x(w) &\leq -\frac{2}{\sigma^2}(\theta - H_x(w) - \epsilon)\left(1 + 2\frac{\epsilon}{-\frac{2}{\sigma^2}\left(\theta - H_x(w) - \epsilon\right)}\right) \\ &= -\frac{2}{\sigma^2}(\theta - H_x(w) - \epsilon) + 2\epsilon,\ w \geq w_3.\end{aligned}$$

Rearranging the terms yields the following:

$$\hat{\beta}_x(w) + \frac{2}{\sigma^2}(\theta - H_x(w)) \leq \left(\frac{2}{\sigma^2} + 2\right)\epsilon,\ w \geq w_3,$$

which completes the proof.

□

Proof of Lemma 11. Consider the equilibrium quantities $(\hat{\beta}_W, \tilde{J}, H)$. Since $H(w) \geq 0$ for $w \geq 0$, it follows from equation (4.37) that for $w \geq 0$,

$$\begin{aligned}\hat{\beta}_W(w) &= \left(\int_w^\infty \exp\left[\frac{2}{\sigma^2}\left(\theta(s-w) - \int_w^s H(u)\,\mathrm{d}u\right)\right]\,\mathrm{d}s\right)^{-1} \\ &\geq \left(\int_w^\infty \exp\left[\frac{2}{\sigma^2}\theta(s-w)\right]\,\mathrm{d}s\right)^{-1} = -\frac{2\theta}{\sigma^2}.\end{aligned}$$

Since $\hat{\beta}_W = \Psi \circ \hat{\beta}_W$, it follows from property (ii) of Lemma 10 that $\hat{\beta}_W(w) \in [\underline{\beta}(w), \bar{\beta}(w)]$ for $w \geq 0$. It then follows from property (iii) of Lemma 10 that $\tilde{J}_k(w) \in [\underline{J}_k(w), r_k]$ for $w \geq 0$ and $k = 1, \ldots, K$.

It then follows from property (iv) of Lemma 10 that for $k = 1, \ldots, K$,

$$\lim_{w\to\infty} \hat{\beta}_W(w) + \frac{2}{\sigma^2}(\theta - H(w)) = 0 \ \text{ and } \lim_{w\to\infty} \tilde{J}_k(w) = r_k.$$

□

Proof of Lemma 14. It follows from equation (4.70) that for $w \geq 0$,

$$\begin{aligned}
\delta'_{\tilde{\beta}}(w) &= \tilde{\beta}^1(w)\left(\tilde{\beta}^1(w) - \frac{2}{\sigma^2}(\theta - H^1(w))\right) - \frac{2}{\sigma^2}\sum_{k=1}^{K}\frac{\gamma'_k(w)}{2(\hat{J}_k^1(w))^{\delta}} \\
&\quad - \tilde{\beta}^2(w)\left(\tilde{\beta}^2(w) - \frac{2}{\sigma^2}(\theta - H^2(w))\right) + \frac{2}{\sigma^2}\sum_{k=1}^{K}\frac{\gamma'_k(w)}{2(\hat{J}_k^2(w))^{\delta}} \\
&= \left(\tilde{\beta}^1(w) - \frac{2}{\sigma^2}(\theta - H^1(w)) + \tilde{\beta}^2(w)\right)\delta_{\tilde{\beta}}(w) + \frac{2}{\sigma^2}\tilde{\beta}^2(w)\delta_H(w) \\
&\quad + \sum_{k=1}^{K}\frac{\delta\gamma'_k(w)}{\sigma^2(\bar{J}_k(w))^{\delta+1}}\delta_{\tilde{J}_k}(w) \\
&= \left(-\frac{2\theta}{\sigma^2} + c(w) + \tilde{\beta}^2(w)\right)\delta_{\tilde{\beta}}(w) + \frac{2}{\sigma^2}\tilde{\beta}^2(w)\delta_H(w) + \sum_{k=1}^{K}\frac{\delta\gamma'_k(w)}{\sigma^2(\bar{J}_k(w))^{\delta+1}}\delta_{\tilde{J}_k}(w),
\end{aligned}$$

where

$$c(w) = \tilde{\beta}^1(w) - \frac{2}{\sigma^2}(\theta - H^1(w)) + \frac{2\theta}{\sigma^2} = \hat{\beta}^1_W(w) + \frac{2\theta}{\sigma^2}.$$

Substituting (4.76) into $\delta'_{\tilde{\beta}}(w)$, we have that for $w \geq 0$,

$$\delta'_{\tilde{\beta}}(w) = \left(-\frac{2\theta}{\sigma^2} + c(w) + \tilde{\beta}^2(w) - \frac{2\tilde{\beta}^2(w)g_0(w)}{\sigma^2}\right)\delta_{\tilde{\beta}}(w) + \sum_{k=1}^{K}\left(\frac{\delta}{\sigma^2 r_k^{\delta+1}} + l_k(w)\right)\delta_{\tilde{J}_k}(w).$$

Similarly, it follows from (4.71) and (4.76) that for $w \geq 0$ and $k = 1, \ldots K$,

$$\delta'_{\tilde{J}_k}(w) = n_k(w)\delta_{\tilde{\beta}}(w) + \left(-\frac{2\theta}{\sigma^2} + c(w)\right)\delta_{\tilde{J}_k}(w) + \sum_{i=1}^{K} m_i(w)\delta_{\tilde{J}_i}(w).$$

Writing $\delta'_{\tilde{\beta}}(w)$ and $\delta'_{J}(w)$ in the matrix form yields the following: For $w \geq 0$,

$$\begin{bmatrix} \delta'_{\tilde{\beta}}(w) \\ \delta'_{\tilde{J}_1}(w) \\ \cdots \\ \delta'_{\tilde{J}_K}(w) \end{bmatrix} = (A + c(w)I + B(w)) \begin{bmatrix} \delta_{\tilde{\beta}}(w) \\ \delta_{\tilde{J}_1}(w) \\ \cdots \\ \delta_{\tilde{J}_K}(w) \end{bmatrix}.$$

Since $H^1(w) \geq 0$ for $w \geq 0$, it follows from Lemma 11 that

$$\hat{\beta}^1_W(w) \geq -\frac{2}{\sigma^2}(\theta - H^1(w)) \geq -\frac{2\theta}{\sigma^2}, \ w \geq 0.$$

Thus $c(w) \geq 0$.

We complete the proof by showing that $\lim_{w\to\infty} |||B(w)|||_\infty = 0$. It follows from Lemmas 11 that $\bar{J}_k(w) \to r_k$ as $w \to \infty$ for all k. Note that there exists $c_k(w) \in (0,1)$ such that

$$\bar{J}_k(w) = c_k(w)\tilde{J}^1_k(w) + (1 - c_k(w))\tilde{J}^2_k(w), \ k = 1, \ldots, K \ \text{ and } \ w \geq 0,$$

To see this we apply the mean value theorem to the function $1/x^\delta$ and conclude that for $x_1 < x_2$, and some $\bar{x} \in (x_1, x_2)$ that

$$\frac{1}{x_1^\delta} - \frac{1}{x_2^\delta} = -\frac{\delta}{(\bar{x})^{\delta+1}}(x_2 - x_1).$$

Recognizing that this equation and equation (4.81) have the same structure gives the result. Thus, $\bar{J}_k(w) \to r_k$ as $w \to \infty$. Combining this with Assumption 4, we have that

$$\lim_{w\to\infty} \left(\frac{\gamma'_k(w)}{(\bar{J}_k(w))^{\delta+1}} - \frac{\bar{\gamma}'_k}{r_k^{\delta+1}} \right) = 0 \ \text{ for } \ k = 1, \ldots, K. \tag{B.32}$$

It follows from Lemma 11 that

$$\lim_{w\to\infty} \tilde{\beta}^2(w) = \lim_{w\to\infty} \left(r_k - \tilde{J}_k^2(w) \right) = 0 \text{ for } \ k = 1, \ldots, K. \tag{B.33}$$

In addition, it follows from Lemma 13 that there exist constants w_0 and M such that for $w \geq w_0$,

$$|g_i(w)| \leq M, \ w \geq 0 \ \text{ and } \ i = 0, \ldots, K. \tag{B.34}$$

Substituting (B.32)-(B.34) into (4.78)-(4.80), we have that for $k = 1, \ldots, K$,

$$\lim_{w\to\infty} \tilde{\beta}^2(w) \left(1 - \frac{2g_0(w)}{\sigma^2} \right) = 0 \ \text{ and } \ \lim_{w\to\infty} l_k(w) = \lim_{w\to\infty} m_k(w) = \lim_{w\to\infty} n_k(w) = 0.$$

This gives that $\lim_{w\to\infty} |||B(w)|||_\infty = 0$.

□

Prior to proving Lemma 15, we introduce two lemmas that facilitate the proof of Lemma 15.

Lemma 43. *[98] Let $m \geq 0$ be an integer. For any $\lambda < \sigma$, there exists a constant C such that for $t \geq 0$, $t^m e^{\lambda t} < Ce^{\sigma t}$.*

Lemma 44. *Let $\tilde{c}(\cdot)$ be a nonnegative continuous function. In addition, $\tilde{A}$ and $\tilde{B}(t)$ are $(K+1)\times(K+1)$ matrices that satisfy the following: First, the entries of $\tilde{B}(t)$ are functions of t such that $lim_{t\to\infty}|||B(t)|||_\infty = 0$. Second, I is the $(K+1)\times(K+1)$ identity matrix. Lastly, $\tilde{A}$ is an upper triangular matrix of the form*

$$\tilde{A} = aI + \begin{bmatrix} 0 & b_1 & \cdots & b_K \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix},$$

where $a > 0$ and $b_1, \ldots, b_K$ are constants. Then, there exist positive constants C, ϵ and κ

such that if $|||B(t)|||_\infty < \epsilon$ *for* $t \in [t_1, t_2]$*, the following holds: If* $y(\cdot)$ *is a solution to the following system of ODEs:*

$$y'(t) = -\left(\tilde{A} + \tilde{B}(t) + \tilde{c}(t)I\right) y(t), \tag{B.35}$$

then the solution $y(\cdot)$ *satisfies the following inequality:*[2]

$$||y(t_2)||_\infty \leq C||y(t_1)||_\infty e^{-\kappa(t_2 - t_1)}. \tag{B.36}$$

Proof. Let $\Phi(t)$ denote the matrix-valued function defined as follows[3]: For $t \geq 0$,

$$\Phi(t) = \exp\left\{-\int_0^t (a + \tilde{c}(s))\,\mathrm{d}s\right\} \left(I + \begin{bmatrix} 0 & b_1 t & \cdots & b_K t \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \right).$$

It is immediate to see that $\Phi'(t) = -(\tilde{A} + \tilde{c}(t)I)\Phi(t)$ for $t \geq 0$.

We then derive an equation that characterizes $y(t)$ in terms of $\Phi(t)$. We write $y(t)$ as $y(t) = \Phi(t)z(t)$ for some function $z(t)$, $t \geq 0$. Substituting $y(t) = \Phi(t)z(t)$ into equation (B.35), we have that

$$\Phi'(t)z(t) + \Phi(t)z'(t) = -(\tilde{A} + \tilde{c}(t))\Phi(t)z(t) - \tilde{B}(t)\Phi(t)z(t). \tag{B.37}$$

2. This lemma is a modified form of the Poincare-Lyapunov Theorem. In Poincare-Lyapunov Theorem, $y(t)$ satisfies $y'(t) = -(A + B(t))y(t)$. The proof of this lemma is modified from the proof of Theorem 7.1 in Verhulst [106].

3. In fact, the columns of $\Phi(t)$ are linearly independent solutions to the linear system $y'(t) = -(\tilde{A} + \tilde{c}(t))y(t)$, $t \geq 0$. Thus, $\Phi(t)$ is a fundamental matrix of this linear system and the general solution of this linear system can be written as $\{\Phi(t)c : c \in \mathbb{R}^{K+1}\}$.

Substituting $\Phi'(t) = -(\tilde{A} + \tilde{c}(t)I)\Phi(t)$ into equation (B.37), we obtain that

$$\Phi(t)z'(t) = -\tilde{B}(t)\Phi(t)z(t) = -\tilde{B}(t)y(t). \tag{B.38}$$

Note that $\Phi(t)$ is invertible. Multiplying both sides of equation (B.38) by $\Phi^{-1}(t)$ and integrating $z'(t)$, we have the following equation: For any $t_2 \geq t_1 \geq 0$,

$$z(t_2) = z(t_1) - \int_{t_1}^{t_2} \Phi^{-1}(s)\tilde{B}(s)y(s)\,\mathrm{d}s.$$

Multiplying both sides of the equation by $\Phi(t_2)$ and substituting $z(t_1) = \Phi^{-1}(t_1)y(t_1)$ into the equation, we obtain that

$$y(t_2) = \Phi(t_2)\Phi^{-1}(t_1)y(t_1) - \int_{t_1}^{t_2} \Phi(t_2)\Phi^{-1}(s)\tilde{B}(s)y(s)\,\mathrm{d}s,\ t_2 > t_1 \geq 0.$$

Taking the norm for both sides of the equation yields the following: For $t_2 > t_1 \geq 0$,

$$||y(t_2)||_\infty \leq |||\Phi(t_2)\Phi^{-1}(t_1)|||_\infty ||y(t_1)||_\infty + \int_{t_1}^{t_2} |||\Phi(t_2)\Phi^{-1}(s)|||_\infty |||\tilde{B}(s)|||_\infty ||y(s)||_\infty \,\mathrm{d}s. \tag{B.39}$$

We calculate $\Phi(t_2)\Phi^{-1}(s)$ as follows:

$$\Phi(t_2)\Phi^{-1}(s) = \exp\left\{-\int_s^{t_2} a + \tilde{c}(u)\,\mathrm{d}u\right\} \left(I + \begin{bmatrix} 0 & b_1(t_2-s) & \cdots & b_K(t_2-s) \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \right).$$

Therefore, the norm of $\Phi(t_2)\Phi^{-1}(s)$ satisfies the following:

$$|||\Phi(t_2)\Phi^{-1}(s)|||_\infty \leq \left|\exp\left(-\int_s^{t_2}(a + \tilde{c}(u))\,\mathrm{d}u\right)\right| \left(1 + \sum_{k=1}^{K} b_k(t_2 - s)\right). \tag{B.40}$$

Since $\tilde{c}(t) \geq 0$ for $t \geq 0$, the following holds: For $t_2 > s \geq 0$,

$$\left| \exp\left(-\int_s^{t_2} (a + \tilde{c}(u))\, \mathrm{d}u \right) \right| \leq e^{-a(t_2 - s)}.$$

Substituting this inequality into (B.40), we have the following:

$$|||\Phi(t_2)\Phi^{-1}(s)|||_\infty \leq e^{-a(t_2-s)} \left(1 + \sum_{k=1}^{K} b_k (t_2 - s) \right).$$

It follows from Lemma 43 that there exists a constant C_1 such that

$$|||\Phi(t_2)\Phi^{-1}(s)|||_\infty < C_1 e^{-a(t_2-s)/2}.$$

Substituting this inequality into equation (B.39), we obtain the following: For $t_2 > t_1 \geq 0$,

$$||y(t_2)||_\infty \leq C_1 e^{-a(t_2-t_1)/2} ||y(t_1)||_\infty + \int_{t_1}^{t_2} C_1 |||\tilde{B}(t)|||_\infty e^{-a(t_2-s)/2} ||y(s)||_\infty \, \mathrm{d}s. \quad \text{(B.41)}$$

Let $\epsilon_0 = a/4C_1$. Since $\lim_{t \to 0} |||\tilde{B}(t)|||_\infty = 0$, there exits t_0 such that for $t \geq t_0$, $|||\tilde{B}(t)|||_\infty < \epsilon_0$. Let t_1 and t_2 be constants such that $t_2 > t_1$ and $|||\tilde{B}(t)|||_\infty < \epsilon_0$ for $t \in [t_1, t_2]$. Thus, we have that

$$||y(t_2)||_\infty \leq C_1 e^{-a(t_2-t_1)/2} ||y(t_1)||_\infty + \int_{t_1}^{t_2} \frac{1}{4} a e^{-a(t_2-s)/2} ||y(s)||_\infty \, \mathrm{d}s.$$

Multiplying both sides by $\exp(a(t_2 - t_1)/2)$, we obtain that

$$e^{a(t_2-t_1)/2} ||y(t_2)||_\infty \leq C_1 ||y(t_1)||_\infty + \int_{t_1}^{t_2} \frac{1}{4} a e^{a(s-t_1)/2} ||y(s)||_\infty \, \mathrm{d}s, \ t_2 > t_1 \geq t_0.$$

Applying Gronwall's inequality [Lemma 1.2.1, 76] to the function $e^{a(t-t_1)/2} ||y(t)||_\infty$, we have

that

$$e^{a(t_2-t_1)/2}||y(t_2)||_\infty \leq C_1||y(t_1)||_\infty e^{a(t_2-t_1)/4},\ t_2 > t_1 \geq t_0.$$

Multiplying both sides by $\exp(-a(t_2-t_1)/2)$ and letting $\kappa = a/4 > 0$ yields the following:

$$||y(t_2)||_\infty \leq C_1||y(t_1)||_\infty e^{-\kappa(t_2-t_1)},\ t_2 > t_1 \geq t_0.$$

□

Proof of Lemma 15. We prove this lemma by contradiction. Suppose the solution $x(\cdot)$ to equation (4.82) satisfies both $x(t) \neq 0$ for $t \geq 0$ and $x(t) \to 0$ as $t \to \infty$.

Note that the matrices $\tilde{A}$, $\tilde{B}(\cdot)$ and the function $\tilde{c}(\cdot)$ satisfy the conditions in Lemma 44. Let C, κ and ϵ be the constants in Lemma 44. In addition, let t_0 be a constant such that $|||\tilde{B}(t)|||_\infty < \epsilon$ for $t \geq t_0$. The constant t_0 exists because $\lim_{t\to\infty} |||\tilde{B}(t)|||_\infty = 0$. Let $t_1 = t_0$. In addition, let $t_2 > t_1$ be a constant such that

$$Ce^{-\kappa(t_2-t_1)} \leq \frac{1}{2} \ \text{ and } \ ||x(t_2)||_\infty \leq ||x(t_1)||_\infty/2. \tag{B.42}$$

The second inequality follows from the contradiction assumption, $x(t) \to 0$. Therefore, for t_2 large enough, the second inequality holds. To facilitate the proof, let $y(t) = x(t_1 + t_2 - t)$ for $t \in [t_1, t_2]$. Note that $y(t)$ is a solution to the following initial value problem:

$$y'(t) = -(\tilde{A} + \tilde{c}(t_1 + t_2 - t)I + \tilde{B}(t_1 + t_2 - t))y(t) \ \text{ and } \ y(t_1) = x(t_2).$$

In addition, note that $|||\tilde{B}(t_1 + t_2 - t)|||_\infty \leq \epsilon$ for $t \in [t_1, t_2]$. Thus, it follows from Lemma 44 that $y(\cdot)$ satisfies (B.36). Then it follows from (B.36) that

$$||x(t_1)||_\infty = ||y(t_2)||_\infty \leq C||y(t_1)||_\infty e^{-\kappa(t_2-t_1)}.$$

Combining this inequality and (B.42), we obtain the following:

$$||x(t_1)||_\infty \leq C||y(t_1)||_\infty e^{-\kappa(t_2-t_1)} \leq \frac{1}{2}||y(t_1)||_\infty = \frac{1}{2}||x(t_2)||_\infty \leq \frac{1}{4}||x(t_1)||_\infty.$$

This inequality holds only if $x(t_1) = 0$. This contradicts the assumption that $x(t) \neq 0$ for $t \geq 0$. Thus, we must have that $x \not\to 0$ as $t \to \infty$.

□

B.1.3 Proof of the Results in Section 4.2

Proof of Corollary 12. It follows from equation (4.47) that for $w \geq 0$,

$$H'(w) = \sum_{k=1}^{K} \frac{\gamma_k'(w)}{2(\tilde{J}_k(w))^\delta}.$$

Recall from Assumption 4 and Lemma 11 that for $k = 1, \ldots, K$, $\lim_{w\to\infty} \gamma_k'(w) = \bar{\gamma}_k'$ and $\lim_{w\to\infty} \tilde{J}_k(w) = r_k$. Thus, we have that

$$\lim_{x\to\infty} H'(x) = \sum_{k=1}^{K} \frac{\bar{\gamma}_k'}{2r_k^\delta}.$$

By L'Hôspital's rule [Theorem 5.13, 96], the following holds:

$$\lim_{x\to\infty} \frac{H(x)}{x} = \lim_{x\to\infty} H'(x) = \sum_{k=1}^{K} \frac{\bar{\gamma}_k'}{2r_k^\delta}.$$

Therefore, it follows from Lemma 11 that

$$\lim_{w\to\infty} \frac{\hat{\beta}_W(x)}{x} = \frac{2}{\sigma^2}\left(\sum_{k=1}^{K} \frac{\bar{\gamma}_k'}{2r_k^\delta}\right).$$

□

Proof of Lemma 16. We first construct a truncated equilibrium using the auxiliary functions and lemmas in Appendix B.2. Let $\hat{\beta}_W(w) = \tilde{\beta}_\infty w$ and $\tilde{J}_k(w) = r_k$ for $k = 1, \ldots, K$ and $w \geq T$. In particular, $\hat{\beta}_W(T) = \tilde{\beta}_\infty T$ and $\tilde{J}_k(T) = r_k$ for $k = 1, \ldots, K$. Note that $(\hat{\beta}_W(T), \tilde{J}(T)) \in \mathcal{A}$ where A is defined in (B.69) for a large T[4]. We also let $H(T) = \zeta_T(\hat{\beta}_W(T), \tilde{J}(T))$. The value $H(T)$ is well-defined because $\zeta_T(\cdot)$ is well-defined in $\mathcal{A}$; see Lemma 50. In addition, we define the quantities in the truncated equilibrium as follows:

$$(\hat{\beta}_W(s), \tilde{J}(s), H(s)) = \phi^T(T - s; X_0), \ s \in [0, T], \tag{B.43}$$

where $X_0 = (\hat{\beta}_W(T), \tilde{J}(T), H(T))$ and the function ϕ is defined by the solution of the initial value problem given in Lemma 46.

We first show that $(\hat{\beta}_W(s), \tilde{J}(s), H(s))$, $s \in [0, T]$ satisfies the conditions in Definition 5. The first condition is satisfied by construction. In addition, it follows from Lemma 46 that the following holds: For $s \in [0, T]$ and $k = 1, \ldots, K$,

$$\begin{aligned}
\hat{\beta}'_W(s) &= -f_0^T(T - s, \phi(T - s; T, X_0)), \\
\tilde{J}'_k(s) &= -f_k^T(T - s, \phi(T - s; T, X_0)), \\
H'(s) &= -f_{K+1}^T(T - s, \phi(T - s; T, X_0)),
\end{aligned}$$

where $f_i^T(\cdot)$ is given in (B.50)-(B.52), $i = 0, 1, \ldots, K + 1$. Note that $(\hat{\beta}_W(s), \tilde{J}(s), H(s))$ satisfies (4.45)-(4.47) by comparing the right-hand side of (4.45)-(4.47) to that of (B.50)-(B.52). In addition, it follows from (B.70) that

$$H(0) = \phi_{K+1}^T(T; X_0) = 0.$$

Thus, (4.50) holds as well. Since both conditions hold, $(\hat{\beta}_W(s), \tilde{J}(s), H(s))$, $s \in [0, T]$ is a

4. We approximate the equilibrium with a truncated one with large T. Thus, we can assume that T is large enough such that this condition is satisfied.

truncated equilibrium.

Next, we prove the uniqueness of the truncated equilibrium by showing that the only quantities satisfying the two conditions in Definition 5 are $(\hat{\beta}_W(s), \tilde{J}(s), H(s)),\ s \geq 0$. To be more specific, for any given equilibrium quantities $(\hat{\beta}^0_W(s), \tilde{J}^0(s), H^0(s))$ that satisfy the two conditions in Definition 5, $s \in [0, T]$ we show that the following must hold:

$$(\hat{\beta}^0_W(s), \tilde{J}^0(s), H^0(s)) = (\hat{\beta}_W(s), \tilde{J}(s), H(s)),\ s \in [0, T].$$

To facilitate the proof, denote $\tilde{X}_0 = (\hat{\beta}^0_W(T), \tilde{J}^0(T), H^0(T))$ and let $\tilde{\phi} = (\tilde{\phi}_0, \ldots, \tilde{\phi}_{K+1})$ denote the function as follows:

$$\tilde{\phi}(s) = (\hat{\beta}^0_W(T-s), \tilde{J}^0(T-s), H^0(T-s)),\ s \in [0, T]. \tag{B.44}$$

It follows from the second condition in Definition 5 that $(\hat{\beta}^0_W(s), \tilde{J}^0(s), H^0(s)),\ s \in [0, T]$ satisfies (4.45)-(4.47). Note that $H^0(s) \geq 0$ for $s \in [0, T]$. Thus, the $[\cdot]^+$ in (B.50) is immaterial. It is immediate that $\tilde{\phi}(s)$ satisfies (B.53) with the initial value $\tilde{X}_0$. Thus, it follows from Lemma 46 that

$$\tilde{\phi}(s) = \phi^T(s; \tilde{X}_0) \quad \text{for} \quad s \in [0, T]. \tag{B.45}$$

In addition, the following holds:

$$\phi^T_{K+1}(T; \tilde{X}_0) = \tilde{\phi}_{K+1}(T) = H^0(0) = 0,$$

where the last equality follows from Definition 5 that $H^0(0)$ satisfies (4.50). It follows from (B.70) and Lemma 50 that

$$H^0(T) = \zeta_T(\hat{\beta}^0_W(T), \tilde{J}^0(T)).$$

In addition, it follows from the first condition in Definition 5 that

$$\hat{\beta}_W^0(T) = \hat{\beta}_W(T) = \tilde{\beta}_\infty T, \text{ and } \tilde{J}_k^0(T) = \tilde{J}_k(T),\ k = 1, \ldots, K. \tag{B.46}$$

Thus, the following holds:

$$H^0(T) = \zeta_T(\hat{\beta}_W(T), \tilde{J}(T)) = H(T), \tag{B.47}$$

where the last equality holds by the construction of $H(T)$. Therefore, we conclude from (B.46)-(B.47) that $\tilde{X}_0 = X_0$. It follows from (B.45) that $\tilde{\phi}(s) = \phi^T(s; X_0)$ for $s \in [0, T]$. Thus, it follows from (B.43)-(B.44) that for $s \in [0, T]$,

$$(\hat{\beta}_W^0(s), \tilde{J}^0(s), H^0(s)) = \phi^T(T - s; X_0) = \tilde{\phi}(T - s) = (\hat{\beta}_W(s), \tilde{J}(s), H(s)).$$

□

Proof of Lemma 17. It immediate from (4.87)-(4.89) that if (for $k = 1, \ldots, K$),

$$\hat{\beta}_W^1(w + \Delta) \geq \hat{\beta}_W^2(w + \Delta),\ \tilde{J}_k^1(w + \Delta) \geq \tilde{J}_k^2(w + \Delta) \text{ and } H^1(w + \Delta) > H^2(w + \Delta),$$

then the following holds:

$$\hat{\beta}_W^1(w) \geq \hat{\beta}_W^2(w,\ \tilde{J}_k^1(w) \geq \tilde{J}_k^2(w) \text{ and } H^1(w) > H^2(w). \tag{B.48}$$

Note that we start with the initial values such that (For $k = 1, \ldots, K$,)

$$\hat{\beta}_W^1(T) = \hat{\beta}_W^2(T),\ \tilde{J}_k^1(T) = \tilde{J}_k^2(T) \text{ and } H^1(T) > H^2(T),$$

By applying the inequality (B.48) inductively from the terminal time T, we conclude that $H^1(0) > H^2(0)$.

□

B.2 Auxiliary Technical Lemmas and the Proof of Lemma 13

This section proves Lemma 13 which characterizes the relationship between $\delta_{\tilde{\beta}}$, $\delta_{\tilde{J}}$ and δ_H. We prove this lemma in three steps. First, we construct an auxiliary function $\zeta^w(\cdot)$ and characterize its properties in Appendix B.2.1. Second, we focus on a restricted domain of $\zeta^w(\cdot)$ and provide further properties of $\zeta^w(\cdot)$ in Appendix B.2.2. In particular, we show that the partial derivatives of $\zeta^w(\cdot)$ on the restricted domain are bounded by a constant for w bounded beyond a constant w_0. Finally, we show that $\zeta^w(\cdot)$ characterizes $H(\cdot)$ in terms of $\hat{\beta}_W(\cdot)$ and $\tilde{J}(\cdot)$ in Appendix B.2.3. Appendix B.2.4 proves Lemma 13 using the characterization in Appendix B.2.3 and the properties of $\zeta^w(\cdot)$ provided in Appendices B.2.1 and B.2.2. Appendix B.3 provides a detailed roadmap showing how the proof of Lemma 13 is built up.

B.2.1 The Auxiliary Function $\zeta^w(\cdot)$

This section defines a function $\zeta^w(\cdot)$ which characterizes $\tilde{H}(w)$ in terms of $\hat{\beta}_W(w)$ and $\tilde{J}(w)$. To facilitate the analysis to follow, define the open set $\mathcal{I}$ as follows:

$$\mathcal{I} = \left(-\frac{2\theta}{\sigma^2},\ \infty\right) \times \prod_{k=1}^{K} \left(r_k + c_k \frac{\sigma^2}{2\theta\rho_k} \sup_{t\geq 0} \gamma_k'(t),\ r_k\right) \times \mathbb{R} \subseteq \mathbb{R}^{K+2}. \tag{B.49}$$

In addition, fix $w > 0$ and define a vector-valued function $F^w = (f_0^w, f_1^w, \ldots, f_{K+1}^w) : [0, \infty) \times \mathcal{I} \to \mathbb{R}^{K+2}$ as follows: For $t \in [0, w]$ and $X = (x_0, \ldots, x_{K+1}) \in \mathcal{I}$, let

$$f_0^w(t, X) = -x_0 \left[x_0 + \frac{2}{\sigma^2}(\theta - x_{K+1}^+) \right], \tag{B.50}$$

$$f_k^w(t, X) = -c_k \frac{\gamma_k'(w - t)}{\rho_k} + x_0(r_k - x_k), \ k = 1, \ldots, K, \tag{B.51}$$

$$f_{K+1}^w(t, X) = -\sum_{i=1}^{K} \frac{\gamma_i'(w - t)}{2(x_i)^\delta}, \tag{B.52}$$

where $x^+ = \max(x, 0)$. We define a system of ordinary differential equations using the function F^w as follows: For $t \in [0, w]$,

$$X'(t) = F^w(t, X(t)) \ \text{ with } \ X(0) = X_0. \tag{B.53}$$

The following lemma shows that the solution to (B.53) never leaves the set $\mathcal{I}$ provided that it starts in the set $\mathcal{I}$. This lemma facilitates the proof of Lemma 46, which shows that if the initial value $X_0 \in \mathcal{I}$, then the solution to (B.53) is unique.

Lemma 45. *Fix $w > 0$ and let $X(t) = (x_0(t), x_1(t), \ldots, x_{K+1}(t))$ denote a solution to (B.53) on $[0, t_0]$. If $X_0 = (x_0, x_1, \ldots, x_{K+1}) \in \mathcal{I}$, then $X(t) \in \mathcal{I}$ for $t \in [0, t_0]$. In addition, $\lim_{s \to t^-} |x_0(s)| < \infty$ and $\lim_{s \to t^-} |x_{K+1}(s)| < \infty$ for $t \in [0, t_0]$.*

Proof. The solution to equation (B.50) is given as follows: For $t \in [0, t_0]$,

$$\begin{aligned} &x_0(t) \\ &= \left[\frac{1}{x_0} \exp\left(\int_0^t \frac{2}{\sigma^2}(\theta - (x_{K+1}(s))^+) \, \mathrm{d}s \right) + \int_0^t \exp\left(\int_s^t \frac{2}{\sigma^2}(\theta - (x_{K+1}(u))^+) \, \mathrm{d}u \right) \mathrm{d}s \right]^{-1}. \end{aligned} \tag{B.54}$$

Note from (B.54) that $x_0(t) > 0$ because $x_0 > -2\theta/\sigma^2 \geq 0$. In addition, the following holds:

For $t \in [0, t_0]$,

$$\begin{aligned}
\frac{1}{x_0(t)} &= \frac{1}{x_0} \exp\left(\int_0^t \frac{2}{\sigma^2}(\theta - x_{K+1}^+(s))\,\mathrm{d}s\right) + \int_0^t \exp\left(\int_s^t \frac{2}{\sigma^2}(\theta - x_{K+1}^+(u))\,\mathrm{d}u\right)\,\mathrm{d}s \\
&\leq \frac{1}{x_0} \exp\left(\frac{2\theta t}{\sigma^2}\right) + \int_0^t \exp\left[\frac{2}{\sigma^2}\theta(t-s)\right]\,\mathrm{d}s \\
&< -\frac{\sigma^2}{2\theta} \exp\left(\frac{2\theta t}{\sigma^2}\right) + \frac{\sigma^2}{2\theta}\left[\exp\left(\frac{2\theta t}{\sigma^2}\right) - 1\right] = -\frac{\sigma^2}{2\theta},
\end{aligned}$$

where the last inequality follows from the assumption that $x_0 > -2\theta/\sigma^2 \geq 0$. Thus, we have that $x_0(t) > 2\theta/\sigma^2$ for $t \in [0, t_0]$.

The solution to equation (B.51) is given as follows: For $k = 1, \ldots, K$,

$$x_k(t) = x_k \exp\left(\int_0^t -x_0(s)\,\mathrm{d}s\right) + \int_0^t \exp\left(\int_s^t -x_0(u)\,\mathrm{d}u\right)\left(r_k x_0(s) - c_k \frac{\gamma_k'(s)}{\rho_k}\right)\,\mathrm{d}s. \tag{B.55}$$

Therefore, the following holds: For $k = 1, \ldots, K$ and $t \in [0, t_0]$,

$$\begin{aligned}
x_k(t) &\geq x_k \exp\left(\int_0^t -x_0(s)\,\mathrm{d}s\right) + r_k \int_0^w \exp\left(\int_s^t -x_0(u)\,\mathrm{d}u\right) x_0(s)\,\mathrm{d}s \\
&\quad - c_k \sup_{s\geq 0} \frac{\gamma_k'(s)}{\rho_k} \int_0^t \exp\left(\int_s^t -x_0(u)\,\mathrm{d}u\right)\,\mathrm{d}s \\
&\geq x_k \exp\left(\int_0^t -x_0(s)\,\mathrm{d}s\right) + r_k\left[1 - \exp\left(\int_0^t -x_0(s)\,\mathrm{d}s\right)\right] \\
&\quad - c_k \sup_{s\geq 0} \frac{\gamma_k'(s)}{\rho_k} \int_0^t \frac{x_0(s)}{-2\theta/\sigma^2} \exp\left(\int_s^t -x_0(u)\,\mathrm{d}u\right)\,\mathrm{d}s \\
&= x_k \exp\left(\int_0^t -x_0(s)\,\mathrm{d}s\right) + \left(r_k + \frac{c_k\sigma^2}{2\theta} \sup_{s\geq 0} \frac{\gamma_k'(s)}{\rho_k}\right)\left[1 - \exp\left(\int_0^t -x_0(s)\,\mathrm{d}s\right)\right] \\
&> r_k + \frac{c_k\sigma^2}{2\theta} \sup_{s\geq 0} \frac{\gamma_k'(s)}{\rho_k}.
\end{aligned} \tag{B.56}$$

The second inequality follows from $x_0(s) \geq -2\theta/\sigma^2$. The last inequality follows from the

assumption that

$$x_k < r_k + \frac{c_k \sigma^2}{2\theta} \sup_{s \geq 0} \frac{\gamma_k'(s)}{\rho_k}.$$

In addition, it follows from (B.55) that for $t \in [0, t_0]$ and $k = 1, \ldots, K$,

$$\begin{aligned} x_k(t) &\leq x_k \exp\left(\int_0^t -x_0(s)\,\mathrm{d}s\right) + r_k \int_0^t \exp\left(\int_s^t -x_0(u)\,\mathrm{d}u\right) x_0(s)\,\mathrm{d}s \\ &= x_k \exp\left(\int_0^t -x_0(s)\,\mathrm{d}s\right) + r_k \left[1 - \exp\left(\int_0^t -x_0(s)\,\mathrm{d}s\right)\right] < r_k, \end{aligned} \tag{B.57}$$

where the first inequality follows from $c_k \geq 0$ and the last inequality follows from the assumption that $x_k(0) < r_k$ for $k = 1, \ldots, K$. Thus, we conclude that $X(t) \in \mathcal{I}$.

We complete the proof by showing that $\lim_{s \to t^-} |x_0(s)| < \infty$ and $\lim_{s \to t^-} |x_{K+1}(s)| < \infty$ for $t \in [0, t_0]$. Substituting (B.56)-(B.57) into (B.52), we obtain the following: For $t \in [0, t_0]$,

$$x_{K+1} - \int_0^t \sum_{i=1}^K \frac{\gamma_i'(w-u)}{2\left(r_i + c_i \sigma^2 \sup_{s \geq 0} \gamma_k'(s)/2\theta\rho_k\right)^\delta}\,\mathrm{d}u \leq x_{K+1}(t) \leq x_{K+1} - \int_0^t \sum_{i=1}^K \frac{\gamma_i'(w-u)}{2(r_i)^\delta}\,\mathrm{d}u.$$

This inequality gives that $\lim_{s \to t^-} |x_{K+1}(t)| < \infty$ for $t \in [0, t_0]$. Denote

$$\bar{x}(t) = x_{K+1} - \int_0^t \sum_{i=1}^K \frac{\gamma_i'(w-u)}{2(r_i)^\delta}\,\mathrm{d}u.$$

By substituting $x_{K+1}(t) \leq \bar{x}(t)$ into (B.54), we obtain that for $t \in [0, t_0]$,

$$x_0(t) \leq \left[\frac{1}{x_0} \exp\left(\int_0^t \frac{2}{\sigma^2}(\theta - (\bar{x}(s))^+)\,\mathrm{d}s\right) + \int_0^t \exp\left(\int_s^t \frac{2}{\sigma^2}(\theta - (\bar{x}(u))^+)\,\mathrm{d}u\right)\,\mathrm{d}s\right]^{-1} < \infty.$$

Since we show that $x_0(t) \geq -2\theta/\sigma^2$, it is immediate that $\lim_{s \to t^-} |x_0(t)| < \infty$ for $t \in [0, t_0]$. □

Lemma 46. *Fixing $X_0 \in \mathcal{I}$ and $w \geq 0$, there exists a unique solution $\phi^w(t; X_0) : [0, w] \to \mathbb{R}^{K+2}$ to the initial value problem* (B.53).

Proof. Note that $f_i^w(\cdot)$ is continuously differentiable and thus locally Lipschitz continuous

on $(t, X) \in [0, \infty) \times \mathcal{I}$ for $i = 1, \ldots, K+1$. Fix $t_0 \geq 0$, $X_0 \in \mathcal{I}$ and $\epsilon > 0$. Then the following holds: For $X^i = (x_0^i, x_1^i, \ldots, x_{K+1}^i)$ and t^i such that $||X^i - X_0||_\infty < \epsilon$ and $|t^i - t_0| < \epsilon$, $i = 1, 2$,

$$
\begin{aligned}
&\left| f_0^w(t^1, X^1) - f_0^w(t^2, X^2) \right| \\
&= \left| x_0^1 \left[x_0^1 + \frac{2}{\sigma^2}(\theta - (x_{K+1}^1)^+) \right] - x_0^2 \left[x_0^2 + \frac{2}{\sigma^2}(\theta - (x_{K+1}^2)^+) \right] \right| \\
&\leq \left| x_0^1 \left[x_0^1 + \frac{2}{\sigma^2}(\theta - (x_{K+1}^1)^+) \right] - x_0^2 \left[x_0^2 + \frac{2}{\sigma^2}(\theta - (x_{K+1}^1)^+) \right] \right| \\
&\quad + \left| x_0^2 \left[x_0^2 + \frac{2}{\sigma^2}(\theta - (x_{K+1}^1)^+) \right] - x_0^2 \left[x_0^2 + \frac{2}{\sigma^2}(\theta - (x_{K+1}^2)^+) \right] \right| \\
&\leq |((x_0^1 + x_0^2)(x_0^1 - x_0^2)| + \left| \frac{2}{\sigma^2}(\theta - (x_{K+1}^1)^+)(x_0^1 - x_0^2) \right| \\
&\quad + \left| \frac{2x_0^2}{\sigma^2} \left[(x_{K+1}^1)^+ - (x_{K+1}^2)^+ \right] \right| \\
&\leq 2(x_0 + \epsilon)|x_0^1 - x_0^2| + \left| \frac{2}{\sigma^2}(\theta - (x_{K+1} + \epsilon)^+) \right| |x_0^1 - x_0^2| \\
&\quad + \frac{2(x_0 + \epsilon)}{\sigma^2} \left| x_{K+1}^1 - x_{K+1}^2 \right| \\
&\leq \left(2(x_0 + \epsilon) + \left| \frac{2}{\sigma^2}(\theta - (x_{K+1} + \epsilon)^+) \right| + \frac{2(x_0 + \epsilon)}{\sigma^2} \right) ||X^1 - X^2||_\infty.
\end{aligned}
$$

The first two inequalities follow from the inequality that $|a - b| \leq |a| + |b|$ for $a, b \in \mathbb{R}$. The third inequality follows from $||X^i - X_0||_\infty < \epsilon$ and $|a^+ - b^+| \leq |a - b|$ for $a, b \in \mathbb{R}$. Thus, the function $f_0^w(\cdot)$ is locally Lipschitz continuous as well. In sum, $F^w(\cdot)$ is locally Lipschitz continuous. Thus, it follows from the Picard Existence-Uniqueness Theorem that for $X_0 \in \mathcal{I}$, there exists $\epsilon_X > 0$ such that the initial value problem exists and is unique on $[0, \epsilon_X]$; see Theorem 1.3.1 in Kong [76].

We fix X_0 and extend the solution to the maximal interval of existence of a solution. Let $X(t)$, $t \geq 0$ denote a solution to (B.53). It follows from Lemma 45 that $X(t)$ never reaches the boundary of its (open) domain $\mathcal{I}$ in finite time. Thus, it follows from Theorem 1.4.1 in Kong [76] that the solution to (B.53) exists on $[0, \infty)$. In addition, since $F^W(\cdot)$ is locally Lipschitz continuous, the extension of the solution to $[0, \infty)$ is unique; see Theorem 1.4.1 in

Kong [76] as well. In particular, the solution to (B.53) exists and is unique for $t \in [0, w]$ and $X_0 \in \mathcal{I}$. We let $\phi^w(t; X_0)$, $t \in [0, w]$ denote the solution. □

We call the function $\phi^w(\,\cdot\,; X_0)$ the flow associated with (B.53). The flow $\phi^w(\,\cdot\,; X_0)$ emphasizes the dependence of a solution on the initial value X_0. To facilitate the analysis to follow, let ϕ_i^w denote the $(i+1)^{\text{th}}$ coordinate of ϕ^w for $i = 0, \ldots, K+1$. In other words, we can write ϕ^w as

$$\phi^w = (\phi_0^w, \phi_1^w, \ldots, \phi_K^w, \phi_{K+1}^w).$$

In addition, let $DF^w(\cdot)$ be the Jacobian matrix of function $F^w(\cdot)$. The following two lemmas guarantee the smoothness of ϕ^w and characterize its dependence on the initial value.

Lemma 47. *[Page 149, 63] If $F^w(\cdot)$ is continuously differentiable, $\phi^w(t; X)$ is continuously differentiable. That is, $\partial\phi_i^w/\partial t$ and $\partial\phi_i^w/\partial x_j$ exist and are continous in t and $X = (x_0, \ldots, x_{K+1})$ for $t \in [0, w]$ and $i, j = 0, \ldots, K+1$.*

Lemma 48. *[Page 152, 63] Fix $w \geq 0$ and $X_0 \in \mathcal{I}$. Given $U_0 = (u_0, \ldots, u_{K+1}) \in \mathbb{R}^{K+2}$, let $U(t; U_0)$ be the solution to the following linear differential equation:*

$$U'(t) = DF^w(t, \phi^w(t; X_0))U(t) \quad \textit{with} \quad U(0) = U_0.$$

Then the partial derivative of $\partial\phi_i^w/\partial x_j$ satisfies the following relationship: For $i = 0, 1, \ldots, K+1$ and $t \in [0, w]$,

$$\sum_{j=0}^{K+1} \frac{\partial\phi_i^w(t; X_0)}{\partial x_j} u_j = u_i(t; U_0), \tag{B.58}$$

where $u_j(t; U_0)$ is the $(j+1)^{\text{th}}$ coordinate of $U(t; U_0)$[5]*.*

We use Lemma 48 to characterize the monotonicity of ϕ^w with respect to the initial value, which leads to the following lemma.

5. Hirsch et al. [63] states this lemma in the vector form.

Lemma 49. *Fix $w \geq 0$ and suppose that the initial value $X_0 \in \mathcal{I}$ is such that $\phi^w_{K+1}(w; X_0) = 0$. Then the following holds: For $t \in [0, w]$,*

(i) $\phi^w_{K+1}(t; X_0) \geq 0$.

(ii) $\phi^w(t; X_0)$ *is continuously differentiable.*

(iii) $\partial \phi^w_i(t; X_0)/\partial x_j \geq 0$ *for* $i, j = 0, 1, \ldots, K+1$.

(iv) $\partial \phi^w_{K+1}(t; X_0)/\partial x_{K+1} \geq 1$.

Proof. Fix $w \geq 0$. We first show (i). It follows from Lemma 45 that $\phi^w(t; X_0) \in \mathcal{I}$, $t \in [0, w]$. Thus, the following inequality holds: For $t \in [0, w]$ and $k = 1, \ldots, K$,

$$\phi^w_k(t; X_0) \geq r_k + \frac{c_k \sigma^2}{2\theta\rho_k} \sup_{s \geq 0} \gamma'_k(s) > 0,$$

where the last inequality follows from Assumption 5. Therefore, it follows from (B.52) and $\gamma'_i(t) \geq 0$ (for $t \geq 0$) that $(\phi^w_{K+1})'(t; X_0) = f^w_{K+1}(t, \phi^w(t; X_0)) < 0$ for $t \in [0, w]$. In other words, $\phi^w_{K+1}(t; X_0)$ is decreasing in t for $t \in [0, w]$. Thus, it follows from the assumption $\phi^w_{K+1}(w; X_0) = 0$ that $\phi^w_{K+1}(t; X_0) \geq 0$ for $t \in [0, w]$.

Since $\phi^w_{K+1}(t; X_0) \geq 0$ for $t \in [0, w]$, the $[\,\cdot\,]^+$ function in (B.50) is immaterial. Thus, $F^w(t, \phi(t; X_0))$ is continuously differentiable in both arguments for $t \in [0, w]$. It follows from Lemma 47 that $\phi^w(t; X_0)$ is continuously differentiable for $t \in [0, w]$, which gives (ii).

Next, we use Lemma 48 to show (iii). Let $U(t; e_j) = (u_0(t; e_j), u_1(t; e_j), \ldots, u_{K+1}(t; e_j))$, $t \in [0, w]$ be the solution to

$$U'(t; e_j) = DF^w(t, \phi^w_i(t; X_0))U(t; e_j) \;\text{ with } U(0; e_j) = e_j \;\text{ for }\; j = 0, 1, \ldots, K+1, \quad \text{(B.59)}$$

where e_j is the $(K+2)$-dimensional vector with 1 in the $(j+1)^{\text{th}}$ component and zeros

elsewhere. It follows from Lemma 48 that

$$\frac{\partial \phi_i^w(t; X_0)}{\partial x_j} = u_i(t; e_j), \ t \in [0, w] \ \text{ and } \ i, j = 0, 1, \ldots, K+1. \tag{B.60}$$

Thus, proving property (iii) is equivalent to showing that $u_i(t; e_j) \geq 0$ for $t \in [0, w]$ and $i, j = 0, 1, \ldots, K+1$. To show this, we first look at the properties of the Jacobian matrix $DF^w(\cdot)$. To be more specific, it is easy to show that the signs of the entries of $DF^w(\cdot)$ are given as follows:

$$\text{sign}\left[\left(\frac{\partial f_i^w}{\partial x_j}\right)_{i,j=0,1,\ldots,K+1}\right] = \begin{array}{c} i\backslash j \\ 0 \\ 1 \\ 2 \\ 3 \\ \vdots \\ K \\ K+1 \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2 & 3 & \cdots & K & K+1 \end{array} \\ \begin{bmatrix} ? & 0 & 0 & 0 & \cdots & 0 & + \\ + & - & 0 & 0 & \cdots & 0 & 0 \\ + & 0 & - & 0 & \cdots & 0 & 0 \\ + & 0 & 0 & - & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ + & 0 & 0 & 0 & \cdots & - & 0 \\ 0 & + & + & + & \cdots & + & 0 \end{bmatrix} \end{array}. \tag{B.61}$$

The sign of $\partial f_0^w(t, X(t))/\partial x_0$ is unknown, $\partial f_{K+1}^w(t, X(t))/\partial x_{K+1} = 0$, and all other diagonal entries are negative. The essential observation for the proof is that all the off-diagonal entries are nonnegative. This observation will play a critical role to show that $u_i(t; e_j) \geq 0$ for $t \in [0, w]$. Define t_0 as follows:

$$t_0 = \inf\{t \in [0, w] : u_i(t; e_j) < 0 \ \text{ for some } \ i, j = 0, 1, \ldots, K\}. \tag{B.62}$$

To show that $u_i(t; e_j) \geq 0$ for $t \in [0, w]$ and $i, j = 0, 1, \ldots, K$, it suffices to show that $t_0 = w$. We show this by contradiction. (We assume by convention that $t_0 = w$ when $u_i(t; e_j) \geq 0$ for all $t \in [0, w]$.) Suppose $t_0 < w$. To facilitate the proof, we first derive the inequality

(B.65). Note that the following holds by definition of t_0: For $t \in [0, t_0]$:

$$u_i(t; e_j) \geq 0, \ \ i, j = 0, 1, \ldots, K+1. \tag{B.63}$$

Denote $X(t)$ as $X(t) = \phi^w(t; X_0)$ for $t \in [0, w]$. It follows from (B.59) that for $t \in [0, t_0]$ and $i, j = 0, 1, \ldots, K+1$,

$$\begin{aligned} u_i'(t; e_j) &= \frac{\partial f_i^w(t, X(t))}{\partial x_i} u_i(t; e_j) + \sum_{k \neq i} \frac{\partial f_i^w(t, X(t))}{\partial x_k} u_k(t; e_j) \\ &\geq \frac{\partial f_i^w(t, X(t))}{\partial x_i} u_i(t; e_j), \end{aligned} \tag{B.64}$$

where the inequality follows from equations (B.61) and (B.63). In particular, the non-diagonal entries of DF are nonnegative. Then it follows from this inequality immediately that the following holds: For $t \in [0, t_0]$ and $i, j = 0, 1, \ldots, K+1$,

$$\begin{aligned} &\left(\exp\left(-\int_0^t \frac{\partial f_i^w(s, X(s))}{\partial x_i} \, \mathrm{d}s \right) u_i(t; e_j) \right)' \\ &= \exp\left(-\int_0^t \frac{\partial f_i^w(s, X(s))}{\partial x_i} \, \mathrm{d}s \right) \left(-\frac{\partial f_i^w(t, X(t))}{\partial x_i} u_i(t; e_j) + u_i'(t; e_j) \right) \geq 0. \end{aligned}$$

Integrating the left-hand side of this inequality over $[0, t]$ yields the following: For $t \in [0, t_0]$ and $i, j = 0, 1, \ldots, K+1$,

$$\exp\left(-\int_0^t \frac{\partial f_i^w(s, X(s))}{\partial x_i} \, \mathrm{d}s \right) u_i(t; e_j) - u_i(0; e_j) \geq 0,$$

where the inequality follows because the derivative of the left-hand side is nonnegative as we just show. By rearranging the terms, we obtain the following: For $t \in [0, t_0]$ and $i, j = 0, 1, \ldots, K+1$,

$$u_i(t; e_j) \geq u_i(0; e_j) \exp\left(\int_0^t \frac{\partial f_i^w(s, X(s))}{\partial x_i} \, \mathrm{d}s \right). \tag{B.65}$$

We next discuss all possible combinations of i and j (that arises in (B.62)) and show that

there exists $\epsilon_2 > 0$ such that $u_i(t; e_j) \geq 0$ for $t \in [t_0, t_0 + \epsilon_2]$. This contradicts the definition of t_0. Some cases are discussed first because the discussion of other cases depend on their properties. The following table summarizes the indices of cases that discuss each combination of i and j.

$i \backslash j$	0	$1, \ldots, K$	$K+1$
0	Case 1	Case 3	Case 4
$1, \ldots, K$	Case 5	Case 1 if $i = j$; Case 6 if $i \neq j$	Case 7
$K+1$	Case 8	Case 2	Case 1

Table B.1: The indices of cases that discuss each combination of i and j.

- Case 1: $i = j \in \{0, 1, \ldots, K+1\}$. Since $u_i(0; e_i) = 1$, it follows from equation (B.65) that for $t \in [0, t_0]$,

$$u_i(t; e_i) \geq u_i(0; e_i) \exp\left(\int_0^t \frac{\partial f_i^w(u, X(u))}{\partial x_i} \, \mathrm{d}u \right) > 0.$$

  In particular, $u_i(t_0; e_i) > 0$. It follows from the continuity of $u_i(\cdot; e_i)$ that there exists $\epsilon_1 > 0$ such that $u_i(t; e_i) > 0$ for $t \in [t_0, t_0 + \epsilon_1]$ and $i = 0, 1, \ldots, K+1$.

- Case 2: $i = K+1, j = 1, \ldots, K$. It follows from (B.61) and (B.59) that

$$u'_{K+1}(t_0; e_j) = \sum_{k=1}^{K} \frac{\partial f_{K+1}^w(t_0, X(t_0))}{\partial x_k} u_k(t_0; e_j) \geq \frac{\partial f_{K+1}^w(X(u))}{\partial x_j} u_j(t_0; e_j) > 0. \tag{B.66}$$

  The first inequality follows from (B.61) that $\partial f_{K+1}^w / \partial x_k \geq 0$ and the definition of t_0 that $u_k(t_0; e_j) \geq 0$. The last inequality follows from Case 1 that $u_j(t_0; e_j) > 0$. In the proof of property (ii), we show that $F^w(t, X)$ is continuously differentiable in both t and X. Therefore, $DF^w(t, X)$ is continuous in both t and X. Since $X(t)$ and $U(t)$ are differentiable, the right-hand side of (B.59) is continuous in t. Thus, $u'_{K+1}(t; e_j)$ is continuous in t. By the continuity of $u'_{K+1}(t; e_j)$ and (B.66), there exists $0 < \epsilon_2 \leq \epsilon_1$

such that $u_{K+1}(t;e_j) > 0$ for $t \in [t_0, t_0+\epsilon_2]$. Therefore, the following holds: For $t \in (t_0, t_0+\epsilon_2]$,

$$u_{K+1}(t;e_j) = u_{K+1}(t_0;e_j) + \int_{t_0}^{t} u'_{K+1}(u;e_j)\,\mathrm{d}u > 0, \tag{B.67}$$

where the inequality follows from the assumption that $u_{K+1}(t_0;e_j) \geq 0$.

- Case 3: $i = 0, j = 1, \ldots, K$. It follows from (B.59) and (B.61) that for $t \in [t_0, t_0+\epsilon_2]$

$$u_0(t;e_j) = \frac{\partial f_0^w(t,X(t))}{\partial x_0} u_0(t;e_j) + \frac{\partial f_0^w(t,X(t))}{\partial x_{K+1}} u_{K+1}(t;e_j)\,\mathrm{d}s \geq \frac{\partial f_0^w(t,X(t))}{\partial x_0} u_0(t;e_j),$$

where the inequality follows from (B.61) and (B.67). This inequality is equivalent to (B.64) for $i = 0$. Repeating the same steps to get (B.65), we have that the following holds: For $t \in [t_0, t_0+\epsilon_2]$,

$$u_0(t;e_j) \geq u_0(t_0;e_j) \exp\left(\int_{t_0}^{t} \frac{\partial f_0^w(u,X(u))}{\partial x_0}\,\mathrm{d}u\right) \geq 0, \tag{B.68}$$

where the last inequality follows from the definition of t_0 that $u_0(t_0;e_j) \geq 0$.

- Case 4: $i = 0, j = K+1$. It follows from (B.59) and (B.61) that for $t \in [t_0, t_0+\epsilon_1]$,

$$\begin{aligned} u'_0(t;e_{K+1}) &= \frac{\partial f_0^w(t,X(t))}{\partial x_0} u_0(t;e_{K+1}) + \frac{\partial f_0^w(t,X(t))}{\partial x_{K+1}} u_{K+1}(t;e_{K+1}) \\ &\geq \frac{\partial f_0^w(t,X(t))}{\partial x_0} u_0(t;e_{K+1}), \end{aligned}$$

where the inequality follows from the discussion of Case 1 that $u_{K+1}(t;e_{K+1}) > 0$ and (B.61) that $\partial f_0^w(t;X(t))/\partial x_{K+1} \geq 0$. Repeating the same steps to obtain (B.65), we obtain the following inequality: For $t \in [t_0, t_0+\epsilon_1]$,

$$u_0(t;e_{K+1}) \geq u_0(t_0;e_{K+1}) \exp\left(\int_{t_0}^{t} \frac{\partial f_0^w(u,X(u))}{\partial x_0}\,\mathrm{d}u\right) \geq 0.$$

- Case 5: $i = 1, \ldots, K, j = 0$. It follows from (B.59) and (B.61) that for $t \in [t_0, t_0 + \epsilon_1]$,

$$u_i'(t; e_0) = \frac{\partial f_i^w(t, X(t))}{\partial x_0} u_0(t; e_0) + \frac{\partial f_i^w(t, X(t))}{\partial x_i} u_i(t; e_0) \geq \frac{\partial f_i^w(t, X(t))}{\partial x_i} u_i(t; e_0),$$

where the inequality follows from the discussion of Case 1 that $u_0(t; e_0) > 0$ and (B.61) that $\partial f_i^w(t, X(t))/\partial x_0 \geq 0$. By applying the same steps to obtain (B.68), we conclude that $u_i(t; e_0) \geq 0$ for $t \in [t_0, t_0 + \epsilon_1]$.

- Case 6: $i, j = 1, \ldots, K$ and $i \neq j$. It follows from (B.59) and (B.61) that for $t \in [t_0, t_0 + \epsilon_2]$,

$$u_i'(t; e_j) = \frac{\partial f_i^w(t, X(t))}{\partial x_0} u_0(t; e_j) + \frac{\partial f_i^w(t, X(t))}{\partial x_i} u_i(t; e_j) \geq \frac{\partial f_i^w(t, X(t))}{\partial x_i} u_i(t; e_j),$$

where the inequality follows from Case 3 that $u_0(t; e_j) \geq 0$ for $t \in [t_0, t_0 + \epsilon_2]$ and (B.61) that $\partial f_i^w(t, X(t))/\partial x_0 \geq 0$. Repeating the steps to get (B.68), we conclude that $u_i(t; e_j) \geq 0$ for $t \in [t_0, t_0 + \epsilon_2]$.

- Case 7: $i = 1, \ldots, K, j = K + 1$. it follows from (B.59) and (B.61) that for $t \in [t_0, t_0 + \epsilon_1]$,

$$\begin{aligned} u_i'(t; e_{K+1}) &= \frac{\partial f_i^w(t, X(t))}{\partial x_0} u_0(t; e_{K+1}) + \frac{\partial f_i^w(t, X(t))}{\partial x_i} u_i(t; e_{K+1}) \\ &\geq \frac{\partial f_i^w(t, X(t))}{\partial x_i} u_i(t; e_{K+1}), \end{aligned}$$

where the inequality follows from the discussion in Case 4 that $u_0(t; e_{K+1}) \geq 0$ for $t \in [t_0, t_0 + \epsilon_1]$ and (B.61) that $\partial f_i^w(t, X(t))/\partial x_0 \geq 0$. Repeating the steps to get (B.68), we conclude that $u_i(t; e_{K+1}) \geq 0$ for $t \in [t_0, t_0 + \epsilon_1]$.

- Case 8: $i = K + 1, j = 0$. It follows from (B.59) and (B.61) that for $t \in [t_0, t_0 + \epsilon_1]$,

$$u_{K+1}'(t; e_0) = \sum_{k=1}^{K} \frac{\partial f_{K+1}^w(t, X(t))}{\partial x_k} u_k(t; e_0) \geq 0,$$

where the inequality follows from (B.61) and Case 5 that $u_k(t; e_0) \geq 0$ for $t \in [t_0, t_0 + \epsilon_1]$. Integrating both sides of the equation yields the following inequality: For $t \in [t_0, t_0 + \epsilon_1]$,

$$u_{K+1}(t; e_0) = u_{K+1}(t_0; e_0) + \int_{t_0}^{t} \sum_{k=1}^{K} \frac{\partial f_{K+1}^w(u, X(u))}{\partial x_k} u_k(u; e_0)\, \mathrm{d}u \geq 0,$$

where the inequality follows from $u_{K+1}(t_0; e_0) \geq 0$.

To sum up the discussion above, we conclude that $u_i(t; e_j) \geq 0$ for $t \in [t_0, t_0 + \epsilon_2]$, $i, j = 0, 1, \ldots, K+1$. This contradicts the definition of t_0. Thus, we conclude that for $t \in [0, w]$,

$$\frac{\partial \phi_i^w(t; X_0)}{\partial x_j} = u_i(t; e_j) \geq 0, \ i, j = 0, 1, \ldots, K+1.$$

We complete the proof by showing (iv). In particular, it follows from (B.59) and (B.61) that: For $t \in [0, w]$,

$$\frac{\partial \phi_{K+1}^w(t; X_0)}{\partial x_{K+1}} = u_{K+1}(t; e_{K+1}) = 1 + \int_{0}^{t} \sum_{k=1}^{K} \frac{\partial f_{K+1}^w(u, X(u))}{\partial x_k} u_k(u; e_{K+1})\, \mathrm{d}u \geq 1,$$

where the last inequality follows from property (iv) that $u_k(u; e_{K+1}) \geq 0$. □

Fixing $w \geq 0$, we next define a function $\zeta^w(\cdot)$ implicitly using the flow ϕ^w. We will show in Appendix B.2.3 that the flow ϕ^w characterizes the evolution of the equilibrium quantities $(\hat{\beta}_W, \tilde{J}, H)$. The function $\zeta^w(\cdot)$ defined immediately below characterizes $H(w)$ in terms of $\hat{\beta}_W(w)$ and $\tilde{J}(w)$ for a fixed w; see Appendix B.2.3. To facilitate the analysis to follow, let

$$\mathcal{A} = \left(-\frac{2\theta}{\sigma^2}, \infty \right) \times \prod_{k=1}^{K} \left(r_k + c_k \frac{\sigma^2}{2\theta \rho_k} \sup_{t \geq 0} \gamma_k'(t), \ r_k \right). \tag{B.69}$$

Fix $w \geq 0$ and $(\beta, J) \in \mathcal{A}$, then $\zeta^w(\beta, J)$ is the value such that the following holds:

$$\phi^w_{K+1}(w; X_0) = 0, \text{ where } X_0 = (\beta, J, \zeta^w(\beta, J)). \tag{B.70}$$

The following lemma provides some useful properties of $\zeta^w(\cdot)$.

Lemma 50. *The function $\zeta^w : \mathcal{A} \to \mathbb{R}$ is well-defined for $w \geq 0$. That is, there exists a unique value $\zeta^w(\beta, J)$ satisfying* (B.70) *for $w \geq 0$ and $(\beta, J) \in \mathcal{A}$. In addition, ζ^w is continuously differentiable for $w \geq 0$.*

Proof. Fix $w \geq 0$ and $(\beta, J) \in \mathcal{A}$. We need to show that there exists a unique η such that $\zeta^w(\beta, J) = \eta$ satisfies (B.70). We first show that there exists a such η. It follows from Lemma 45 that for $t \in [0, w]$

$$\phi^w_k(t; X_0) \in \left(r_k + c_k \frac{\sigma^2}{2\theta\rho_k} \sup_{t \geq 0} \gamma'_k(t),\ r_k \right),\ k = 1, \ldots, K.$$

By substituting $\phi^w_k(t; X_0) \leq r_k$ into equation (B.52), we obtain the following: For any $X_0 = (x_0, x_1, \ldots, x_{K+1}) \in \mathcal{I}$,

$$\begin{aligned} \phi^w_{K+1}(w; X_0) &\leq x_{K+1}(0) - \int_0^w \sum_{k=1}^K \frac{\gamma'_i(w-t)}{2(r_k)^\delta} \, \mathrm{d}t \\ &= x_{K+1}(0) - \int_0^w \sum_{k=1}^K \frac{\gamma'_k(t)}{2(r_k)^\delta} \, \mathrm{d}t = \phi^w_{K+1}(0) - \underline{H}(w), \end{aligned}$$

where $\underline{H}(w)$ is given in (4.55). Similarly, by substituting $\phi^w_k(t; X_0) \geq r_k + c_k\sigma^2 \sup_{s \geq 0} \gamma'_k(s)/2\theta\rho_k$ into equation (B.52), we obtain that $\phi^w_{K+1}(w; X_0) \geq x_{K+1}(0) - \bar{H}(w)$ where $\bar{H}(w)$ is given in (4.56). In sum, the following holds: For any $X_0 = (x_0, x_1, \ldots, x_{K+1}) \in \mathcal{I}$,

$$x_{K+1}(0) - \bar{H}(w) \leq \phi^w_{K+1}(w; X_0) \leq x_{K+1}(0) - \underline{H}(w).$$

Thus, by substituting $X_0 = (\beta, J, \bar{H}(w))$ and $X_0 = (\beta, J, \underline{H}(w))$ into this inequality, we

obtain the following two inequalities, respectively:

$$\phi^w_{K+1}\left(w;(\beta,J,\bar{H}(w))\right)\geq 0 \text{ and } \phi^w_{K+1}\left(w;(\beta,J,\underline{H}(w))\right)\leq 0.$$

By the continuity of ϕ^w_{K+1}, there exists η such that

$$\phi^w_{K+1}(w;(\beta,J,\eta))=0,\ \eta\in\left[\underline{H}(w),\ \bar{H}(w)\right]. \tag{B.71}$$

It follows from property (iv) of Lemma 49 that $\phi^w_{K+1}(w;(\beta,J,\eta))$ is strictly increasing in η. Therefore, such η satisfying (B.71) is unique. Thus, $\zeta^w(\cdot)$ is well-defined.

Since $\phi^w_{K+1}(w;(\beta,J,\eta))$ is continuously differentiable by Lemma 49, $\zeta^w(\beta,J)$ is continuously differentiable by the implicit function theorem; see Theorem 9.28 in Rudin [96]. □

The rest of this section provides auxiliary properties of the function $\zeta^w(\cdot)$. The following lemma provides equations that characterize the partial derivatives of $\zeta^w(\cdot)$ in terms of the partial derivatives of the flow ϕ^w.

Lemma 51. *Fix $w\geq 0$ and $(\beta,J)\in\mathcal{A}$. Let $X_0=(\beta,J,\zeta^w(\beta,J))$. Then the following*

holds: For $k = 1, \ldots, K$ *and* $s \in [0, w]$,

$$\frac{\partial \zeta^w(\beta, J)}{\partial \beta}$$
$$= -\frac{\dfrac{\partial \zeta^{w-s}(\beta(s), J(s))}{\partial \beta}\dfrac{\partial \phi_0^w(s; X_0)}{\partial x_0} + \displaystyle\sum_{i=1}^{K} \dfrac{\partial \zeta^{w-s}(\beta(s), J(s))}{\partial J_i}\dfrac{\partial \phi_i^w(s; X_0)}{\partial x_0} - \dfrac{\partial \phi_{K+1}^w(s; X_0)}{\partial x_0}}{\dfrac{\partial \zeta^{w-s}(\beta(s), J(s))}{\partial \beta}\dfrac{\partial \phi_0^w(s; X_0)}{\partial x_{K+1}} + \displaystyle\sum_{i=1}^{K} \dfrac{\partial \zeta^{w-s}(\beta(s), J(s))}{\partial J_i}\dfrac{\partial \phi_i^w(s; X_0)}{\partial x_{K+1}} - \dfrac{\partial \phi_{K+1}^w(s; X_0)}{\partial x_{K+1}}}, \tag{B.72}$$

$$\frac{\partial \zeta^w(\beta, J)}{\partial J_k}$$
$$= -\frac{\dfrac{\partial \zeta^{w-s}(\beta(s), J(s))}{\partial \beta}\dfrac{\partial \phi_0^w(s; X_0)}{\partial x_k} + \displaystyle\sum_{i=1}^{K} \dfrac{\partial \zeta^{w-s}(\beta(s), J(s))}{\partial J_i}\dfrac{\partial \phi_i^w(s; X_0)}{\partial x_k} - \dfrac{\partial \phi_{K+1}^w(s; X_0)}{\partial x_k}}{\dfrac{\partial \zeta_{w-s}(\beta(s), J(s))}{\partial \beta}\dfrac{\partial \phi_0^w(s; X_0)}{\partial x_{K+1}} + \displaystyle\sum_{i=1}^{K} \dfrac{\partial \zeta^{w-s}(\beta(s), J(s))}{\partial J_k}\dfrac{\partial \phi_k(s; X_0)}{\partial x_{K+1}} - \dfrac{\partial \phi_{K+1}^w(s; X_0)}{\partial x_{K+1}}}, \tag{B.73}$$

where $\beta(s) = \phi_0^w(s; X_0)$ *and* $J(s) = (\phi_1(s; X_0), \ldots, \phi_K(s; X_0))$.

Proof. Fix $w \geq 0$ and $(\beta, J) \in \mathcal{A}$. Note that for any $s \in [0, w]$, the function $\phi^w(t + s; X_0)$, $t \in [0, w - s]$ satisfies (B.53) for $w - s$ with initial value $\phi^w(s; X_0)$. Thus, it follows from Lemma 46 that for $s \in [0, w]$,

$$\phi^w(t + s; X_0) = \phi^{w-s}(t; \phi^w(s; X_0)),\ t \in [0, w - s].$$

In particular, letting $t = w - s$, we have that $\phi^w(w; X_0) = \phi^{w-s}(w - s; \phi^w(s; X_0))$. In addition, it follows from (B.70) and the definition of X_0 that

$$0 = \phi_{K+1}^w(w; X_0) = \phi_{K+1}^{w-s}(w - s; \phi^w(s; X_0)).$$

This equation shows that $\phi^w_{K+1}(s; X_0)$ satisfies (B.70) for $w - s$ and $(\beta(s), J(s))$. Thus, it follows from Lemma 50 that

$$\zeta^{w-s}(\beta(s), J(s)) = \phi^w_{K+1}(s).$$

To be more specific, we substitute $\beta(s)$, $J(s)$ and X_0 into this equation and obtain the following:

$$\zeta^{w-s}(\phi^w_0(s; (\beta, J, \zeta^w(\beta, J))), \cdots, \phi^w_K(s; (\beta, J, \zeta^w(\beta, J)))) = \phi^w_{K+1}(s; (\beta, J, \zeta^w(\beta, J))). \tag{B.74}$$

Taking the partial derivative of both sides of equation (B.74) with respect to β yields the following equation:

$$\begin{aligned}
&\frac{\partial \zeta^{w-s}(\beta(s), J(s))}{\partial \beta} \left(\frac{\partial \phi^w_0(s; X_0)}{\partial x_0} + \frac{\partial \phi^w_0(s; X_0)}{\partial x_{K+1}} \frac{\partial \zeta^w(\beta, J)}{\partial \beta} \right) \\
&+ \sum_{i=1}^{K} \frac{\partial \zeta^{w-s}(\beta(s), J(s))}{\partial J_i} \left(\frac{\partial \phi^w_i(s; X_0)}{\partial x_0} + \frac{\partial \phi^w_i(s; X_0)}{\partial x_{K+1}} \frac{\partial \zeta^w(\beta, J)}{\partial \beta} \right) \\
&= \frac{\partial \phi^w_{K+1}(s; X_0)}{\partial x_0} + \frac{\partial \phi^w_{K+1}(s; X_0)}{\partial x_{K+1}} \frac{\partial \zeta^w(\beta, J)}{\partial \beta}.
\end{aligned}$$

Rearranging the terms, we obtain equation (B.72). Similarly, we can take the partial derivative of both sides of equation (B.74) with respect to J_k and rearrange the terms to obtain equation (B.73) for $k = 1, \ldots K$. □

The following lemma shows the sign of the partial derivatives of $\zeta^w(\cdot)$.

Lemma 52. *The following inequalities hold: For $w \geq 0$ and $(\beta, J) \in \mathcal{A}$,*

$$\frac{\partial \zeta^w(\beta, J)}{\partial \beta} \leq 0 \quad \textit{and} \quad \frac{\partial \zeta^w(\beta, J)}{\partial J_k} \leq 0, \ k = 1, \ldots, K.$$

Proof. Note that $\zeta^0(\beta, J) = 0$ for any $(\beta, J) \in \mathcal{A}$. Thus the following holds:

$$\frac{\partial \zeta^0(\beta, J)}{\partial \beta} = \frac{\partial \zeta^0(\beta, J)}{\partial J_k} = 0, \ k = 1, \ldots, K.$$

Substituting $s = w$ in equations (B.72) and (B.73), we obtain that for $w \geq 0$ and $k = 1, \ldots, K$,

$$\frac{\partial \zeta^w(\beta, J)}{\partial \beta} = -\frac{\partial \phi^w_{K+1}(w; X_0)/\partial x_0}{\partial \phi^w_{K+1}(w; X_0)/\partial x_{K+1}} \text{ and } \frac{\partial \zeta^w(\beta, J)}{\partial J_k} = -\frac{\partial \phi^w_{K+1}(w; X_0)/\partial x_k}{\partial \phi^w_{K+1}(w; X_0)/\partial x_{K+1}},$$

where $X_0 = (\beta, J, \zeta^s(\beta, J))$. It follows from Lemma 49 that

$$\frac{\partial \phi^w_{K+1}(w; X_0)}{\partial x_i} \geq 0, \ i = 0, 1, \ldots, K+1.$$

Thus, we have that for $w \geq 0$ and $k = 1, \ldots, K$,

$$\frac{\partial \zeta^w(\beta, J)}{\partial \beta} \leq 0 \text{ and } \frac{\partial \zeta^w(\beta, J)}{\partial J_k} \leq 0.$$

□

B.2.2 Properties of $\zeta^w(\cdot)$ on a Restricted Domain

In this section, we restrict our focus on $(\beta, J) \in \mathcal{A}(w)$, where $\mathcal{A}(w)$ is given in (4.61)[6]. And, we provide further auxiliary properties of $\zeta^w(\beta, J)$ for $(\beta, J) \in \mathcal{A}(w)$. The following lemma shows that the solution $\phi^w(\cdot; X_0)$ of (B.53) (which corresponds to the time-reversed equilibrium quantities as we will show in Appendix B.2.3) lives in a set characterized (in part) by the collection of sets $\mathcal{A}(s)$ for $s \in [0, w]$.

6. It follows from (4.60) that $\mathcal{A}(w) \subseteq \mathcal{A}$.

Lemma 53. *If $(\beta, J) \in \mathcal{A}(w) \subseteq \mathcal{A}$, then the following holds:*

$$\phi^w(s; X_0) \in \mathcal{A}(w-s) \times [\underline{H}(w-s), \bar{H}(w-s)],\ s \in [0, w],$$

where $X_0 = (\beta, J, \zeta^w(\beta, J))$, $\underline{H}(w-s)$ and $\bar{H}(w-s)$ are given by (4.55) *and* (4.56)*, respectively.*

Proof. Fix w and $(\beta, J) \in \mathcal{A}(w)$. For notational brevity, let $x_i(s) = \phi_i^w(s; X_0)$, $s \in [0, w]$ and $i = 0, \ldots, K+1$.

We first show that $x_{K+1}(s) \in [\underline{H}(w-s), \bar{H}(w-s)]$. It follows from (B.52) that for $s \in [0, w]$,

$$\begin{aligned}
x_{K+1}(w) &= x_{K+1}(s) - \int_s^w \sum_{k=1}^K \frac{\gamma_k'(w-u)}{2(x_k(u))^\delta}\,\mathrm{d}u \\
&\leq x_{K+1}(s) - \int_s^w \sum_{k=1}^K \frac{\gamma_k'(w-u)}{2r_k^\delta}\,\mathrm{d}u \\
&= x_{K+1}(s) - \int_0^{w-s} \sum_{k=1}^K \frac{\gamma_k'(u)}{2r_k^\delta}\,\mathrm{d}u \\
&= x_{K+1}(s) - \underline{H}(w-s),
\end{aligned}$$

where the inequality follows from Lemma 45 that $x_k(u) \leq r_k$. Similarly, the following holds:

$$x_{K+1}(w) \geq x_{K+1}(s) - \bar{H}(w-s),\ s \in [0, w].$$

It follows from (B.70) that $x_{K+1}(w) = 0$. Substituting $x_{K+1}(w) = 0$ into the left-hand side of these two inequalities yields that $x_{K+1}(s) \in [\underline{H}(w-s), \bar{H}(w-s)]$.

Next, we show that $x_0(s) \in [\underline{\beta}(w-s; w), \bar{\beta}(w-s; w)]$. Recall from Lemma 49 that $x_{K+1}(s) \geq 0$ for $s \in [0, w]$. In addition, it follows from Lemma 45 that $x_0(s) > -2\theta/\sigma^2 > 0$

for $s \in [0, w]$. Thus, it follows from equation (B.54) that for $s \in [0, w]$,

$$\begin{aligned}
\frac{1}{x_0(s)} &= \frac{1}{x_0(0)} \exp\left[\int_0^s \frac{2}{\sigma^2}(\theta - x_{K+1}(t))\,\mathrm{d}t\right] + \int_0^s \exp\left[\int_u^s \frac{2}{\sigma^2}(\theta - x_{K+1}(t))\,\mathrm{d}t\right]\,\mathrm{d}u \\
&\leq \left(\int_w^\infty \exp\left[\int_w^u \frac{2}{\sigma^2}(\theta - \underline{H}(t))\,\mathrm{d}t\right]\,\mathrm{d}u\right) \exp\left[\int_0^s \frac{2}{\sigma^2}(\theta - \underline{H}(w-t))\,\mathrm{d}t\right] \\
&\quad + \int_0^s \exp\left[\int_u^s \frac{2}{\sigma^2}(\theta - \underline{H}(w-t))\,\mathrm{d}t\right]\,\mathrm{d}u \\
&= \int_w^\infty \exp\left[\int_{w-s}^u \frac{2}{\sigma^2}(\theta - \underline{H}(t))\,\mathrm{d}t\right]\,\mathrm{d}u + \int_{w-s}^w \exp\left[\int_{w-s}^u \frac{2}{\sigma^2}(\theta - \underline{H}(t))\,\mathrm{d}t\right]\,\mathrm{d}u \\
&= \int_{w-s}^\infty \exp\left[\int_{w-s}^u \frac{2}{\sigma^2}(\theta - \underline{H}(t))\,\mathrm{d}t\right]\,\mathrm{d}u = \frac{1}{\underline{\beta}(w-s)},
\end{aligned}$$

where the inequality follows from the assumption that $x_0(0) = \beta \geq \underline{\beta}(w)$ and the inequality that $x_{K+1}(s) \geq \underline{H}(w-s)$. Thus, we have that $x_0(s) \geq \underline{\beta}(w-s)$ for $s \in [0, w]$. Similarly, by substituting $x_0(0) \leq \bar{\beta}(w)$ and $x_{K+1}(s) \leq \bar{H}(w-s)$ into (B.54), we have that for $s \in [0, w]$,

$$\frac{1}{x_0(s)} \geq \frac{1}{\bar{\beta}(w-s)}.$$

Thus, we have that $x_0(s) \leq \bar{\beta}(w-s)$. In sum, $x_0(s) \in [\underline{\beta}(w-s), \bar{\beta}(w-s)]$.

We end the proof by showing that $x_k(s) \in [\underline{J}_k(w-s), r_k]$ for $k = 1, \ldots, K$. It follows

from equation (B.55) that for $s \in [0, w]$ and $k = 1, \ldots, K$,

$$
\begin{aligned}
& x_k(s) \\
&= x_k(0) \exp\left(\int_0^s -x_0(u)\,\mathrm{d}u\right) + \int_0^s \exp\left(\int_u^s -x_0(t)\,\mathrm{d}t\right)\left(r_k x_0(u) - c_k \frac{\gamma_k'(w-u)}{\rho_k}\right)\mathrm{d}u \\
&= x_k(0) \exp\left(\int_0^s -x_0(u)\,\mathrm{d}u\right) + r_k \int_0^s \exp\left(\int_u^s -x_0(t)\,\mathrm{d}t\right) x_0(u)\,\mathrm{d}u \\
&\quad - c_k \int_0^s \exp\left(\int_u^s -x_0(t)\,\mathrm{d}t\right) \frac{\gamma_k'(w-u)}{\rho_k}\,\mathrm{d}u \\
&\geq x_k(0) \exp\left(\int_0^s -x_0(u)\,\mathrm{d}u\right) + r_k\left[1 - \exp\left(\int_0^s -x_0(u)\,\mathrm{d}u\right)\right] \\
&\quad - c_k \int_0^s \frac{x_0(u)}{\underline{\beta}(w-s)} \exp\left(\int_u^s -x_0(t)\,\mathrm{d}t\right) \frac{\gamma_k'(w-u)}{\rho_k}\,\mathrm{d}u \\
&= x_k(0) \exp\left(\int_0^s -x_0(u)\,\mathrm{d}u\right) + \left(r_k - \frac{c_k}{\underline{\beta}(w-s)\rho_k} \sup_{t\geq 0} \gamma_k'(t)\right)\left[1 - \exp\left(\int_0^s -x_0(u)\,\mathrm{d}u\right)\right] \\
&\geq r_k - \frac{c_k}{\underline{\beta}(w-s)\rho_k} \sup_{t\geq 0} \gamma_k'(t) = \underline{J}_k(w-s).
\end{aligned}
$$

The first inequality follows from $x_0(t) \geq \underline{\beta}(w-t) \geq \underline{\beta}(w-s)$ for $t \in [0, s]$. The last inequality follows from the assumption that $x_k(0) = J_k \geq \underline{J}_k(w) \geq \underline{J}_k(w-s)$ for $k = 1, \ldots, K$. In addition, it is immediate from Lemma 45 that $x_k(s) \leq r_k$. Thus, we have that $x_k(s) \in [\underline{J}_k(w-s), r_k]$ for $k = 1, \ldots, K$. □

Lemma 53 gives upper and lower bounds of the flow $\phi^w(s; X_0)$. An important property of the bounds in Lemma 53 is that they only depend on the value of $w-s$. In particular, the first K dimensions of the flow stay in the restricted domain $\mathcal{A}(w-s)$. The proof of the following two lemmas relies on this property. Lemmas 54 provides an intermediate result needed for the proof of Lemma 55, which in turn provides the bounds on the partial derivatives of ϕ^w.

Lemma 54. *There exists $w_1 \geq 0$ such that the following holds: For $w, s \geq 0$ and $(\beta, J) \in \mathcal{A}(w)$, if $w - s \geq w_1$, then*

$$
\frac{3}{2}\phi_0^w(s; X_0) + \frac{2}{\sigma^2}\left(\theta - \phi_{K+1}^w(s; X_0)\right) \geq 0, \tag{B.75}
$$

where $X_0 = (\beta, J, \zeta^w(\beta, J))$.

Proof. To facilitate the proof, we first derive the inequalities (B.76)-(B.78). Note that $\underline{J}_k(w) \to r_k$ as $w \to \infty$. In addition, it follows from Assumption 4 that $\gamma_k'(w) \to \bar{\gamma}_k'(w) \to \bar{\gamma}_k$ as $w \to \infty$ for $k = 1, \dots K$. Thus, there exists t_1 such that for $w \geq t_1$ and $k = 1, \dots, K$,

$$\frac{\gamma_k'(w)}{(\underline{J}_k(w))^\delta} \leq \frac{7}{6}\frac{\bar{\gamma}_k'}{r_k^\delta} \quad \text{and} \quad \gamma_k'(w) \geq \frac{4}{5}\bar{\gamma}_k. \tag{B.76}$$

It follows from the second inequality in (B.76) that for $w \geq t_1$,

$$\underline{H}(w) = \underline{H}(t_1) + \int_{t_1}^{w} \sum_{i=1}^{K} \frac{\gamma_i'(s)}{2r_i^\delta}\, \mathrm{d}s \geq \underline{H}(t_1) + \frac{4}{5}\sum_{i=1}^{K} \frac{\bar{\gamma}_i'}{2r_i^\delta}(w - t_1). \tag{B.77}$$

In addition, there exists $t_2 \geq t_1$ such that for $w \geq t_2$, the following holds:

$$\underline{H}(t_1) - \bar{H}(t_1) + \frac{1}{30}\sum_{i=1}^{K} \frac{\bar{\gamma}_i'}{2r_i^\delta}(w - t_1) \geq 0 \tag{B.78}$$

The first two terms in the left-hand side of (B.78) are constants. The last goes to infinity as $w \to \infty$. Thus, inequality (B.78) holds for w large enough.

Now fixing $w \geq t_2$ and $(\beta, J) \in \mathcal{A}(w)$, we show that equation (B.75) holds. It follows

from equation (B.52) and Lemma 53 that for $s \geq 0$ such that $w - s \geq t_2$,

$$\begin{aligned}
\phi^w_{K+1}(s; X_0) &= \phi^w_{K+1}(w; X_0) + \int_s^w \sum_{i=1}^K \frac{\gamma_i'(w-u)}{2(\phi_i^w(u; X_0))^\delta} \, \mathrm{d}u \\
&\leq \int_0^{w-s} \sum_{i=1}^K \frac{\gamma_i'(w-u)}{2(\underline{J}_i(w-u))^\delta} \, \mathrm{d}u \\
&= \int_0^{w-s} \sum_{i=1}^K \frac{\gamma_i'(u)}{2(\underline{J}_i(u))^\delta} \, \mathrm{d}u \\
&\leq \int_0^{t_1} \sum_{i=1}^K \frac{\gamma_i'(u)}{2(r_i)^\delta} \, \mathrm{d}u + \int_{t_1}^{w-s} \sum_{i=1}^K \frac{\gamma_i'(u)}{2(\underline{J}_i(u))^\delta} \, \mathrm{d}u \\
&= \bar{H}(t_1) + \int_{t_1}^{w-s} \sum_{i=1}^K \frac{\gamma_i'(u)}{2(\underline{J}_i(u))^\delta} \, \mathrm{d}u \\
&\leq \bar{H}(t_1) + \frac{7}{6} \frac{\bar{\gamma}_k'}{2r_k^\delta} (w - s - t_1).
\end{aligned} \tag{B.79}$$

The first inequality follows from Lemma 53 that $\phi_i^w(u; X_0) \geq \underline{J}_i(w-u)$ and (B.70) that $\phi^w_{K+1}(w; X_0) = 0$. The second inequality follows from $\underline{J}_i(u) \leq r_i$. The last inequality follows from equation (B.76). Since $\underline{H}(w)$ is increasing in w, the following holds: For $w \geq t_2$,

$$\begin{aligned}
\underline{\beta}(w) &= \left(\int_w^\infty \exp\left[\int_w^s \frac{2}{\sigma^2} (\theta - \underline{H}(u)) \, \mathrm{d}u \right] \mathrm{d}s) \right)^{-1} \\
&\geq \left(\int_w^\infty \exp\left[\int_w^s \frac{2}{\sigma^2} (\theta - \underline{H}(w)) \, \mathrm{d}u \right] \mathrm{d}s) \right)^{-1} \\
&= -\frac{2}{\sigma^2} (\theta - \underline{H}(w)) \\
&\geq -\frac{2}{\sigma^2} \left(\theta - \underline{H}(t_1) - \frac{4}{5} \sum_{i=1}^K \frac{\bar{\gamma}_i'}{2r_i^\delta} (w - t_1) \right),
\end{aligned} \tag{B.80}$$

where the last inequality follows from equation (B.77). Thus, it follows from equations

(B.79)-(B.80) that for $s \geq 0$ such that $w - s \geq t_2$,

$$\begin{aligned}
&\frac{3}{2}\phi_0^w(s;X_0) + \frac{2}{\sigma^2}\left(\theta - \phi_{K+1}^w(s;X_0)\right) \\
&\geq \frac{3}{2}\underline{\beta}(w-s) + \frac{2}{\sigma^2}\left(\theta - \phi_{K+1}^w(s;X_0)\right) \\
&\geq -\frac{3}{\sigma^2}\left(\theta - \underline{H}(t_1) - \frac{4}{5}\sum_{i=1}^{K}\frac{\bar{\gamma}_i'}{2r_i^\delta}(w-s-t_1)\right) + \frac{2}{\sigma^2}\left(\theta - \bar{H}(t_1) - \frac{7}{6}\frac{\bar{\gamma}_k'}{2r_k^\delta}(w-s-t_1)\right) \\
&= -\frac{1}{\sigma^2}(\theta - \underline{H}(w_1)) + \frac{2}{\sigma^2}\left(\underline{H}(t_1) - \bar{H}(t_1) + \sum_{i=1}^{K}\frac{1}{30}\frac{1}{2r_i^\delta}(w-s-t_1)\right) \geq 0.
\end{aligned}$$

The first inequality follows from Lemma 53. The second inequality follows from (B.79)-(B.80). The last inequality follows from inequality (B.78) and $\theta - \underline{H}_1(w) \leq 0$. Letting $w_1 = t_2$ completes the proof. □

Lemma 55. *For any $\epsilon \in (0,1)$, there exists w_2 such that the following holds: For $(\beta, J) \in \mathcal{A}(w)$, $w \geq w_2$, $s \in [0, s_1]$,*

$$\frac{\partial \phi_i^w(s;X_0)}{\partial x_i} \leq 1 - \frac{s}{4},\ i = 0,1,\ldots,K, \tag{B.81}$$

$$\frac{\partial \phi_i^w(s;X_0)}{\partial x_0} \leq \epsilon,\ i = 1,\ldots,K, \tag{B.82}$$

$$\frac{\partial \phi_i^w(s;X_0)}{\partial x_j} \leq \epsilon,\ i,j = 1,\ldots,K \quad \textit{and} \quad i \neq j, \tag{B.83}$$

$$\frac{\partial \phi_0^w(s;X_0)}{\partial x_j} \leq M_1,\ j = 1,\ldots,K, \tag{B.84}$$

$$\frac{\partial \phi_{K+1}^w(s;X_0)}{\partial x_j} \leq M_1,\ j = 0,1,\ldots,K, \tag{B.85}$$

where $X_0 = (\beta, J, \zeta^w(\beta, J))$ and

$$c_1 = \sum_{k=1}^{K}\frac{\delta\bar{\gamma}_k'}{r_k^{\delta+1}},\ s_1 = \min\left\{1, \frac{\sigma}{2\sqrt{c_1}}, \frac{\sigma^2}{8c_1}\right\} \quad \textit{and} \quad M_1 = \max\left\{c_1, \frac{4c_1}{\sigma^2}\right\}.$$

Proof. Fix $w \geq 0$ and $(\beta, J) \in \mathcal{A}(w)$. In addition, let $X(s) = \phi^w(s;X_0)$ for $s \in [0,w]$ and

$x_i(s)$ denote the i^{th} component of $X(s)$ for $i = 0, 1, \ldots, K+1$.

We first derive the inequalities (B.86)-(B.88) to facilitate the proof. Define the function f_x as follows: For $s \geq 0$,

$$f_x(s) = -2x_0(s) - \frac{2}{\sigma^2}(\theta - x_{K+1}(s)).$$

Let w_1 be a constant such that the following holds: For $w \geq w_1$ and $s \in [0, s_1]$,

$$\sum_{k=1}^{K} \frac{\delta\gamma_k'(w-s)}{2(\underline{J}_k(w-s))^{\delta+1}} \leq c_1, \;\; x_0(s) \leq -2f_x(s) \;\; \text{and} \;\; x_0(s) \geq 2. \tag{B.86}$$

The first inequality follows from $\gamma_k'(t) \to \bar{\gamma}_k'$ (see Assumption 4) and $\underline{J}_k(t) \to r_k$ as $t \to \infty$. Therefore, the following holds for w large enough:

$$\sum_{k=1}^{K} \frac{\delta\gamma_k'(w-s)}{2(\underline{J}_k(w-s))^{\delta+1}} \leq 2\sum_{k=1}^{K} \frac{\delta\bar{\gamma}_k'}{2r_k^{\delta+1}} = c_1.$$

The second inequality in (B.86) follows from Lemma 54. To be more specific, multiplying both sides of (B.75) by 2 yields the following:

$$3x_0(s) + \frac{4}{\sigma^2}(\theta - x_{K+1}(s)) = -x_0(s) - 2f_x(s) \geq 0$$

for $w - s$ large enough. Rearranging the terms yield the second inequality in (B.86). In addition, it follows from Lemma 53 that $x_0(s) \geq \underline{\beta}(w-s) \geq \underline{\beta}(w-s_1) \to \infty$ as $w \to \infty$. Thus, the third inequality in (B.86) holds if w is large enough. Since s is bounded above by s_1, we can find such w_1 large enough such that (B.86) holds. In addition, the following holds: For $s \in [0, s_1]$,

$$\frac{c_1 s^2}{\sigma^2} < \frac{1}{4} \;\; \text{and} \;\; \exp(-s) \leq 1 - \frac{s}{2}. \tag{B.87}$$

Both inequalities follow from the definition of s_1. In particular, the second inequality holds

for $s \leq 1$. Note also that the following holds: For $k = 1, \ldots, K$,

$$0 \leq r_k - x_k(s) \leq r_k - \underline{J}_k(w - s) \leq r_k - \underline{J}_k(w - s_1),$$

where the first two inequalities follows from Lemma 53 and the third one follows because $\underline{J}_k(w)$ is increasing in w. Note that $r_k - \underline{J}_k(w - s_1) \to 0$ as $w \to \infty$. Thus, for any given $\epsilon \in (0, 1)$, there exists $w_2 \geq w_1$ such that the following holds: For $w \geq w_2$, $s \in [0, s_1]$ and $k = 1, \ldots, K$,

$$r_k - x_k(s) < \frac{\epsilon}{2}. \tag{B.88}$$

Let $U(t; e_j) = (u_0(t; e_j), u_1(t; e_j), \ldots, u_{K+1}(t; e_j))$, $t \in [0, w]$, be the solution to (B.59) with $U_0 = e_j$ for $j = 0, 1, \ldots, K + 1$. Recall from (B.60) that for $i, j = 0, \ldots, K + 1$,

$$\frac{\partial \phi_i^w(s; X_0)}{\partial x_j} = u_i(s; e_j).$$

Thus, we show (B.81)-(B.85) using $u_i(s; e_j)$ for $i, j = 0, 1, \ldots, K$.

We first show (B.81) for $i = 0$, (B.82) and (B.85) for $j = 0$. In particular, we fix $w \geq w_2$ and show that for $s \leq s_1$,

$$u_0(s; e_0) \leq 1 - \frac{s}{4},\ u_j(s; e_0) \leq \epsilon,\ u_{K+1}(s; e_0) \leq M_1.$$

First, we show that $u_0(s; e_0) \leq 1$ for $s \in [0, s_1]$ by contradiction. Suppose this is not true. Define s_2 as follows:

$$s_2 = \inf\{s \in [0, s_1] : u_0(s; e_0) > 1\}.$$

By assumption, we have that $s_2 < s_1$. Note that $u_0(0; e_0) = 1$. In addition, it follows from (B.50) and (B.59) that

$$u_0'(0; e_0) = \left(-2x_0(0) - \frac{2}{\sigma^2}(\theta - x_{K+1}(0))\right) = 2f_x(0) \leq -x_0(0) \leq -2 < 0,$$

where the first equality follows from $u_0(0;e_0) = 1$ and $u_i(0;e_0) = 0$ for $i = 1,\ldots,K$ and the inequalities follow from (B.86). Recall that $u_0'(s;e_0)$ is continuous in s because the right-hand side of (B.59) is continuous. There exists $\epsilon_1 > 0$ such that $u_0'(s;e_0) < 1$. It is immediate that $u_0(s;e_0) \le u_0(0;e_0) = 1$ for $s \in [0,\epsilon_1]$. Therefore, $s_2 > 0$. By definition of s_2 and continuity of $u_0(s;e_0)$, it follows that $u_0(s;e_0) \le 1$ for $s \le s_2$ and $u_0(s_2;e_0) = 1$. We now show the contradiction by showing that $u_0(s_2;e_0) < 1$. It follows from (B.51) and (B.59) that for $s \in [0,s_2]$ and $k = 1,\ldots,K$,

$$u_k'(s;e_0) = (r_k - x_k(s))u_0(s;e_0) - x_0(s)u_k(s;e_0)$$

with $u_k(0;e_0) = 0$. Solving this ODE, we obtain that for $s \in [0,s_2]$ and $j = 1,\ldots,K$,

$$\begin{aligned} u_k(s;e_0) &= \int_0^s \exp\left(-\int_u^s x_0(t)\,\mathrm{d}t\right)[r_k - x_k(u)]u_0(u;e_0)\,\mathrm{d}u \\ &\le \int_0^s \exp\left(-\int_u^s x_0(t)\,\mathrm{d}t\right)\frac{\epsilon}{2}\,\mathrm{d}u \le \frac{\epsilon s}{2}, \end{aligned} \tag{B.89}$$

where the first inequality follows from (B.88) that $r_k - x_k(s) \le \epsilon/2$ and $u_0(s;e_0) \le 1$ for $s \le s_2$. It follows from (B.52) and (B.59) that for $s \in [0,s_2]$,

$$\begin{aligned} u_{K+1}(s;e_0) &= \int_0^s \sum_{j=1}^K \frac{\delta\gamma_j'(w-u)}{2(x_j(u))^{\delta+1}}u_j(u;e_0)\,\mathrm{d}u \\ &\le \int_0^s \sum_{j=1}^K \frac{\delta\gamma_j'(w-u)}{2(\underline{J}_j(w-u))^{\delta+1}}u_j(u;e_0)\,\mathrm{d}u \\ &\le c_1\int_0^s \frac{\epsilon}{2}u\,\mathrm{d}u \le \frac{c_1\epsilon s^2}{2}. \end{aligned} \tag{B.90}$$

The inequality in the second line follows from Lemma 53 that $x_j(u) \ge \underline{J}_j(w-u)$ for $j = 1,\ldots,K$. The first inequality in the third line follows from equations (B.86) and (B.89). In

addition, it follows from (B.50) and (B.59) that

$$\begin{aligned} u_0'(s;e_0) &= \left[-2x_0(s) - \frac{2}{\sigma^2}(\theta - x_{K+1}(s))\right] u_0(s;e_0) + \frac{2}{\sigma^2}x_0(s)u_{K+1}(s;e_0) \\ &= f_x(s)u_0(s;e_0) + \frac{2}{\sigma^2}x_0(s)u_{K+1}(s;e_0) \end{aligned}$$

with $u_0(0;e_0) = 1$. Solving this ODE yields the following: For $s \in (0, s_2]$,

$$\begin{aligned} u_0(s;e_0) &= \exp\left(\int_0^s f_x(u)\,\mathrm{d}u\right) + \int_0^s \exp\left(\int_u^s f_x(t)\,\mathrm{d}t\right)\frac{2}{\sigma^2}x_0(u)u_{K+1}(u;e_0)\,\mathrm{d}u \\ &\le \exp\left(\int_0^s f_x(u)\,\mathrm{d}u\right) + \int_0^s \exp\left(\int_u^s f_x(t)\,\mathrm{d}t\right)\frac{2}{\sigma^2}(-2f_x(u))\frac{c_1\epsilon u^2}{2}\,\mathrm{d}u \\ &\le \exp\left(\int_0^s f_x(u)\,\mathrm{d}u\right) + \frac{1}{2}\int_0^s \exp\left(\int_u^s f_x(t)\,\mathrm{d}t\right)(-f_x(u))\,\mathrm{d}u \\ &= \frac{1}{2}\left[1 + \exp\left(\int_0^s f_x(u)\,\mathrm{d}u\right)\right] \\ &\le \frac{1}{2}[1 + \exp(-s)] < 1. \end{aligned} \tag{B.91}$$

The the first inequality follows from (B.86) that $x_0(s) \le -2f_x(s)$ and (B.90). The second inequality follows from (B.87) that $c_1u^2/\sigma^2 < 1/4$ and $\epsilon < 1$. The third inequality follows (B.86) that $f_x(s) \le -x_0(s)/2 \le -1$. The last inequality follows from $s > 0$. In particular, $u_0(s_2;e_0) < 1$. This contradicts the definition of s_2 and continuity of u_0, i.e. $u_0(s_2;e_0) = 1$. Therefore, we have that $s_2 = s_1$. In other words, $u_0(s;e_0) \le 1$ for $s \le s_1$.

Note that (B.89)-(B.91) holds for $s \in [0, s_1]$. Thus, it follows from (B.87) and (B.91) that for $s \in [0, s_1]$,

$$u_0(s;e_0) \le \frac{1}{2}[1 + \exp(-s)] \le \frac{1}{2}\left[1 + 1 - \frac{s}{2}\right] = 1 - \frac{s}{4},$$

In addition, it follows from (B.89)-(B.90) that for $s \in [0, s_1]$

$$\begin{aligned} u_k(s;e_0) &\le \frac{\epsilon s}{2} \le \epsilon,\ k = 1, \ldots, K, \\ u_{K+1}(s;e_0) &\le \frac{c_1\epsilon s^2}{2} \le c_1 \le M_1, \end{aligned}$$

where the inequalities hold due to $s \leq s_1 \leq 1$ and $\epsilon \leq 1$. This gives (B.81) for $w \geq w_2$ and $s \in [0, s_1]$.

Next, we show that (B.81) for $i \neq 0$, (B.83), (B.84) and (B.85) for $j \neq 0$ holds. In particular, we show that the following holds:

$$u_i(s; e_i) \leq 1 - \frac{s}{4},\ u_i(s; e_j) \leq \epsilon,\ i, j = 1, \ldots, K,\ i \neq j,$$
$$u_0(s; e_j) \leq M_1,\ u_{K+1}(s; e_j) \leq M_1\ j = 1, \ldots, K.$$

Note from the definition of s_1 that the following holds: For $s \in [0, s_1]$,

$$\frac{2c_1 s}{\sigma^2} \leq \frac{1}{4}. \tag{B.92}$$

We then show that equation (B.85) holds for $w \geq w_2$, $s \in [0, s_1]$ and $k = 1, \ldots, K$. Now fix k. We first show that $u_k(s; e_k) \leq 1$ for $s \in [0, s_1]$ by contradiction. Suppose this is not true. Define s_3 as follows:

$$s_3 = \inf\{s \in [0, s_1] : u_j(s; e_k) > 1 \ \text{ for some } \ j \in \{1, \ldots, K\}\}.$$

By assumption, $s_3 \leq s_1$. Note that $s_3 > 0$. The reason is given as follows. It follows from (B.51), (B.59) and $u_k(0; e_k) = 1$ that

$$u_k'(0; e_k) = -x_0(0) = -\beta < 0.$$

Thus, $u_k(s; e_k) \leq 1$ for s small enough. In addition, $u_j(0; e_k) = 0$ for $j = 1, \ldots, K$ and $j \neq k$. By the continuity of $u_j(s; e_k)$, $u_j(s; e_k) \leq 1$ for s small enough. Thus, we have that $s_3 > 0$. By definition of s_3 and continuity of $u_k(s; e_k)$, it holds that $u_j(s; e_k) \leq 1$ for $s \leq s_3$ and $j = 1, \ldots, K$. In addition, $u_{j_0}(s_3; e_k) = 1$ for some $j_0 \in \{1, \ldots, K\}$. It follows from

(B.52), (B.59) and $u_{K+1}(0;e_k)=0$ that for $s\in[0,s_3]$,

$$u_{K+1}(s;e_k)=\int_0^s\sum_{j=1}^K\frac{\delta\gamma_j'(w-u)}{2(x_j(u))^{\delta+1}}u_j(u;e_k)\,\mathrm{d}u\le\int_0^s\sum_{j=1}^K\frac{\delta\gamma_j'(w-u)}{2(\underline{J}_j(u))^{\delta+1}}u_j(u;e_k)\,\mathrm{d}u\le c_1 s. \tag{B.93}$$

The first inequality follows from Lemma 53 that $x_j(u)\ge\underline{J}_j(w-u)$ and $u_j(u;e_k)\le 1$ for $u\le s_3$. The second inequality follows from (B.86). In addition, it follows from (B.50) and (B.59) that for $s\in[0,s_3]$,

$$u_0'(s;e_k)=f_x(s)u_0(s;e_k)+\frac{2}{\sigma^2}x_0(s)u_{K+1}(s;e_0).$$

By solving this ODE and substituting $u_0(0;e_k)=0$ into the solution, we have the following: For $s\in[0,s_3]$,

$$\begin{aligned}u_0(s;e_k)&=\int_0^s\left(\exp\int_u^s f_x(t)\,\mathrm{d}t\right)\frac{2}{\sigma^2}x_0(u)u_{K+1}(u;e_k)\,\mathrm{d}u\\&\le\frac{2c_1 s}{\sigma^2}\int_0^s\left(\exp\int_u^s f_x(t)\,\mathrm{d}t\right)(-2f_x(u))\,\mathrm{d}u\\&=\frac{4c_1 s}{\sigma^2}\left[1-\exp\int_0^s f_x(u)\,\mathrm{d}u\right]\le\frac{4c_1 s}{\sigma^2}\le M_1,\end{aligned} \tag{B.94}$$

where the first inequality follows from (B.86). Similarly, it follows from (B.51), (B.59) and $u_k(0;e_k)=1$ that for $s\in[0,s_3]$,

$$\begin{aligned}u_k(s;e_k)&=\exp\left(\int_0^s -x_0(u)\,\mathrm{d}u\right)+\int_0^s\left(\exp\int_u^s -x_0(t)\,\mathrm{d}t\right)(r_k-x_0(u))u_0(u;e_k)\,\mathrm{d}u\\&\le\exp\left(\int_0^s -x_0(u)\,\mathrm{d}u\right)+\int_0^s\epsilon\frac{4c_1 s}{\sigma^2}\,\mathrm{d}s\\&\le\exp(-s)+\frac{2\epsilon c_1 s^2}{\sigma^2}\\&\le 1-\frac{s}{2}+\frac{s}{4}=1-\frac{s}{4}<1.\end{aligned} \tag{B.95}$$

The first inequality follows from (B.88), (B.94) and $x_0(t) \geq 0$. The second inequality follows from (B.86) that $x_0(s) \geq 2$. The inequality in the fourth line follows from (B.87), (B.92) and $\epsilon \in (0,1)$. In addition, it follows from (B.51), (B.59) and $u_k(0;e_k) = 1$ that for $j = 1,\ldots,K$, $j \neq k$,

$$\begin{aligned} u_j(s;e_k) &= \int_0^s \left(\exp \int_u^s -x_0(t)\,\mathrm{d}t\right)(r_j - x_0(u))u_0(u;e_k)\,\mathrm{d}u \\ &\leq \int_0^s \frac{\epsilon}{2}\frac{4c_1 u}{\sigma^2}\,\mathrm{d}s \\ &= \frac{\epsilon c_1 s^2}{\sigma^2} \leq \epsilon < 1. \end{aligned} \tag{B.96}$$

The inequality in the second line follows from (B.88) and (B.94). The inequalities in the last lines follows from (B.88) that $c_1 s^2/\sigma^2 \leq c_1 s_1^2/\sigma^2 < 1$ and $\epsilon \in (0,1)$ for $s \in [0, s_3]$. Equations (B.95)-(B.96) show that $u_j(s;e_k) < 1$ for $s \in [0,s_3]$ and $j = 1,\ldots,K$. In particular, $u_j(s_3;e_k) < 1$ for $j = 1,\ldots,K$. This contradicts the definition of s_3 and the continuity of $u_k(s;e_k)$. Thus, $u_k(s;e_k) \leq 1$ for $s \in [0,s_1]$. In addition, (B.93)-(B.96) holds for $s \in [0,s_1]$. In particular, the following holds: For $s \in [0,s_1]$, $j,k = 1,\ldots,K$ and $j \neq k$,

$$u_k(s;e_k) \leq 1 - \frac{s}{4},\ u_j(s;e_k) \leq \epsilon,\ u_0(s;e_k) \leq M_1,\ u_{K+1}(s;e_k) \leq c_1 s \leq c_1 \leq M_1.$$

□

The following lemma shows that the partial derivatives of $\zeta^w(\cdot)$ are bounded for $(\beta, J) \in \mathcal{A}(w)$.

Lemma 56. *There exist constants w_0 and M such that for $w \geq w_0$ and $(\beta,J) \in \mathcal{A}(w)$, the following inequalities hold:*

$$\left|\frac{\partial \zeta^w(\beta,J)}{\partial \beta}\right| \leq M \quad \textit{and} \quad \left|\frac{\partial \zeta^w(\beta,J)}{\partial J}\right| \leq M.$$

Proof. To facilitate the proof, define the functions $y_i(w)$, $i = 0,1,\ldots,K$, as follows: For

$w \geq 0$,

$$y_0(w) = \sup\left\{-\frac{\partial \zeta^w(\beta, J)}{\partial \beta} : (\beta, J) \in \mathcal{A}(w)\right\},$$
$$y_k(w) = \sup\left\{-\frac{\partial \zeta^w(\beta, J)}{\partial J} : (\beta, J) \in \mathcal{A}(w)\right\},\ k = 1, \ldots, K.$$

It follows from Lemma 52 that $y_i(w) \geq 0$, $i = 0, 1, \ldots, K$. Thus, it suffices to show that there exist constants w_0 and M such that the following holds:

$$y_i(w) \leq M \quad \text{for} \quad i = 0, 1, \ldots, K \quad \text{and} \quad w \geq w_0.$$

Fix w and $(\beta, J) \in \mathcal{A}(w)$. We write $\phi^w(s) = \phi^w(s; X_0)$ in short where $X_0 = (\beta, J, \zeta^w(\beta, J))$. It follows from Lemma 51 that

$$\begin{aligned}
-\frac{\partial \zeta^w(\beta, J)}{\partial \beta} &= \frac{\dfrac{\partial \zeta^{w-s}(\beta(s), J(s))}{\partial \beta}\dfrac{\partial \phi_0^w(s)}{\partial x_0} + \displaystyle\sum_{i=1}^{K} \dfrac{\partial \zeta^{w-s}(\beta(s), J(s))}{\partial J_i}\dfrac{\partial \phi_i^w(s)}{\partial x_0} - \dfrac{\partial \phi_{K+1}^w(s)}{\partial x_0}}{\dfrac{\partial \zeta^{w-s}(\beta(s), J(s))}{\partial \beta}\dfrac{\partial \phi_0^w(s)}{\partial x_{K+1}} + \displaystyle\sum_{i=1}^{K} \dfrac{\partial \zeta^{w-s}(\beta(s), J(s))}{\partial J_i}\dfrac{\partial \phi_i^w(s)}{\partial x_{K+1}} - \dfrac{\partial \phi_{K+1}^w(s)}{\partial x_{K+1}}} \\
&= \frac{-\dfrac{\partial \zeta^{w-s}(\beta(s), J(s))}{\partial \beta}\dfrac{\partial \phi_0^w(s)}{\partial x_0} - \displaystyle\sum_{i=1}^{K} \dfrac{\partial \zeta^{w-s}(\beta(s), J(s))}{\partial J_i}\dfrac{\partial \phi_i^w(s)}{\partial x_0} + \dfrac{\partial \phi_{K+1}^w(s)}{\partial x_0}}{-\dfrac{\partial \zeta^{w-s}(\beta(s), J(s))}{\partial \beta}\dfrac{\partial \phi_0^w(s)}{\partial x_{K+1}} - \displaystyle\sum_{i=1}^{K} \dfrac{\partial \zeta^{w-s}(\beta(s), J(s))}{\partial J_i}\dfrac{\partial \phi_i^w(s)}{\partial x_{K+1}} + \dfrac{\partial \phi_{K+1}^w(s)}{\partial x_{K+1}}} \\
&\leq \frac{-\dfrac{\partial \zeta^{w-s}(\beta(s), J(s))}{\partial \beta}\dfrac{\partial \phi_0^w(s)}{\partial x_0} - \displaystyle\sum_{i=1}^{K} \dfrac{\partial \zeta^{w-s}(\beta(s), J(s))}{\partial J_i}\dfrac{\partial \phi_i^w(s)}{\partial x_0} + \dfrac{\partial \phi_{K+1}^w(s)}{\partial x_0}}{\dfrac{\partial \phi_{K+1}^w(s)}{\partial x_{K+1}}} \\
&\leq y_0(w-s)\frac{\partial \phi_0^w(s)}{\partial x_0} + \sum_{i=1}^{K} y_i(w-s)\frac{\partial \phi_i^w(s)}{\partial x_0} + \frac{\partial \phi_{K+1}^w(s)}{\partial x_0},
\end{aligned} \tag{B.97}$$

where $\beta(s) = \phi_0^w(s)$ and $J(s) = (\phi_1(s), \ldots, \phi_K(s))$. It follows from Lemmas 49 and 52 that

every term in both the numerator and denominator of the right-hand side of second line is positive. Thus the inequality in the third line holds. The inequality in the fourth line follows from proporty (iv) of Lemma 49 and Lemma 53 that $(\beta(s), J(s)) \in \mathcal{A}(w-s)$. Taking the supremum over all $(\beta, J) \in \mathcal{A}(w)$ yields the following: For $w \geq$ and $s \in [0, w]$,

$$y_0(w) \leq \sum_{i=0}^{K} y_i(w-s) \frac{\partial \phi_i^w(s)}{\partial x_0} + \frac{\partial \phi_{K+1}^w(s)}{\partial x_0}. \tag{B.98}$$

Similarly, it also follows from Lemma 51 that the following hold: For $w \geq 0$, $s \in [0, w]$ and $k = 1, \ldots, K$,

$$y_k(w) \leq \sum_{i=0}^{K} y_i(w-s) \frac{\partial \phi_i^w(s)}{\partial x_k} + \frac{\partial \phi_{K+1}^w(s)}{\partial x_k}. \tag{B.99}$$

Let $\epsilon \in (0, 1)$ be a constant such that the following holds:

$$(K-1)\epsilon + \sqrt{KM_1\epsilon} < \frac{s_1}{8}, \tag{B.100}$$

where s_1 and M_1 are the constants in Lemma 55. It follows from Lemma 55 that there exists w_2 such that (B.81)-(B.85) hold for $w \geq w_2$ and $s \in [0, s_1]$. Let $w_0 = w_2$. Next, we find the constant M such that $y_i(w) \leq M$ for $w \geq w_0$ and $i = 0, 1 \ldots, K$. To facilitate the proof, define the sequences $\{a_i(n), n \geq 1\}$ as follows: For $i = 0, 1, \ldots, K$ and $n \geq 1$,

$$a_i(n) = y_i(w_0 + (n-1)s_1).$$

Substituting (B.81)-(B.85) into (B.98) and (B.99) yields that for $n \geq 1$,

$$a_0(n+1) \leq \left(1 - \frac{s_1}{4}\right) a_0(n) + \sum_{i=1}^{K} \epsilon a_i(n) + M_1, \tag{B.101}$$

$$a_k(n+1) \leq M_1 a_0(n) + \left(1 - \frac{s_1}{4}\right) a_k(n) + \sum_{i \neq k} \epsilon a_i(n) + M_1, \ k = 1, \ldots, K. \tag{B.102}$$

Multiplying both sides of equation (B.101) by $\sqrt{KM_1/\epsilon}$ and adding it to equation (B.102)

yields the following: For $n \geq 1$,

$$
\begin{aligned}
&\sqrt{\frac{KM_1}{\epsilon}} a_0(n+1) + \sum_{i=1}^{K} a_i(n+1) \\
&\leq \left(\sqrt{\frac{KM_1}{\epsilon}} \left(1 - \frac{s_1}{4}\right) + KM_1 \right) a_0(n) + \sum_{i=1}^{K} \left(1 - \frac{s_1}{4} + (K-1)\epsilon + \sqrt{\frac{KM_1}{\epsilon}}\epsilon \right) a_i(n) \\
&\quad + \left(\sqrt{\frac{KM_1}{\epsilon}} + K \right) M_1 \\
&= \left(1 - \frac{s_1}{4} + \sqrt{KM_1\epsilon}\right) \sqrt{\frac{KM_1}{\epsilon}} a_0(n) + \sum_{i=1}^{K} \left(1 - \frac{s_1}{4} + (K-1)\epsilon + \sqrt{KM_1\epsilon}\right) a_i(n) \\
&\quad + \left(\sqrt{\frac{KM_1}{\epsilon}} + K \right) M_1 \\
&\leq \left(1 - \frac{s_1}{4} + (K-1)\epsilon + \sqrt{KM_1\epsilon}\right) \left(\sqrt{\frac{KM_1}{\epsilon}} a_0(n) + \sum_{i=1}^{K} a_i(n) \right) + \left(\sqrt{\frac{KM_1}{\epsilon}} + K \right) M_1 \\
&\leq \left(1 - \frac{s_1}{8}\right) \left(\sqrt{\frac{KM_1}{\epsilon}} a_0(n) + \sum_{i=1}^{K} a_i(n) \right) + \left(\sqrt{\frac{KM_1}{\epsilon}} + K \right) M_1,
\end{aligned}
$$

where the last inequality follows from (B.100). Applying this inequality recursively, we obtain the following: For $n \geq 1$:

$$
\begin{aligned}
&\sqrt{\frac{KM_1}{\epsilon}} a_0(n) + \sum_{i=1}^{K} a_i(n) \\
&\leq \left(1 - \frac{s_1}{8}\right)^n \left(\sqrt{\frac{KM_1}{\epsilon}} a_0(0) + \sum_{i=1}^{K} a_i(0) \right) + \sum_{j=1}^{n} \left(\sqrt{M_1}\epsilon + K \right) M_1 \left(1 - \frac{s_1}{8}\right)^{j-1} \\
&\leq \left(\sqrt{\frac{KM_1}{\epsilon}} a_0(0) + \sum_{i=1}^{K} a_i(0) \right) + \sum_{j=1}^{\infty} \left(\sqrt{M_1}\epsilon + K \right) M_1 \left(1 - \frac{s_1}{8}\right)^{j-1}.
\end{aligned}
$$

Denote the constant on right-hand side of the third line as M_2. Recall that $y_i(w) \geq 0$ for $w \geq 0$ and $i = 0, \ldots, K$. Thus, the sequences $a_i(n) \geq 0$ for $i = 0, \ldots, K$ and $n \geq 1$.

Therefore, it follows from the inequality we just show that the following holds:

$$a_0(n) \leq \sqrt{\frac{\epsilon}{KM_1}} M_2 \quad \text{and} \quad a_k(n) \leq M_2 \quad \text{for} \quad n \geq 1 \quad \text{and} \quad k = 1, \ldots, K. \tag{B.103}$$

Fix $w \geq w_0$ and let $n_w = \sup\{n \in \mathbb{N} : w_0 + ns_1 \leq w\}$. In addition, let $\delta_w = w - (w_0 + n_w s_1)$. It is immediate that $\delta_w \leq s_1$. Thus, substituting (B.81)-(B.85) and (B.103) into (B.98)-(B.99), we obtain the following: For $w \geq w_1$,

$$\begin{aligned}
y_0(w) &\leq \left(1 - \frac{\delta_w}{4}\right) a_0(n_w) + \sum_{i=1}^{K} \epsilon a_i(n_w) + M_1 \leq \sqrt{\frac{\epsilon}{KM_1}} M_2 + K\epsilon M_2 + M_1, \\
y_k(w) &\leq M_1 a_0(n_w) + \left(1 - \frac{\delta_w}{4}\right) a_k(n_w) + \sum_{i \neq k} \epsilon a_i(n_w) + M_1 \\
&\leq \left(\sqrt{\frac{\epsilon}{KM_1}} M_1 + 1 + (K-1)\epsilon\right) M_2 + M_1,
\end{aligned}$$

where the first inequality in each line follows from (B.81)-(B.85) and the second inequality in each line follows from (B.103). We complete the proof by letting

$$M = \max\left\{\sqrt{\frac{\epsilon}{KM_1}} M_2 + K\epsilon M_2 + M_1, \left(\sqrt{\frac{\epsilon}{KM_1}} M_1 + 1 + (K-1)\epsilon\right) M_2 + M_1\right\}.$$

□

B.2.3 Characterizing the Function $H(\cdot)$ in Terms of $\hat{\beta}_W(\cdot)$ and $\tilde{J}$ in Equilibrium Using $\zeta^w(\cdot)$

In this subsection, we show that the function $\zeta^w(\cdot)$ analyzed in previous sections characterizes the relationships of the equilibrium quantities $(\hat{\beta}_W, \tilde{J}, H)$. To be more specific, we characterize $H(w)$ in terms of $\hat{\beta}_W(w)$ and $\tilde{J}(w)$ using the function $\zeta^w(\cdot)$ for $w \geq 0$.

We first show that the flow ϕ^w characterizes the (time-reversed) evolution of the equilibrium quantities. Fix $w \geq 0$ and let $(\hat{\beta}_W, \tilde{J}, H)$ denote the equilibrium quantities, i.e.

$(\hat{\beta}_W, \tilde{J}, H)$ satisfies (4.45)-(4.50). In addition, define a function $Y : [0, w] \to \mathbb{R}^{K+2}$ as follows: For $t \in [0, w]$,

$$Y(t) = (\hat{\beta}_W(w-t), \tilde{J}(w-t), H(w-t)). \tag{B.104}$$

It is immediate that $Y'(t) = -(\hat{\beta}'_W(w-t), \tilde{J}'(w-t), H'(w-t))$ for $t \in [0, w]$. By comparing (4.45)-(4.47) to (B.50)-(B.52) and letting $X_0 = (\hat{\beta}_W(w), \tilde{J}(w), H(w))$, we conclude that $Y(t)$, $t \in [0, w]$ satisfies equation (B.53)[7]. Thus, it follows from Lemma 46 that $Y(t) = \phi^w(t; X_0)$ for $t \in [0, w]$. By substituting (B.104) into this equation, we obtain that

$$\phi^w(w; X_0) = Y(w) = (\hat{\beta}_W(0), \tilde{J}(0), H(0)).$$

Recall from (4.50) that $H(0) = 0$. In other words, we have that $\phi^w_{K+1}(w; X_0) = 0$. Thus, it follows from Lemma 50 that the following holds:

$$H(w) = \zeta^w(\hat{\beta}_W(w), \tilde{J}(w)). \tag{B.105}$$

B.2.4 Proof of Lemma 13

Recall that $(\hat{\beta}^1_W, \tilde{J}^1, H^1)$ and $(\hat{\beta}^2_W, \tilde{J}^2, H^2)$ denote the quantities in two different equilibria and their differences are defined in (4.65)-(4.67). It follows from (B.105) immediately that for $w \geq 0$ and $i = 1, 2$,

$$H^i(w) = \zeta^w(\hat{\beta}^i_W(w), \tilde{J}^i(w)).$$

Note from Lemma 50 that $\zeta^w(\cdot)$ is continuously differentiable. Thus, it follows from the mean-value theorem [Theorem 8.4, 7] that for $w \geq 0$, there exist a constant $c_w \in (0, 1)$ such

7. Since $H(t) \geq 0$ for $t \in [0, w]$, the truncation of x_{K+1} in (B.50) is immaterial.

that

$$\begin{aligned}\delta_H(w) &= H^1(w) - H^2(w) \\ &= \zeta^w(\hat{\beta}^1_W(w), \tilde{J}^1(w)) - \zeta^w(\hat{\beta}^2_W(w), \tilde{J}^2(w)) \\ &= \frac{\partial \zeta^w(\bar{\beta}(w), \bar{J}(w))}{\partial \beta}\delta_{\hat{\beta}}(w) + \sum_{k=1}^{K} \frac{\partial \zeta^w(\bar{\beta}(w), \bar{J}(w))}{\partial J_k}\delta_{\tilde{J}_k}(w),\end{aligned} \tag{B.106}$$

where $\bar{\beta}(w) = c_w\hat{\beta}^1_W(w) + (1-c_w)\hat{\beta}^2_W(w)$ and $\bar{J}(w) = c_w\tilde{J}^1(w) + (1-c_w)\tilde{J}^2(w)$. Define functions $\tilde{g}_0(\cdot)$ and $\tilde{g}_k(\cdot)$, $k = 1, \ldots, K$ as follows: For $w \geq 0$,

$$\tilde{g}_0(w) = -\frac{\partial \zeta^w(\bar{\beta}(w), \bar{J}(w))}{\partial \beta} \quad \text{and} \quad \tilde{g}_k(w) = -\frac{\partial \zeta^w(\bar{\beta}(w), \bar{J}(w))}{\partial J_k}. \tag{B.107}$$

Thus, equation (B.106) is simplified as follows: For $w \geq 0$,

$$\delta_H(w) = -\tilde{g}_0(w)\delta_{\hat{\beta}}(w) - \sum_{k=1}^{K} \tilde{g}_k(w)\delta_{\tilde{J}_k}(w). \tag{B.108}$$

It is immediate from Lemma 52 that $\tilde{g}_i(\cdot)$, $i = 0, 1, \ldots, K$ are non-negative. In addition, it follows from (4.68) and (4.69) that

$$\delta_{\tilde{\beta}}(w) = \delta_{\hat{\beta}}(w) - \frac{2\delta_H(w)}{\sigma^2}, \ w \geq 0.$$

Substituting this equation into equation (B.108) and re-arranging the terms, we obtain that

$$\delta_H(w) = -\frac{\tilde{g}_0(w)}{1 + 2\tilde{g}_0(w)/\sigma^2}\delta_{\tilde{\beta}}(w) - \sum_{k=1}^{K} \frac{\tilde{g}_k(w)}{1 + 2\tilde{g}_0(w)/\sigma^2}\delta_{\tilde{J}_k}(w), \ w \geq 0.$$

Define non-negative functions $g_i(\cdot)$, $i = 0, \ldots, K$, as follows:

$$g_i(w) = \frac{\tilde{g}_i(w)}{1 + 2\tilde{g}_0(w)/\sigma^2}, \ w \geq 0. \tag{B.109}$$

This yields equation (4.76).

Note from Lemma 11 that $(\hat{\beta}^i(w), \hat{J}^i(w)) \in \mathcal{A}(w)$ for $i = 1, 2$. Thus, it follows from the convexity of set $\mathcal{A}(w)$ that $(\bar{\beta}(w), \bar{J}(w)) \in \mathcal{A}(w)$. Therefore, we conclude from Lemma 56 that there exist w_0 and M such that $\tilde{g}_i(w) \leq M$ for $w \geq w_0$ and $i = 0, \ldots, K$. Since $\tilde{g}_0(w)$ is nonnegative, it is immediate that for $w \geq w_0$ and $i = 0, \ldots, K$, we have that $g_i(w) \leq \tilde{g}_i(w) \leq M$.

□

B.3 The Roadmap of the Uniqueness Proof

This appendix provides a detailed roadmap of the uniqueness proof (Proposition 9). The proof is done by contradiction. In what follows, we first provide an overview of the key steps that lead to the contradiction using various auxiliary lemmas (see Figure B.1). We then summarize the key steps to proving Lemma 13, which is an important technical lemma for the uniqueness proof. Two auxiliary functions, denoted by $\phi^w(\cdot)$ and $\zeta^w(\cdot)$, and several lemmas in Appendix B.2 facilitate the proof of Lemma 13. We also explain the relationship between the lemmas and how they are used to prove Lemma 13 (see Figure B.2).

The proof (of uniqueness) by contradiction proceeds as follows: Suppose that there are two different equilibria and define their difference as $(\delta_{\tilde{\beta}}, \delta_{\tilde{J}})$ (see equations (4.66) and (4.69)). The contradiction is built on the limiting properties of the difference $(\delta_{\tilde{\beta}}(w), \delta_{\tilde{J}}(w))$, $w \geq 0$. Figure B.1 shows how the the contradiction is constructed.

On the one hand, Lemma 11 provides the limits of equilibrium quantities (in any potential equilibrium). One immediate conclusion from Lemma 11 (also see equation (4.68) that defines $\tilde{\beta}$) is that the difference of the equilibrium quantities vanishes as w goes to infinity, i.e.

$$(\delta_{\tilde{\beta}}(w), \delta_{\tilde{J}}(w)) \to 0 \quad \text{as} \quad w \to \infty.$$

On the other hand, Lemmas 14 and 15 show that this convergence statement cannot

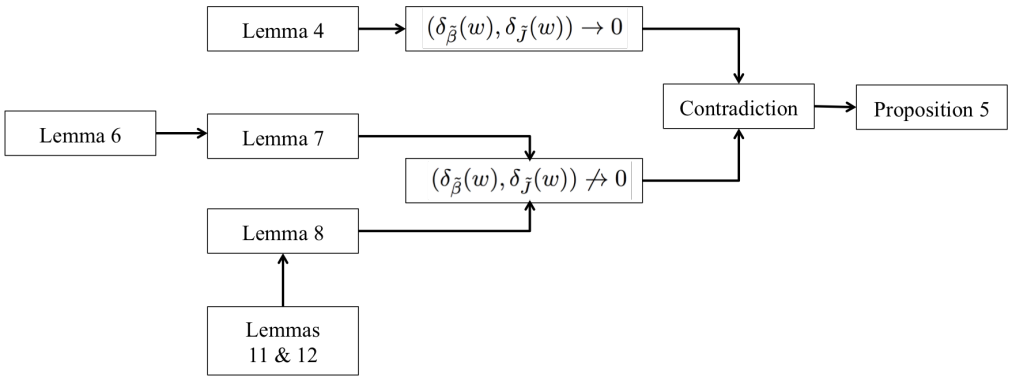


Figure B.1: The logic flow for proving uniqueness of the equilibrium

hold. Lemma 14 shows the difference of the equilibrium quantities $(\delta_{\tilde{\beta}}(w), \delta_{\tilde{J}}(w))$, $w \geq 0$ are characterized by a system of ODEs; see equation (4.82). To be more specific, the system of ODEs characterizing the difference $(\delta_{\tilde{\beta}}, \delta_{\tilde{J}})$ has three parts: A constant matrix A with a special structure, a nonnegative function $c(\cdot)$, multiplied by the identity matrix and a perturbation matrix function $B(\cdot)$. One important property of the perturbation matrix $B(w)$ is that it vanishes as w goes to infinity, i.e. $B(w) \to 0$ as $w \to \infty$. This property is proved with the help of Lemma 13.

Then Lemma 15 shows that the solution to the system of ODEs given in equation (4.82) cannot converge to zero. Lemmas 43 and 44 in Appendix B.1.2 facilitate the proof of Lemma 15. Lemma 43 provides an auxiliary technical result useful for proving Lemma 44. In turn, Lemma 44 provides a useful result on how (the sup-norm of) the solution to the time-reversed version of (4.82) decays over time. Lemma 44 (and its proof) are appropriately modified from the Poincare-Lyapunov theorem, which provides sufficient conditions for the stability of the solution to a system of ODEs. We then prove Lemma 15 by constructing the solution to the system of ODEs given in equation (4.82) using the solution to the time-reversed system of ODEs given in Lemma 44. Combining Lemmas 14 and 15, we conclude that the difference $(\delta_{\tilde{\beta}}(w), \delta_{\tilde{J}}(w))$ cannot converge to zero, which leads to the contradiction.

The rest of this section summarizes the critical steps to prove Lemma 13 in Appendix B.2. Lemma 13 characterizes $\delta_H(w)$ in terms of $\delta_{\tilde{\beta}}(w)$ and $\delta_{\tilde{J}}(w)$ for $w \geq 0$ using functions

$g_i(\cdot)$, $i = 0, 1, \ldots, K$; see equation (4.76). Figure B.2 shows how various lemmas are used (and relate to one another) to prove Lemma 13.

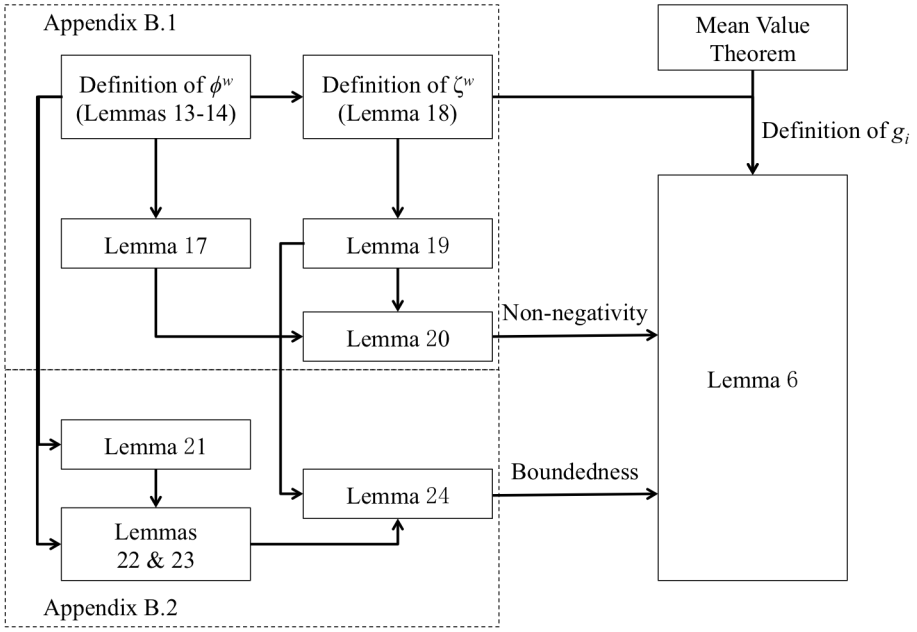


Figure B.2: The logic flow for proving Lemma 13

We first construct the functions $g_i(\cdot)$, $i = 0, 1, \ldots, K$ in two steps. In the first step, we use the auxiliary function $\zeta^w(\cdot)$ defined in Appendix B.2.1 to characterize $H(w)$ in terms of $\hat{\beta}_W(w)$ and $\tilde{J}(w)$, i.e. $H(w) = \zeta^w(\hat{\beta}_W(w), \tilde{J}(w))$ for $w \geq 0$; see Appendix B.2.3. In the second step, we apply the mean-value theorem and construct the functions $g_i(\cdot)$, $i = 0, 1, \ldots, K$ using the partial derivatives of $\zeta^w(\cdot)$; see equations (B.107) and (B.109) in Appendix B.2.4.

The auxiliary function $\zeta^w(\beta, J)$ is defined implicitly though the following equation (B.70):

$$\phi^w_{K+1}(w; (\beta, J, \zeta^w(\beta, J))) = 0. \tag{B.110}$$

The astute reader will recognize that to make sense of this definition, the ($(K+2)$-dimensional) function ϕ^w (indexed through $0, 1, \ldots, K+1$), needs to be introduced. This itself involves an intricate construction: To this end, consider the time-evolution of the equilibrium quantities given by the system of $(K+2)$ ODEs in equations (4.45)-(4.47), rewritten here for

convenience: For $t \in [0, \infty)$,

$$\begin{aligned}
\hat{\beta}'_W(t) &= \hat{\beta}_W(t)\left(\hat{\beta}_W(t) + \frac{2}{\sigma^2}(\theta - H(t))\right), \\
\tilde{J}'_k(t) &= c_k \frac{\gamma'_k(t)}{\rho_k} - \hat{\beta}_W(t)(r_k - \tilde{J}_k(t)),\ k = 1, \ldots, K, \\
H'(t) &= \sum_{k=1}^{K} \frac{\gamma'_k(t)}{2(\tilde{J}_k(t))^{\delta}}.
\end{aligned}$$

Next, fix $w \geq 0$ and consider a time-reversed version of these over $[0, w]$ as given in (B.53). To be more specific, the system of ODEs in (B.53) is an initial value problem with the initial condition $X_0 \in \mathbb{R}^{K+2}$. Let $\phi^w(t; X_0)$ for $t \in [0, w]$ denote the solution to the initial value problem given in (B.53), parametrized by w and X_0. This function $\phi^w(\cdot; X_0)$ is called the flow associated with (B.53) and it emphasizes the solution's dependence on the initial value X_0. Lemmas 45 and 46 establish that the flow is well-defined. Lemma 47 proves that it is suitably differentiable. *Recall that the flow $\phi^w(\cdot)$ is defined through* (B.53) *which is a time-reversed version of the system of ODEs of the equilibrium quantities. If we substitute the equilibrium quantities into the initial value X_0, the flow $\phi^w(\cdot; X_0)$ gives time-reversed equilibrium quantities.* Appendix B.2.3 derives this formally: If $X_0 = (\hat{\beta}_W(w), \tilde{J}(w), H(w))$, then the flow $\phi^w(\cdot)$ satisfies the following identity: For $s \in [0, w]$,

$$\phi^w(s; X_0) = (\hat{\beta}_W(w - s), \tilde{J}(w - s), H(w - s)), \tag{B.111}$$

where $(\hat{\beta}_W(w), \tilde{J}(w), H(w))$ are the equilibrium quantities.

To put (B.110) in perspective, note that it corresponds to the equilibrium condition that $H(0) = 0$ (see equation (4.50)). To see this, note that (B.110) sets the terminal value of the last component of $\phi^w(\cdot; X_0)$ to zero. Indeed, that component corresponds to H; see equation (B.111). And, due to the time-reversal the condition on the terminal value corresponds to a condition on the initial value of H, i.e. $H(0) = 0$. Lastly, Lemma 50 shows that $\zeta^w(\beta, J)$ satisfying (B.110) is unique, hence well-defined, and continuously differentiable.

Recall that Lemma 13 states two properties of the function $g_i(\cdot)$ for $i = 0, 1, \ldots, K$. First, they are non-negative. Second, the values of $g_i(w)$ are bounded for w bounded beyond a constant w_0 for all i. Also recall from the preceding discussion that the functions $g_i(\cdot)$ are defined using the partial derivatives of $\zeta^w(\cdot)$; see (B.107) and (B.109). Therefore, showing these two properties boils down to showing the related properties of the partial derivatives of the function $\zeta^w(\cdot)$.

Showing the non-negativity of $g_i(\cdot)$, $i = 0, 1, \ldots, K$ is equivalent to showing the non-positivity of the partial derivatives of $\zeta^w(\cdot)$. The recursive equations (B.72)-(B.73) given in Lemma 51 characterize the partial derivatives of the function ζ^w using the partial derivatives of ϕ^w. In turn, Lemma 49 shows that the partial derivatives of ϕ^w are nonnegative; see part (iii) of Lemma 49. Substituting the signs of the partial derivatives of ϕ^w into equations (B.72)-(B.73) in Lemma 51 gives the non-positivity of the partial derivatives of ζ^w; see Lemma 52. Then the non-negativity of $g_i(\cdot)$ (for $i = 0, 1, \ldots, K$) follows from substituting the signs of the partial derivatives of ζ^w into equations (B.107) and (B.109).

We complete the proof of Lemma 13 by showing that the values of $g_i(w)$ (for $i = 0, 1, \ldots, K$) are bounded for w larger than a constant w_0[8]. Note from equations (B.107) and (B.109) that $g_i(w)$ (for $i = 0, 1, \ldots, K$) is defined by calculating the partial derivatives of ζ^w at $(\bar{\beta}(w), \bar{J}(w))$ which is a convex combinations of two (potential) equilibria. Because all potential equilibria live in the set of $\mathcal{A}(w)$ (see Lemma 11), we have that $(\bar{\beta}(w), \bar{J}(w)) \in \mathcal{A}(w)$. Consequently, it suffices to show that the partial derivatives of $\zeta^w(\beta, J)$ are bounded by a constant M for all $(\beta, J) \in \mathcal{A}(w)$ (and $w \geq w_0$). This, in turn, is proved in Lemma 56. The proof of Lemma 56 proceeds from the recursive equations (B.72)-(B.73) for the partial derivatives of ζ^w (see Lemma 51). Lemma 53 is a technical lemma showing that the flow $\phi^w(\cdot; X_0)$ (which corresponds to the time-reversed equilibrium quantities) lives in a set characterized (in part) by the collection of sets $\mathcal{A}(s)$ for $s \in [0, w]$. It facilitates the proofs

8. We only need the boundedness for $w \geq w_0$ because the proof of uniqueness rests on the limiting properties of the equilibrium quantities.

of Lemmas 54 and 55, which in turn provide bounds on the partial derivatives of $\phi^w(\cdot; X_0)$ for specific values of X_0 defined using the values in the set $\mathcal{A}(w)$ and large w. Substituting these bounds into equations (B.72)-(B.73) gives Lemma 56.

B.4 The Parameter Estimation and Statistical Test in Section 4.2.2

B.4.1 The Maximum Likelihood Estimation of the Parameters Used in Section 4.2.2

This section estimates the reward r_k and the waiting cost c_k (for $k = 1, \ldots, 4$) from the data.

We follow Aksin et al. [3] and assume that customers in the system in every 5 seconds. To be specific, we assume that the time length of one period in the system is 5 seconds, i.e. $\delta t = 1/12$ min and the customers make their abandonment decisions at the beginning of every period. Recall that $q_k(w)$ is the probability that a class-k customer abandons in the next period if she has been waited for w periods and $\beta_k(w)$ is the probability that a class-k customer enters in the next period given that she has been waited for w periods. Thus, the probability that we observe a class k customer waiting for w_0 periods and then abandoning is given as follows:

$$q_k(w_0) \prod_{w=1}^{w_0-1} (1 - q_k(w)).$$

The probability that a class-k customer waits for w_0 periods and enters service is $\prod_{w=1}^{w_0}(1 - q_k(w))$.

To facilitate the analysis to follow, let k_i denote the class of the observed customer i and w_i denote the observed waiting time of customer i (for $i = 1, \ldots, I$). In addition, let d_i denote customer i's final abandonment result. To be specific, let $d_i = 1$ if customer i abandons eventually and $d_i = 0$ otherwise. Thus, the likelihood of the sample given the

parameters $\Theta = (r_1, \ldots, r_4, c_1, \ldots, c_4)$ is given as follows:

$$L(\Theta) = \prod_{i=1}^{I} P_i(d_i, w_i),$$

where

$$P_i(d_i, w_i) = \begin{cases} q_{k_i}(w_i) \prod_{w=1}^{w_i-1} (1 - q_{k_i}(w)), & \text{if } d_i = 1, \\ \prod_{w=1}^{w_1} (1 - q_{k_i}(w)), & \text{if } d_i = 0. \end{cases} \tag{B.112}$$

Therefore, we can write the optimization problem that maximizes the log-likelihood of the sample as follows:

$$\text{maximize}_{\Theta} \ \log L(\Theta) = \sum_{i=1}^{I} \log P_i(d_i, w_i), \tag{B.113}$$

$$\text{subject to (B.112)}, \tag{B.114}$$

$$q_k(w) = \frac{1}{2^n} \hat{q_k}(\sqrt{2^n} w), \ w \geq 1, \tag{B.115}$$

$$\hat{q_k}(w) = \frac{1}{2(\hat{J}_k(w))^{\delta}}, \ w \geq 1, \tag{B.116}$$

$$\hat{J}_k(w) = r_k - c_k \int_w^{\infty} \exp\left(\int_w^s -\hat{\beta}_k(u) \, \mathrm{d}u \right) \mathrm{d}s, \ w \geq 1, \tag{B.117}$$

$$\hat{\beta}_k(w) = \sqrt{2^n} \beta_k \left(\frac{w}{\sqrt{2^n}} \right), \ w \geq 1. \tag{B.118}$$

Equations (B.115) and (B.118) follow from (4.85)-(4.86), which provide the scaling relationship between the observed system and the heavy traffic approximation. Equations (B.116)-(B.117) follow from (4.32)-(4.33), which compute the abandonment rates. The probability of abandonment $\beta_k(\cdot)$ can be computed directly from the observed data. In sum, the optimization problem (B.113)-(B.118) gives the maximum likelihood estimates of the observed data. The estimates are provided in Table B.2. In addition, Table B.2 provides a comparison between our estimates and the estimates in Aksin et al. [3].

Since we assume a different distribution of the random shocks in the abandonment model, the estimates are different from the ones in Table 4 of Aksin et al. [3]. To be specific, we

Table B.2: The maximum likelihood estimates and the log-likelihoods of our estimates and the estimates Aksin et al. [3]

	Our estimates			Est. in Aksin et al. [3]		
	r_k	c_k	Log-likelihood	r_k	c_k	Log-likelihood
High Priority	5.302	4.072	-24,538.83	6.309	1.067	-24,413.71
Medium Priority	5.162	1.048	-123,322.84	6.175	0.506	-123,322.84
Low Priority	4.868	0.000	-108,523.45	5.299	5.45×10^{-4}	-108,495.35
No Priority	3.766	0.000	-487,626.29	4.211	0.122	-487,718.57

assume that the random shocks follow Assumption 6 with $\delta = 4$ while Aksin et al. [3] assumes that the random shocks follow Type-I extreme value distributions. In addition, we assume that the reward from service and the waiting cost of customers are homogenous within a group whereas Aksin et al. [3] assumes random rewards from service and waiting costs. However, in most cases, their variance estimates are zero, essentially coinciding with our model. The estimates on the reward from service and the waiting cost should be interpreted as the relative magnitude comparing to the random shocks. Though the estimates of the parameters are different, the resulting predictions of the two models on the abandonment probabilities are close; see the log-likelihood values in Table B.2.

B.4.2 The Kolmogorov-Smirnov test for the numerical example in Section 4.2.2

This subsection conducts a Kolmogorov-Smirnov test (K-S test) to compare the predicted (steady-state) distributions of the VOWTs from the endogenous and exogenous models. The K-S test is conducted following the steps below: For each priority group,

1. Run the simulations for both the exogenous and endogenous models. Record the VOWT of each customer. The sample sizes of the exogenous and endogenous models are m_x and m_n, respectively.

2. Compute the empirical distributions of the steady-state VOWT from the exogenous and endogenous models, their CDFs denoted by $F_x(\cdot)$ and $F_n(\cdot)$, respectively.

3. Compute the supremum of the difference of the two empirical distributions, denoted by c'. To be specific, the value of c' is given by $c' = \sup_w |F_x(w) - F_n(w)|$.

4. Scaling the supremum of the difference define $c = c'\sqrt{m_x m_n/(m_x + m_n)}$.

5. The p-value of the test is given by $1 - KS(c)$ where $KS(\cdot)$ is the cumulative distribution function of Kolmogorov-Smirnov distribution.

The steps outlined above are done using the kstest2 function in Matlab. The value of c of each priority group under the service polices considered in Table 4.4 are summarized in Table B.3 as follow:

Table B.3: The values of the test statistic c under the service polices considered in Table 4.4

	High Priority	Medium Priority	Low Priority	No Priority
FCFC	18.83	39.25	27.169	33.76
Static Priority	0.578	0.646	2.321	10.65
Threshold (75secs)	2.069	5.308	21.74	8.142
Threshold (15secs)	0.883	2.119	37.79	6.64
Reversed strict priority	115.044	11.796	0.780	1.544
Reversed point-update	47.32	78.65	48.97	2.766

The corresponding p-values of the K-S test are provided in the Table B.4.

Table B.4: The p-values of the K-S test under the service polices considered in Table 4.4. (** indicates the p-value is less then 5%.)

FCFC	4.4E-17**	8.1E-35**	2.5E-24**	4.7E-30**
Static Priority	0.304	0.269	0.0096**	5.6E-10**
Threshold (75secs)	0.016**	2.5E-5**	1.3E-19**	8.5E-8**
Threshold (15secs)	0.170	0.014**	1.5E-33**	1.7E-6**
Reversed strict priority	1.2E-100**	5.7E-11**	0.208	0.046**
Reversed point-update	7.9E-42**	4.8E-69**	4.2E-43**	0.0040**

The K-S test shows (at the significance level of 5%) that the endogenous and exogenous distributions are different in all but four cases. For example, the difference of the distributions for the high priority group under the static priority policy. However, note that the waiting times of the corresponding priority groups are small and the abandonment rates are low in these cases. They would have less impact on the performance measure.

APPENDIX C

APPENDIX OF CHAPTER 5

C.1 The Characterization of the System Dynamics

Assume that the system is empty initially. Denoting the cumulative amount of time the server is busy over $[0, t]$ by $T(t)$, the number of customers in the system at time t, denoted by $Q(t)$, is given as follows:

$$Q(t) = A(t) - S(T(t)) \geq 0, \ t \geq 0. \tag{C.1}$$

We restrict attention to work-conserving policies. That is,

$$T(t) \ \text{ increases if and only if } \ Q(t) > 0. \tag{C.2}$$

Recall that the system is stable because $b < \mu$. We restrict attention to work-conserving policies. For $t \geq 0$, let $T(t)$ denote the cumulative amount of time the server is busy over $[0, t]$. Clearly, we have that

$$T(\cdot) \ \text{ is nondecreasing with } \ T(0) = 0, \tag{C.3}$$

$$0 \leq T(t) - T(s) \leq t - s, \ 0 \leq s \leq t. \tag{C.4}$$

In addition, let $Q(t)$ denote the total number of customers in the system at time $t \geq 0$. It is nonnegative, and its evolution is governed by the following equation:

$$Q(t) = A(t) - S(T(t)), \ t \geq 0, \tag{C.5}$$

$$Q(t) \geq 0, \ t \geq 0. \tag{C.6}$$

It follows from Equation (C.4) that T is Lipschitz continuous. Thus, it is absolutely continuous and differentiable almost everywhere with respect to the Lebesgue measure on $[0, \infty)$. A time $t > 0$ is called a regular point if T is differentiable at time t. Recall that we restrict attention to work-conserving policies. Thus, the following hold at regular times: For $t \geq 0$,

$$\dot{T}(t) = 1 \quad \text{whenever} \quad Q(t) > 0, \tag{C.7}$$

where $\dot{T}(\cdot)$ denotes the derivative of $T(\cdot)$.

Given a routing policy $\mathcal{I}$, let $A_1(t)$ and $A_2(t)$ denote the cumulative numbers of customers routed to the online and offline queues up to time t, respectively. The cumulative number of customers routed to the online queue up to time t, i.e. $A_1(t)$, is given as follows:

$$A_1(t) = \sup\left\{k : i_k \in \mathcal{I}_1,\ \tau_{i_k} \leq t\right\},\ t \geq 0, \tag{C.8}$$

where we set $\sup \emptyset = 0$ for notational convenience as before. In addition, we have that $A_2(t) = A(t) - A_1(t)$ for $t \geq 0$.

To formally describe the evolution of the online and offline queue lengths, let $S_1(t)$ and $S_2(t)$ denote the total number of online and offline customers served by time t, respectively. Thus, the online and offline queue lengths, denoted by $Q_1(t)$ and $Q_2(t)$, respectively, are given as follows: For $k = 1, 2$,

$$Q_k(t) = A_k(t) - S_k(t),\ t \geq 0, \tag{C.9}$$

$$Q_k(t) \geq 0,\ t \geq 0. \tag{C.10}$$

Because we restrict attention to the work-conserving policy that gives strictly priority to the online queue, the following hold at all regular times: For $t \geq 0$,

$$S_1(t) = \int_0^t \mathbb{I}_{\{Q_1(t)>0\}}\, \mathrm{d}S(T(t)), \tag{C.11}$$

$$S_2(t) = S(T(t)) - S_1(t), \tag{C.12}$$

where $\mathbb{I}_{\{\cdot\}}$ is the indicator function. In words, Equation (C.11) says that as long as there are customers in the online queue, the server works on that queue, i.e. it gives strict priority to the online queue. Similarly, Equation (C.12) implies whenever the online queue is empty, the server works on the offline queue (provided it is not empty). Note that the evolution of the total number of customers $Q(t)$ in the system depends only on the arrival and service processes. It is independent of the routing and service policies as long as the latter is work-conserving.

C.2 Proofs of the Lemmas in Section 5.4

This section consists of the proofs of the lemmas in Section 5.4.

C.2.1 Proofs of the lemmas in Section 5.4.1

Proof of Lemma 20. We prove the statement by induction. This is true for $k = 0$ by the definition of Q_1^0, i.e. $Q_1^0 = Q_1^{\emptyset} = Q$.

As the inductive assumption, suppose that the statement is true for $k-1$ and $i \in \mathcal{I}_{k-1}^C$ (for $k = 1, \ldots, n$), i.e. $w_i^{k-1} = w_i$. We then show that it is true for k and $i \in \mathcal{I}_k^C$, i.e. $w_i^k = w_i$. Recall that the greedy rule defined in Definition 7 picks i_k^* in the k^{th} iteration. Thus, $\mathcal{I}_k^C = \mathcal{I}_{k-1}^C \backslash \{i_k^*\}$. The proof proceeds by considering the following two cases: $i > i_k^*$ and $i < i_k^*$.

Case 1: $i > i_k^*$. That is, customer i arrives after customer i_k^*. Since the customers are served in the LCFS fashion, the waiting time of customer i is independent of whether customer i_k^* is in the online or offline queue. Thus, the waiting time of customer i is unchanged after removing customer i_k^*, i.e.

$$w_i^k = w_i^{k-1} = w_i \quad \text{for} \quad i \in \mathcal{I}_k^C \cap \{i_k^* + 1, \ldots, n\},$$

where the second equality follows from the inductive assumption.

Case 2: $i < i_k^*$. That is, customer i arrives before customer i_k^*. We discuss two sub-cases

in this case. The first sub-case is when customer i enters service before customer i_k^* arrives, i.e.

$$\tau_{i_k^*} > \tau_i + w_i^{k-1} = \tau_i + w_i = s_i,$$

where the first equality follows from the inductive assumption $w_i^{k-1} = w_i$ and the second equality follows from Lemma 19. Thus, removing customer i_k^* from queue Q_1^{k-1} does not affect the waiting time of customer i because customer i has left the queue by time $\tau_{i_k^*}$. To be specific, we have that $w_i^k = w_i^{k-1} = w_i$.

The second sub-case is when customer i enters service after customer i_k^* arrives, i.e. $\tau_i < \tau_{i_k^*} < s_i$. We show next that such a customer does not exist, i.e. $i \notin \mathcal{I}_k^C$. We prove it by contradiction. Suppose there exists an index $i \in \mathcal{I}_k^C$ such that

$$\tau_i < \tau_{i_k^*} < s_i = \tau_i + w_i^{k-1}.$$

Now consider the online queue Q_1^{k-1} after $k-1$ deletions, and note that customer i will be in the queue at least during the period $[\tau_i, s_i]$. In particular, we have that $Q_1^{k-1}(t) \geq 1$ for $t \in [\tau_i, s_i)$. Define

$$t_1 = \inf\{t > \tau_i : Q_1^{k-1}(t) = 0\}$$

as the first time when Q_1^{k-1} hits zero after customer i arrives at time τ_i. Note that $t_1 > s_1 > \tau_{i_k^*}$. Thus, the following holds:

$$t_1 = \inf\{t > \tau_{i_k^*} : Q_1^{k-1}(t) = 0\},$$

because customer i_k^* arrives after customer i. Essentially, we will argue that the greedy policy would pick customer i to remove from the online queue in step k (instead of customer i_k^*) which would be a contradiction. To see this, we will show next that the reduction in the area under the online queue length process due to removing customer i from $\mathcal{I}_{k-1}^C$ is precisely $t_1 - \tau_i$. That is,

$$H(Q_1^{k-1}) - H(\Phi(Q_1^{k-1}, \{i\})) = t_1 - \tau_i.$$

Figure C.1 illustrates the change in the online queue length dynamics due to the removal of customer i. Similarly, the reduction in the area under the online queue length process due

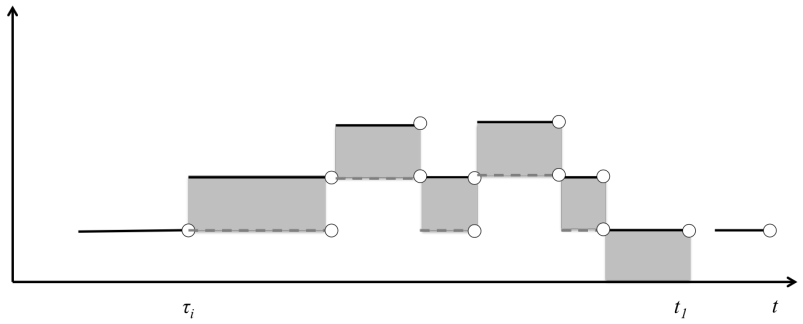


Figure C.1: The solid line shows the queue length process $Q_1^{k-1}(t)$. The dash line shows the resulting queue length process after deleting a customer arriving at time τ_i. The time t_1 is the first time when the queue hits zero. The shadowed area $(t_1 - \tau_i)$ is the savings from removing this customer.

to removing customer i_k^* from $\mathcal{I}_{k-1}^C$ is $t_1 - \tau_{i_k^*}$, i.e.

$$H(Q_1^{k-1}) - H(\Phi(Q_1^{k-1}, \{i_k^*\})) = t_1 - \tau_{i_k^*}.$$

Note that both customers i and i_k^* are in the set $\mathcal{I}_{k-1}^C$. In addition, we have that

$$H(Q_1^{k-1}) - H(\Phi(Q_1^{k-1}, \{i\})) = t_1 - \tau_i > t_1 - \tau_{i_k^*} = H(Q_1^{k-1}) - H(\Phi(Q_1^{k-1}, \{i_k^*\})),$$

where the inequality follows from the assumption that $\tau_i < \tau_{i_k^*}$. This contradicts to the definition of i_k^* given in Equation (5.7). In other words, we should have removed customer i instead of customer i_k^* in the k^{th} step. This completes the proof. □

C.2.2 Proofs of the lemmas in Section 5.4.2

Proof of Lemma 21. Note that for $\mathcal{I} \subseteq \mathcal{A}$,

$$\Phi\left(Q_1^{\emptyset}, \mathcal{I}\right) = \Phi\left(\tilde{Q}_r + \tilde{Q}_a, \mathcal{I}\right) = \tilde{Q}_r + \Phi\left(\tilde{Q}_a, \mathcal{I}\right),$$

where the first equity follows from $Q_1^{\emptyset}(t) = Q(t) = \tilde{Q}_r(t) + \tilde{Q}_a(t)$ for $t \geq 0$ and the second one follows because queue r has strict priority and no customer $i \in \mathcal{R}$ belongs to $\mathcal{I} \subseteq \mathcal{A}$. Substituting this into Equation (5.4), we obtain for $\mathcal{I} \subseteq \mathcal{A}$ that

$$p|\mathcal{I} \cap \mathcal{A}| + H(\Phi(Q_1^{\emptyset}, \mathcal{I} \cap \mathcal{A})) = p|\mathcal{I}| + H(\tilde{Q}_r + \Phi(\tilde{Q}_a, \mathcal{I})) = H(\tilde{Q}_r) + p|\mathcal{I}| + H(\Phi(\tilde{Q}_a, \mathcal{I})),$$

where the last equality follows from the additivity of the integral used to calculate it. (Recall that $H(Q)$ is the total area under the queue process Q, i.e. the integral of the queue process Q). It follows from Theorem 3 that the resulting set $\tilde{\mathcal{I}}$ of the first auxiliary p/h-lookahead policy minimizes $p|\mathcal{I}| + H(\Phi(\tilde{Q}_a, \mathcal{I}))$. Because $H(\tilde{Q}_r)$ is fixed, i.e. it does not depend on $\mathcal{I}$, the set $\tilde{\mathcal{I}} \subseteq \mathcal{A}$ prescribed by the first auxiliary p/h-lookahead policy minimizes the objective given in Equation (5.4). Hence, it is optimal for the first auxiliary system. □

Proof of Lemma 22. We prove this lemma in two steps. We first show that $Q(\tilde{s}_i) = Q(\tau_i-) - \tilde{Q}_r(\tau_i-)$. We then show that $Q(t) > Q(\tau_i-) - \tilde{Q}_r(\tau_i-),\ t \in [\tau_i, \tilde{s}_i)$.

By definition of $\tilde{s}_i$ in Equation (5.17), the time $\tilde{s}_i$ is the time when a customer in queue $\tilde{Q}_a$ (either $\tilde{Q}_n$ or $\tilde{Q}_2$) enters the service. Since the customers in queue $\tilde{Q}_r$ enjoys the static priority (with preemption), we must have that $\tilde{Q}_r(\tilde{s}_i) = 0$. Substituting this equation into Equations (5.16) and (5.17), we have the following: For $i \in \mathcal{A}$,

$$Q(\tilde{s}_i) = \tilde{Q}_a(\tilde{s}_i) = \tilde{Q}_a(\tau_i-) = Q(\tau_i-) - \tilde{Q}_r(\tau_i-).$$

The first equality follows from Equation (5.16) and that $\tilde{Q}_r(\tilde{s}_i) = 0$. The second equality follows from Equation (5.17) and the fact that $\tilde{Q}_r(\cdot)$ is right-continuos while the last equality follows from Equation (5.16).

In addition, the following holds for $t \in [\tau_i, \tilde{s}_i)$, which completes the proof:

$$Q(t) \geq \tilde{Q}_a(t) > \tilde{Q}_a(\tau_i-) = Q(\tau_i-) - \tilde{Q}_r(\tau_i-).$$

The first inequality follows from Equation (5.16) and the fact that $\tilde{Q}_r(t) \geq 0$. The second equality follows from the definition of $\tilde{s}_i$ given in Equation (5.17), whereas the equality fol-

lows from Equation (5.16). □

Proof of Lemma 23. Let j be such that $\tau_j \in (\tau_i, \tilde{s}_i)$. It follows from Equation (5.17) that $\tilde{Q}_a(\tilde{s}_i) = \tilde{Q}_a(\tau_i-)$. In addition, since $\tau_j \in (\tau_i, \tilde{s}_i)$, it also follows from Equation (5.17) that $\tilde{Q}_a(\tau_j-) > \tilde{Q}_a(\tau_i-)$. Thus, we have that $\tilde{Q}_a(\tilde{s}_i) = \tilde{Q}_a(\tau_i-) < \tilde{Q}_a(\tau_j-)$. Once again, by the definition of $\tilde{s}_j$ in Equation (5.17), we conclude that $\tilde{s}_j < \tilde{s}_i$. Therefore, the following holds:

$$\tilde{s}_j < \tilde{s}_i < \tau_i + w < \tau_j + w,$$

where the second inequality follows from the fact that $i \notin \tilde{\mathcal{I}}$ and the last inequality follows from $\tau_j > \tau_i$ by assumption. Therefore, $j \notin \tilde{\mathcal{I}}$.

We next show that $\tilde{Q}_r(\tilde{s}_i) = 0$. Note that at time $\tilde{s}_i$ the service of a customer in either queue n or queue 2 (the offline queue) is completed. Queue r must be empty at time $\tilde{s}_i$ for that to happen because it has strict preemptive priority over the queue n and the offline queue. Thus, $\tilde{Q}_r(\tilde{s}_i) = 0$. □

Proof of Lemma 24. (i) Suppose customers $i = i_0, i_1, \ldots, i_k$ arrive in $[\tau_i, \hat{s}_i)$. We proceed with a proof by induction. Note by assumption that $i = i_0 \notin \hat{\mathcal{I}}$, which constitutes the induction basis. As the induction hypothesis, we assume that (i) holds for $i_0, \ldots, j$ and we show that it also holds for customer $j+1$, i.e. $j+1 \notin \hat{\mathcal{I}}$. Note that there are no arrivals to queue r during (τ_i, τ_{j+1}) because all arriving customers join queue n by the induction hypothesis. Then the only potential changes to $\hat{Q}_r$ during (τ_i, τ_{j+1}) are due to service completions (of customers in queue r). We consider two possible cases: $\hat{Q}_r(\tau_{j+1}-) > 0$ and $\hat{Q}_r(\tau_{j+1}-) = 0$.

If $\hat{Q}_r(\tau_{j+1}-) > 0$, then all service effort during (τ_i, τ_{j+1}) is dedicated to serving queue r. Thus, all jobs who depart the system during (τ_i, τ_{j+1}) belong to queue r. Consequently, during (τ_i, τ_{j+1}) the total number of customers in the system decreases by the same amount

$\hat{Q}_r$ decreases. That is,

$$Q(\tau_i) - Q(\tau_{j+1}-) = \hat{Q}_r(\tau_i) - \hat{Q}_r(\tau_{j+1}-).$$

Rearranging the terms then gives the following:

$$Q(\tau_{j+1}-) - \hat{Q}_r(\tau_{j+1}-) = Q(\tau_i) - \hat{Q}_r(\tau_i) = Q(\tau_i-) - \hat{Q}_r(\tau_i-) + 1 > Q(\tau_i-) - \hat{Q}_r(\tau_i-).$$

The last equality follows from the fact that customer i enters the system but does not join queue r at time τ_i. Thus, it follows from the definition of $\hat{s}_i$ (see Equation (5.20)) that

$$Q(\hat{s}_i) = Q(\tau_i-) - \hat{Q}_r(\tau_i-) < Q(\tau_{j+1}-) - \hat{Q}_r(\tau_{j+1}-). \tag{C.13}$$

Combining the inequality in the preceding equation with the definitions of $\hat{s}_i$ and $\hat{s}_{j+1}$ (see Equation (5.20)), we conclude that $\hat{s}_{j+1} < \hat{s}_i$. Therefore, the following holds:

$$\hat{s}_{j+1} < \hat{s}_i < \tau_i + w < \tau_{j+1} + w, \tag{C.14}$$

where the second inequality follows from the fact that $i \notin \hat{\mathcal{I}}$ and the definition of $\hat{\mathcal{I}}$ (see Equation (5.21)). Thus, we conclude from the definition of $\hat{\mathcal{I}}$ and Equation (C.14) that customer $j+1$ is not offered the callback option, i.e. $j+1 \notin \hat{\mathcal{I}}$.

If $\hat{Q}_r(\tau_{j+1}-) = 0$, then

$$Q(\tau_{j+1}-) - \hat{Q}_r(\tau_{j+1}-) = Q(\tau_{j+1}-) > Q(\tau_i-) - \hat{Q}_r(\tau_i-) = Q(\hat{s}_i),$$

where the inequality follows from the definition of $\hat{s}_i$ in Equation (5.20) and that $\tau_{j+1} < \hat{s}_i$. By the definition of $\hat{s}_{j+1}$, we conclude that $\hat{s}_{j+1} < \hat{s}_i$. Therefore, Equation (C.14) holds as well. Thus, we conclude that customer $j+1$ is not offered the callback option, i.e. $j+1 \notin \hat{\mathcal{I}}$, in this case as well, concluding the proof of (i).

(ii) Suppose there are $k_i \geq 0$ arrivals during $[\tau_i, \hat{s}_i]$. It follows from part (i) of this lemma that all of them join queue n. It also follows from the definition of $\hat{s}_i$ (see Equation (5.20)) that

$$Q(\hat{s}_i) = Q(\tau_i-) - \hat{Q}_r(\tau_i-).$$

In other words, the total number of customers in the system decreases by $\hat{Q}_r(\tau_i-)$. Therefore, there are $k_i + \hat{Q}_r(\tau_i-)$ service completions during $[\tau_i, \hat{s}_i]$. Since queue r has the highest priority, all customers in queue r will be served first so that $\hat{Q}_r(\hat{s}_i) = 0$.

In addition, the following holds for $t \in [\tau_i, \hat{s}_i)$:

$$\begin{aligned}
\hat{Q}_1(t) &= Q(t) - \hat{Q}_2(t) \\
&= Q(t) - \hat{Q}_2(\tau_i-) \\
&= Q(t) - Q(\tau_i-) + \hat{Q}_1(\tau_i-) \\
&\geq Q(t) - Q(\tau_i-) + \hat{Q}_r(\tau_i-) \\
&> 0,
\end{aligned} \tag{C.15}$$

where the equality in the second line follows from the fact that no customers were routed to the offline queue during (τ_i, s_i), which in turn follows from part (i) of this lemma. The inequality in line four of Equation (C.15) follows because

$$\hat{Q}_1(\tau_i-) = \hat{Q}_r(\tau_i-) + \hat{Q}_n(\tau_i-) \geq \hat{Q}_r(\tau_i-).$$

The definition of $\hat{s}_i$ implies that

$$Q(t) > Q(\tau_i-) - \hat{Q}_r(\tau_i-)$$

because $t \in (\tau_i, \hat{s}_i)$ from which the last inequality in Equation (C.15) follows.

In particular, Equation (C.15) shows that the online queue is always nonempty in $(\tau_i, \hat{s}_i)$. Therefore, the $k + \hat{Q}_r(\tau_i-)$ customers who leave the queue during are $(\tau_i, \hat{s}_i)$ are from the online queue, with k of them from queue n and $\hat{Q}_r(\tau_i-)$ of them from queue r. Thus, the queue length of queue n remains the same, i.e. $\hat{Q}_n(\hat{s}_i) = \hat{Q}_n(\tau_i-)$. Since the queue n is served under the LCFS discipline, the last customer who leaves the queue is the first one among the k_i arrivals, i.e. customer i. In other words, customer i leaves the queue at time $\hat{s}_i$.

(iii) We proceed with a proof by contradiction. Suppose $\hat{Q}_n(\tau_i-) > 0$. In particular,

there exist $i' \notin \hat{\mathcal{I}}$ such that $\tau_i \in (\tau_{i'}, \hat{s}_{i'})$. Then it follows from part (i) of this lemma that $i \notin \hat{\mathcal{I}}$, which contradicts to the assumption that $i \in \hat{\mathcal{I}}$. Thus, $\hat{Q}_n(\tau_i-) = 0$. Moreover, because $i \in \hat{\mathcal{I}}$, customer i does not join queue n. Instead, he joins queue r if $i \in \mathcal{R}$; and he joins the offline queue if $i \in \mathcal{A}$. So we conclude that $\hat{Q}_n(\tau_i) = \hat{Q}_n(\tau_i-) = 0$. □

Proof of Lemma 25. As done throughout the paper, we consider a busy period and suppose that there are n customer arrivals during it. Assuming at most one event can happen at any fixed time, there are $2n$ events in the busy period. That is, the state of the system changes at $2n$ points in time. Recall that the customer arrival times are denoted by $0 < \tau_1 < \cdots < \tau_n$. It will be convenient, however, to denote the event times as $0 < t_1 < t_2 < \cdots < t_{2n}$. (Note that the set of arrival times $\{\tau_1, \ldots, \tau_n\}$ is a subset of the event times $\{t_1, \ldots, t_{2n}\}$; and the set $\{t_1, \ldots, t_{2n}\} \backslash \{\tau_1, \ldots, \tau_n\}$ corresponds to the departure or service completion times. We also let $t_0 = 0$ for notational convenience.

Note that it suffices to show that Equation (5.22) at $t = t_i$ for $i = 0, 1, \ldots, 2n$, because the system state changes only at those times. We proceed by induction. As the induction basis, we note that Equation (5.22) holds for $i = 0$, because both the original system and the auxiliary system are empty at time zero.

As the inductive hypothesis, we assume that Equation (5.22) holds for $l = 0, 1, \ldots, i$ and show that it holds for $l = i + 1$ as well. Because there is no event in between (t_i, t_{i+1}), the following holds:

$$Q_k(t_{i+1}-) = \hat{Q}_k(t_{i+1}-) \quad \text{for} \quad k = r, n, 2. \tag{C.16}$$

In which follows, we consider two cases depending on whether the event at time t_{i+1} is an arrival or a departure.

Case 1: Suppose a customer, say customer j arrives at time t_{i+1}. In particular, $\tau_j = t_{i+1}$. Note that because $Q_r(t_{i+1}-) = \hat{Q}_r(t_{i+1}-)$, it follows from Equations (5.13) and (5.17) that $s_j^r = \hat{s}_j$. Consequently, in both system, customer j is routed to the customers to the same queue. To be specific, if $s_j^r = \hat{s}_j \geq \tau_j + w$, then both system offers the callback option to

the customer j. If he accepts it and joins the offline queue, then the offline queue increases by one in both systems. That is,

$$Q_2(t_{i+1}) = \hat{Q}_2(t_{i+1}) = Q_2(t_{i+1}-) + 1 = \hat{Q}_2(t_{i+1}-) + 1.$$

Other quantities remain unchanged. If customer j rejects the offer, then both Q_r and $\hat{Q}_r$ increase by one while other quantities remain unchanged. On the other hand, if $s_j^r = \hat{s}_j < \tau_j + w$, then neither system offers the callback option; and both Q_n and $\hat{Q}_n$ increase by one, whereas all other quantities remain unchanged. Thus, we conclude that Equation (5.22) holds at time t_{i+1} in Case 1.

Case 2: In this case, we consider four further sub-cases:

(a) The online queue is empty in both systems at time $t_{i+1}-$. That is, $Q_1(t_{i+1}-) = \hat{Q}_1(t_{i+1}-)$. Equivalently, we have that

$$Q_r(t_{i+1}-) = Q_n(t_{i+1}-) = 0 \text{ and } \hat{Q}_r(t_{i+1}-) = \hat{Q}_n(t_{i+1}-) = 0.$$

Therefore, the offline queue decreases by one in both systems, i.e. Q_2 and $\hat{Q}_2$ decreases by one, whereas the other quantities remain the same.

(b) Queue n is empty whereas queue r is not empty in both systems. That is

$$Q_r(t_{i+1}-) = \hat{Q}_r(t_{i+1}-) > 0 \text{ and } Q_n(t_{i+1}-) = \hat{Q}_n(t_{i+1}-) = 0.$$

the second auxiliary system picks a customer in $\hat{Q}_r$ to enter service because it has the highest priority. We argue that the original system also picks one customer in Q_r to enter service. Note that the same set of customers are routed to Q_r and $\hat{Q}_r$ so far; see Case 1. Since $Q_r(t_{i+1}-) = \hat{Q}_r(t_{i+1}-)$, the same numbers of customers have left the queues $\hat{Q}_r$ and Q_r. Since both queues are served in the FCFS fashion, the set of customers who have entered service in Q_r and $\hat{Q}_r$ are the same. Therefore, Q_r and $\hat{Q}_r$ consist of the same customers at time $t_{i+1}-$.

(c) Queue r is empty whereas queue n is not empty in both systems. That is

$$Q_r(t_{i+1}-) = \hat{Q}_r(t_{i+1}-) = 0 \text{ and } Q_n(t_{i+1}-) = \hat{Q}_n(t_{i+1}-) > 0.$$

By an argument similar to the one in Case 2(b), we conclude that both Q_n and $\hat{Q}_n$ decrease by one.

(d) Both queue n and queue r are nonempty in both systems. That is,

$$Q_r(t_{i+1}-) = \hat{Q}_r(t_{i+1}-) > 0 \text{ and } Q_n(t_{i+1}-) = \hat{Q}_n(t_{i+1}-) > 0.$$

The second auxiliary system picks a customer in queue r to enter service because queue r has the highest priority in that system. Next, we argue that the original system also picks a customer in queue r to enter service. To this end, we first note that the same set of customers are routed to queue r in both systems so far by the induction hypothesis. Because we also have $Q_r(t_{i+1}-) = \hat{Q}_r(t_{i+1}-)$, the same number of customers have left queue r in both systems by time $t_{i+1}-$. Moreover, since both queues are served with the FCFS service discipline, the set of customers who have entered service from queue r in both systems are the same. Therefore, queue r in both systems consist of the same customers at time $t_{i+1}-$.

Let index j correspond to the customer with the smallest index present in queue n of the second auxiliary system at time $t_{i+1}-$. He arrived at time τ_j and will leave the system at time $\hat{s}_j$ in the second auxiliary system by part (iii) of Lemma 24. Since customer j has not left the system yet, we have that $t_{i+1} \in [\tau_j, s_j)$. Thus, it follows from part (i) of Lemma 24 that all customers arriving during $[\tau_j, s_j)$ are routed to queue n in the second auxiliary system. The same is true for the original system by the induction hypothesis. Therefore, all customers in queue r at time $t_{i+1}-$ (in either system) must have arrived before τ_j. Moreover, all customers in queue n at time $t_{i+1}-$ (in either system) must have arrived at or after time τ_j. Then, because the online queue (combination of queue r and queue n) is served in a FCFS basis in the original system, customers in queue r

have priority over customers in queue n. Therefore, a customer in queue r is picked to enter service at time t_{i+1} in the original system. Thus, Equation (5.22) holds for t_{i+1}.

This completes the proof. □

Proof of Lemma 26. We first show that

$$\tilde{Q}_1(t) \leq \hat{Q}_1(t),\ t \geq 0. \tag{C.17}$$

Note that all customers $i \in \mathcal{R}$ join the queue r in the first auxiliary system. Moreover, all customers who join the queue r in the second auxiliary system belong to the set $\mathcal{R}$. Thus, if a customer joins queue r in the second auxiliary system, he joins the queue r as well in the first auxiliary system. Since queue r in both systems has the highest priority among all queues in either system, we conclude that

$$\tilde{Q}_r(t) \geq \hat{Q}_r(t) \geq 0. \tag{C.18}$$

In particular, the following holds:

$$Q(t) - \tilde{Q}_r(t) \leq Q(t) - \hat{Q}_r(t),\ t \geq 0.$$

Using this, we conclude from Equations (5.18) and (5.20) that for $i \in \mathcal{A}$, $\hat{s}_i \leq \tilde{s}_i$. Consequently, if customer i in $\mathcal{A}$ is routed to the online queue in the first auxiliary system, i.e. $\tilde{s}_i < \tau_i + w$, then he is routed to the online queue in the second auxiliary system as well, because $\hat{s}_i < \tilde{s}_i < \tau_i + w$. In addition, also note that all customers in the set $\mathcal{R}$ are routed to the online queue in both systems. To summarize, all customers who are routed to the online queue in the first auxiliary system are also routed to the online queue in the second auxiliary system. Because the online queue has the strict preemptive priority over the offline queue in both systems, we conclude that $\tilde{Q}_1(t) \leq \hat{Q}_1(t)$ for $t \geq 0$.

Next, we turn to proving $\tilde{Q}_1(t) = \hat{Q}_1(t)$ for $t \geq 0$. Because $\tilde{Q}_1(t) \leq \hat{Q}_1(t)$ for $t \geq 0$ as proved immediately above, it suffices to show $\tilde{Q}_1(t) = \hat{Q}_1(t)$ for all busy periods of the online queue in the second auxiliary system. Furthermore, since the online queue has the

strict preemptive priority in both systems, it suffices to show that the same set of customers are routed to the online queue during $[t_1, t_2]$ in both systems, where $[t_1, t_2]$ is one busy period of the online queue in the second auxiliary system.

Consider the second auxiliary system and let $i_1, \ldots, i_k$ denote the customers arriving in $[t_1, t_2]$. We consider the following two cases. First, assume that all of them receive the callback option, i.e. $i_l \in \hat{\mathcal{I}}$ for $l = 1, \ldots, k$. In this case, customer i_l joins either queue r (if $i_l \in \mathcal{R}$) or the offline queue (if $i_l \in \mathcal{A}$). In particular, no customer joins queue n and we conclude that $\hat{Q}_n(t) = 0$ for $t \in [t_1, t_2]$. In this case, we also conclude the following for $t \in [t_1, t_2]$:

$$\tilde{Q}_1(t) \leq \hat{Q}_1(t) = \hat{Q}_r(t) \leq \tilde{Q}_r(t) \leq \tilde{Q}_1(t). \tag{C.19}$$

where the first inequality follows from (C.17). The equality follows from Equation (5.19) and that $\hat{Q}_n(t) = 0$ for $t \in [t_1, t_2]$. The second inequality follows from (C.18), whereas the last one follows from (5.15) and that $\tilde{Q}_n(t) \geq 0$ for $t \in [t_1, t_2]$. Therefore, in this case, it follows Equation (C.19) that $\hat{Q}_1(t) = \tilde{Q}_1(t)$ for $t \in [t_1, t_2]$.

In the second case, there exists a customer who does not receive the callback option. Let $i^* \leq i_k$ be the first such customer. Then by definition, we have that

$$\hat{Q}_n(t) = 0 \quad \text{for} \quad t \in [t_1, \tau_{i^*}) \quad \text{and} \quad \hat{Q}_n(\tau_{i^*}) = 1. \tag{C.20}$$

Thus, Equation (C.19) holds for $t \in [t_1, \tau_{i^*})$. This implies that $\hat{Q}_1(t) = \tilde{Q}_1(t)$ for $t \in [t_1, \tau_{i^*})$. Moreover, Equation (C.19) also implies that $\hat{Q}_r(t) = \tilde{Q}_r(t)$ for $t \in [t_1, \tau_{i^*})$. Therefore, the following holds:

$$\tilde{Q}_n(t) = \tilde{Q}_1(t) - \tilde{Q}_r(t) = \hat{Q}_1(t) - \hat{Q}_r(t) = \hat{Q}_n(t) = 0, \ t \in [t_1, \tau_{i^*}). \tag{C.21}$$

The first and third equalities follow from Equations (5.15) and (5.19). The last equality follows from Equation (C.20).

We complete the proof by showing that $\hat{Q}_1(t) = \tilde{Q}_1(t)$ for $t \in [\tau_{i^*}, t_2]$. To this end, we first show that $\hat{s}_{i^*} = t_2$. That is, the busy period of the online queue ends when customer

i^* departs the second auxiliary system. To see this, note from part (ii) of Lemma 24 and Equation (C.21) that

$$\hat{Q}_r(\hat{s}_{i^*}) = 0 \text{ and } \hat{Q}_n(\hat{s}_{i^*}) = \hat{Q}_n(\tau_{i^*}-) = 0.$$

Thus, it follows from (5.19) that $\hat{Q}_1(\hat{s}_{i^*}) = 0$. It also follows Part (ii) of Lemma 24 that $\hat{s}_{i^*}$ is the time when customer i^* leaves the system, which implies that $\hat{Q}_1(t) > 0$ for $t \in [\tau_{i^*}, \hat{s}_{i^*})$. Thus, we conclude that $\hat{s}_{i^*} = t_2$. Second, we note from part (i) of Lemma 24 that $i^*, \ldots, i_k \notin \hat{\mathcal{I}}$, i.e. all customers arriving after i^* (customers $i^*+1, \ldots, i_k$) join the online queue. (in particular, they join queue n) in the second auxiliary system. Therefore, to show that $\hat{Q}_1(t) = \tilde{Q}_1(t)$ for $t \in [\tau_{i^*}, t_2]$, it suffices to show that all customers in $\{i^*, \ldots, i_k\}$ joins the online queue in the first auxiliary system as well. To show this, we discuss two cases.

Case 1: Customers $i^*, \ldots, i_k$ belong to the set $\mathcal{R}$, i.e. $\{i^*, \ldots, i_k\} \subseteq \mathcal{R}$. Recall that all customers in the set $\mathcal{R}$ join queue r in the first auxiliary system. Therefore, in this case, all customers $i^*, \ldots, i_k$ in the first auxiliary system join the queue r (and hence the online queue) in the first auxiliary system.

Case 2: There exists a customer $j \in \{i^*, \ldots, i_k\}$ such that $j \in \mathcal{A}$. Let j^* be the smallest such index and note that $\hat{Q}_1(\tau_{j^*}-) = \tilde{Q}_1(\tau_{j^*}-)$, because for customers $i^*, i^*+1, \ldots, j^*-1$ belong to set $\mathcal{R}$ and they all join the online queue in both systems. Also note that $\hat{Q}_1(t) > 0$ for $t \in [\tau_{i^*}, \tau_{j^*})$ because customer i^* leaves the system at $\hat{s}_{i^*} = t_2$, i.e. when the current busy period of the online queue in the second auxiliary system ends. Consequently, no customer in the offline queue enters service during $[\tau_{i^*}, \tau_{j^*})$ in the second auxiliary system. Moreover, as discussed above, customers $i^*, i^*+1, \ldots, j^*-1$ all join the online queue in the second auxiliary system, i.e. no one joins the offline queue during $[\tau_{i^*}, \tau_{j^*})$ in the second auxiliary system. In particular, $\hat{Q}_2(\tau_{j^*}-) = \hat{Q}_2(\tau_{i^*}-)$ for $t \in [\tau_{i^*}, \tau_{j^*})$. Therefore, the following holds:

$$Q(\tau_{j^*}-) - \hat{Q}_1(\tau_{j^*}-) = \hat{Q}_2(\tau_{j^*}-) = \hat{Q}_2(\tau_{i^*}-) = Q(\tau_{j^*}-) - \hat{Q}_1(\tau_{i^*}-). \tag{C.22}$$

It remains to show that customers $j^*, \ldots, i_k$ all join the online queue $\tilde{Q}_1$ in the first auxiliary

system. Note that j^* is the first customer in $\{i^*, \ldots, i_k\}$ such that $j \in \mathcal{A}$, i.e. $\{i^*, \ldots, j^* - 1\} \subseteq \mathcal{R}$. In other words, customers $i^*, \ldots, j^* - 1$ all join $\tilde{Q}_r$ in the first auxiliary system. Therefore, it follows from Equation (C.21) that $\tilde{Q}_n(\tau_{j^*}-) = 0$. In addition, the following holds:

$$\begin{aligned} Q(\hat{s}_{i^*}) =& Q(\tau_{i^*}-) - \hat{Q}_r(\tau_{i^*}-) \\ =& Q(\tau_{i^*}-) - \hat{Q}_1(\tau_{i^*}-) \\ =& Q(\tau_{j^*}-) - \hat{Q}_1(\tau_{j^*}-) \\ =& Q(\tau_{j^*}-) - \tilde{Q}_1(\tau_{j^*}-) \\ =& Q(\tau_{j^*}-) - \tilde{Q}_r(\tau_{j^*}-). \end{aligned} \tag{C.23}$$

The first equality follows from definition of $\hat{s}_{i^*}$ in Equation (5.20). The second equality follows from Equation (C.20). The third equality follows from Equation (C.22). The fourth equality follows because $\hat{Q}_1(\tau_{j^*}-) = \tilde{Q}_1(\tau_{j^*}-)$. The last equality follows from Equation (5.15) and that $\tilde{Q}_n(\tau_{j^*}-) = 0$. In addition, it follows from Equation (5.20) that

$$Q(t) > Q(\hat{s}_{i^*}) = Q(\tau_{j^*}-) - \hat{Q}_r(\tau_{i^*}-) \text{ for } \ t \in [\tau_{j^*}, \hat{s}_{i^*}).$$

Combining this with Equation (C.23), we conclude that

$$Q(t) > Q(\hat{s}_{i^*}) = Q(\tau_{j^*}-) - \tilde{Q}_r(\tau_{j^*}-) \text{ for } \ t \in [\tau_{j^*}, \hat{s}_{i^*}).$$

It follows from this and Equation (5.18) that $\tilde{s}_{j^*} = \hat{s}_{i^*}$, which implies that

$$\tilde{s}_{j^*} \leq \hat{s}_{i^*} < \tau_{i^*} + w < \tau_{j^*} + w.$$

Therefore, we have that $j^* \notin \tilde{\mathcal{I}}$ and $j^* \in \mathcal{A}$, i.e. customer j^* does not receive the callback offer and joins the online queue (in particular, queue n) in the first auxiliary system. Then, it follows from Lemma 23 that all customers in $\mathcal{A}$ arriving during $[\tau_j, \tilde{s}_j]$ join the online queue in the first auxiliary system, where $\tilde{s}_{j^*} = \hat{s}_{i^*} = t_2$. In addition, all customers in $\mathcal{R}$ arriving during $[\tau_j, \tilde{s}_j]$ join the online queue in the first auxiliary system by definition. Thus, because

$\tilde{s}_{j^*} = t_2$, we conclude that all customers arriving during $[\tau_j, t_2]$ join the online queue in the first auxiliary system, completing the proof. □

C.3 A Bayesian Approach to Estimate the Parameters of the Arrival Process

This section describes the steps to estimate the parameters characterizing the arrival process using a Markov Chain Monte Carlo method, proposed by Zhang [119] to estimate a similar model of call center arrivals.

We use the individual call level data of a US bank call center to study the system with the callback option. To be specific, we analyze the call arrival data of brokerage customers in February 2003. To eliminate the day-of-week effect, we focus on those customers who arrive during the peak hours (9am-2pm) in the weekdays and request the service from the agents. There are $D = 19$ days of workdays in February 2003. In addition, we pick the time unit, denoted by δ, to be 10 seconds, i.e. $\delta = 10$ seconds. Since we focus on the peak hours (9am - 2pm), we have that there are $T = 3000$ time units in each day.

Although we focus on the arrivals during the peak hours in the weekdays, the arrival process still has the hour-of-day effect. To incorporate the hour-of-day effect, we modify the model of the arrival rate process in Equation (5.1) by scaling the arrival rate process by the hour-of-day effect. To be specific, we assume that the arrival process follows a Poisson process with its intensity follows

$$\lambda(t) = c(t)x(t), \quad \text{for} \quad t \in [0, T]$$

where $c(t)$ (for $t \geq 0$) is a deterministic function and $x(t)$ is a diffusion process which follows[1]

$$\mathrm{d}x(t) = \alpha(1 - x(t))\,\mathrm{d}t + \sigma\sqrt{x(t)}\,\mathrm{d}W(t),\ t \in [0, T]. \tag{C.24}$$

In each day, the realized diffusion process $x_d(t)$ (for $d = 1, \ldots, D$) draws one sample path from the the diffusion process $x(t)$. We follow Glynn et al. [50] and assume that the deterministic hour-of-day effect function $c(t)$ is a constant within each 30-minute interval, i.e. for $i = 1, 2, \ldots, 10$,

$$c(t) = \theta_i, \quad \text{for} \quad t \in [(i-1)\Delta, i\Delta),$$

where $\Delta = 300$. Thus, the parameters to estimate are $\Xi = (a, \sigma, \Theta)$ where $\Theta = (\theta_1, \ldots, \theta_{10})$.

The observed data is the counts of the incoming customers[2]. Let $Y_{k,d}$ (for $k = 1, \ldots, T$ and $d = 1, \ldots, D$) denote the observed number of arrivals within in the k-th time unit in day d. Since we assume that the arrival process follows a Poisson process given its intensity $\lambda(t)$, the following holds: For $k = 1, \ldots, T$ and $d = 1, \ldots, D$,

$$\begin{aligned}\mathbb{P}(Y_{k,d} = j | x_{k,d}, \Xi) =& \frac{\left(\int_{k-1}^{k} c(t)x_d(t)\,\mathrm{d}t\right)^j}{j} \exp\left(-\int_{(k-1)}^{k} c(t)x_d(t)\,\mathrm{d}t\right) \\ \approx& \frac{(\theta_{I(k)}x_{k,d})^j}{j} \exp(-\theta_{I(k)}x_{k,d}),\end{aligned} \tag{C.25}$$

where $I(k) = i$ if $c(k) = \theta_i$. The goal is to estimate the parameters Ξ given the observations $Y = (Y_{k,d})$ (for $k = 1, \ldots, T$ and $d = 1, \ldots, D$).

We use a Bayesian approach to estimate the parameters Ξ. To be specific, we assume that the parameters follow specific prior distributions. We then simulate the joint posterior distributions given the observations, i.e. $\mathbb{P}(\Xi|Y)$. In other words, we run a Monte Carlo simulation and sample the parameters Ξ which follow the posterior distribution $\mathbb{P}(\Xi|Y)$.

1. We normalize the constant b in Equation (5.1) to be one, so the long-term mean of the diffusion process $x(t)$ is one, i.e. $\lim_{t\to\infty} \mathbb{E}[x(t)] = 1$.

2. The data set recorded the arrival time of each customer at the accuracy of one second. Therefore, there may be multiple arrivals within one second.

The following assumption provides the prior distributions of the parameters Ξ.

Assumption 7. *The prior of* Ξ *are given by the follows:*

$$
\begin{aligned}
a &\sim \mathcal{N}(\mu_a, \sigma_a^2) \\
\sigma^2 &\sim InverseGamma(c_\sigma, d_\sigma) \\
\theta_i &\sim Gamma(c_i, d_i),\ i = 1, \ldots, 10,
\end{aligned}
$$

where c_σ *and* c_i *are the shape parameters of the inverse Gamma distribution and Gamma distribution, respectively. The constants* d_σ *and* d_i *are the rate parameters of the inverse Gamma distribution and Gamma distribution, respectively. In addition, the priors of* Ξ *are mutually independent.*

The joint posterior distribution of the parameters Ξ is difficult to compute. Therefore, it is difficult to sample the parameters Ξ from it's joint posterior distribution directly. However, the Gibbs sampler provides an iterative approach to sample the joint posterior distribution via sampling the marginal posterior distributions of the parameters. To be specific, the Gibbs sampler draws one sample of the parameters to be estimated in each iteration and update the marginal posterior distribution of each parameter given the current sample. The sequence of the samples constitutes a Markov chain, whose stationary distribution is the joint posterior distribution of the parameters to be estimated; see Gamerman [43] for more discussion. In the rest of this section, we first derive the marginal posterior distribution of the parameters Ξ. Then we describe the specific steps to implement the Gibbs sampler. In addition, we report the estimates of the parameters at the end of this section.

Prior to deriving the marginal posterior distributions of the parameters, we follow Zhang [119] and approximate the one-step transition distribution of the Markov Chain $x_{1,d}, \ldots, x_{T,d}$ (for $d = 1, \ldots, D$) by the Euler discretization scheme of the stochastic differential equation (C.24). To be specific, let $f(\cdot|x_{k,d}, \Xi)$ denote the probability density function of the value of $x_{k+1,d}$ given the values of $x_{k,d}$ and the parameters. Thus, the pdf $f(\cdot)$ can be approximated

as follows:

$$f(x_{k+1,d}|x_{k,d},\Xi) \approx \frac{1}{\sqrt{2\sigma^2 x_{k,d}}} \exp\left(-\frac{(x_{k+1,d}-a(1-x_{k,d}))^2)}{2\sigma^2 x_{k,d}}\right). \tag{C.26}$$

From now on, we replace the posterior distributions of $Y_{k,d}|x_{k,d},\Xi$ and $x_{k+1,d}|x_{k,d},\Xi$ with the approximation (C.25)-(C.26). The following lemma provides the marginal posterior distributions of the parameters Ξ.

Lemma 57. *If the parameters Ξ follow the following distributions:*

$$\begin{aligned}
&a \sim \mathcal{N}(\tilde{\mu}_a, \tilde{\sigma}_a^2)\\
&\sigma^2 \sim InverseGamma(\tilde{c}_\sigma, \tilde{d}_\sigma)\\
&\theta_i \sim Gamma(\tilde{c}_i, \tilde{d}_i),\ i = 1,\dots,10.
\end{aligned}$$

Thus, under the approximation (C.25)-(C.26)*, the following holds: For $i = 1,\dots,10$,*

$$a|\sigma^2,\Theta,X,Y \sim \mathcal{N}(B/A, 1/A) \tag{C.27}$$

$$\sigma^2|a,\Theta,X,Y \sim InverseGamma\left(\tilde{c}_\sigma + D(T-1)/2,\ \bar{d}_\sigma\right), \tag{C.28}$$

$$\theta_i|a,\sigma^2,X,Y \sim Gamma\left(\tilde{c}_i + \sum_{d=1}^{D}\sum_{k=(i-1)\Delta+1}^{i\Delta} Y_{k,d},\ \tilde{d}_i + \sum_{d=1}^{D}\sum_{k=(i-1)\Delta+1}^{i\Delta} x_{k,d}\right), \tag{C.29}$$

where A, B, and $\bar{d}_\sigma$ are given as follows:

$$\begin{aligned}
A &= \frac{1}{\tilde{\sigma}_a^2} + \frac{1}{\sigma^2}\sum_{d=1}^{D}\sum_{k=1}^{T-1}\frac{(1-x_{k,d})^2}{x_{k,d}},\\
B &= \frac{\tilde{\mu}_a}{\tilde{\sigma}_a^2} + \frac{1}{\sigma^2}\sum_{d=1}^{D}\sum_{k=1}^{T-1}\frac{(1-x_{k,d})x_{k+1,d}}{x_{k,d}},\\
\bar{d}_\sigma &= \tilde{d}_\sigma + \sum_{d=1}^{D}\sum_{k=1}^{T-1}\frac{(x_{k+1,d}-a(1-x_{k,d}))^2}{2x_{k,d}}.
\end{aligned}$$

Proof. We first prove Equation (C.27). Substituting the priority distribution of the parameter a into the posterior distribution of $a|\sigma^2, \Theta, X, Y$, we obtain its pdf as follows:

$$
\begin{aligned}
&f(a|\sigma^2, \Theta, X, Y) \\
&\propto f(\Xi, X, Y) = P(Y|\Xi, X) f(X|\Xi) f(\Xi) \\
&\propto f(X|\Xi) f(a) \propto f(a) \prod_{d=1}^{D} \prod_{k=1}^{T-1} f(x_{k+1,d}|x_{k,d}, \Xi) \\
&= \frac{1}{\sqrt{2\pi\tilde{\sigma}_a^2}} \exp\left(-\frac{(a-\tilde{\mu}_a)^2}{2\tilde{\sigma}_a^2}\right) \prod_{d=1}^{D} \prod_{k=1}^{T-1} \frac{1}{\sqrt{2\pi\sigma^2 x_{k,d}}} \exp\left(-\frac{(x_{k+1,d} - a(1-x_{k,d}))^2}{2\sigma^2 x_{k,d}}\right) \\
&\propto \exp\left[a\left(\frac{\tilde{\mu}_a}{\tilde{\sigma}_a^2} + \sum_{d=1}^{D}\sum_{k=1}^{T-1} \frac{(1-x_{k,d})x_{k+1,d}}{\sigma^2 x_{k,d}}\right) - a^2\left(\frac{1}{\tilde{\sigma}_a^2} + \frac{1}{\sigma^2}\sum_{d=1}^{D}\sum_{k=1}^{T-1} \frac{(1-x_{k,d})^2}{x_{k,d}}\right)\right].
\end{aligned}
$$

The second line follows from Equation (C.25) that conditioning on the process X, the arrival count Y does not depend on the parameter a, so it is a constant independent of a. The third line follows from Equation (C.26). Thus, it follows from the last line that $a|\sigma^2, \Theta, X, Y$ follows a Normal distribution with mean B/A and standard deviation $1/A$.

We prove Equation (C.28) by substituting the prior distribution of σ^2 into the posterior distribution of $\sigma^2|a, \Theta, X, Y$. To be specific, the following holds:

$$
\begin{aligned}
&f(\sigma^2|a, \Theta, X, Y) \\
&\propto f(\Xi, X, Y) = P(Y|\Xi, X) f(X|\Xi) f(\Xi) \\
&\propto f(X|\Xi) f(\sigma^2) \propto f(\sigma^2) \prod_{d=1}^{D} \prod_{k=1}^{T-1} f(x_{k+1,d}|x_{k,d}, \Xi) \\
&\propto \frac{\tilde{c}_\sigma^{\tilde{d}_\sigma}}{\Gamma(\tilde{c}_\sigma)} (\sigma^2)^{-\tilde{c}_\sigma - 1} \exp\left(-\frac{\tilde{d}_\sigma}{\sigma^2}\right) \prod_{d=1}^{D} \prod_{k=1}^{T-1} \frac{1}{\sqrt{2\pi\sigma^2 x_{k,d}}} \exp\left(-\frac{(x_{k+1,d} - a(1-x_{k,d}))^2}{2\sigma^2 x_{k,d}}\right) \\
&\propto (\sigma^2)^{-\tilde{c}_\sigma - 1 - D(T-1)/2} \exp\left[-\frac{1}{\sigma^2}\left(\tilde{d}_\sigma + \sum_{d=1}^{D}\sum_{k=1}^{T-1} \frac{(x_{k+1,d} - a(1-x_{k,d}))^2}{2x_{k,d}}\right)\right].
\end{aligned}
$$

The second line follows from Equation (C.25) that conditioning on the process X, the arrival count Y does not depend on the parameter σ^2, so it is a constant independent of the value of σ^2. The third line follows from Equation (C.26). Therefore, the last line implies that $\sigma^2|a, \Theta, X, Y \sim InverseGamma\left(\tilde{c}_\sigma + D(T-1)/2,\ \bar{d}_\sigma\right)$.

We end the proof by showing Equation (C.29). The following holds:

$$
\begin{aligned}
& f(\theta_i|a, \sigma^2, X, Y) \\
& \propto f(\Xi, X, Y) = P(Y|\Xi, X) f(X|\Xi) f(\Xi) \\
& \propto f(Y|\theta_i, X) f(\theta_i) \propto f(\theta_i) \prod_{d=1}^{D} \prod_{k=(i-1)\Delta+1}^{i\Delta} \mathbb{P}(Y_{k,d}|x_{k,d}, \theta_i) \\
& \propto \frac{\theta_i^{\tilde{c}_i - 1}}{\Gamma(\tilde{c}_i)} \tilde{d}_i^{-\tilde{c}_i} \exp(-\tilde{d}_i \theta_i) \prod_{d=1}^{D} \prod_{k=(i-1)\Delta+1}^{i\Delta} \frac{(\theta_i x_{k,d})^j}{j} \exp(-\theta_i x_{k,d}) \\
& \propto \theta_i^{\tilde{c}_i + \sum_{d=1}^{D} \sum_{k=(i-1)\Delta+1}^{i\Delta} Y_{k,d}} \exp\left(-\theta_i \left(\tilde{d}_i + \sum_{d=1}^{D} \sum_{k=(i-1)\Delta+1}^{i\Delta} x_{k,d}\right)\right).
\end{aligned}
$$

The second line follows from the fact that the diffusion process $x(t)$ and thus its discretization X is independent of the hour-of-day effect θ_i. This proves Equation (C.29) and thus completes the proof. □

In addition, we can also compute the posterior distribution of the Markov process $x_{k,d}$ (for $k = 1, \ldots, T$ and $d = 1, \ldots, D$) given the parameters Ξ and the observed data Y. Let $X = (x_{k,d})$ denote the underlying arrival rate process guiding the arrival process. In addition, let $X_{-(k,d)} = \{x_{k',d'} : k' \neq k \text{ and } d' \neq d\}$. It follows from Equation (11) in Zhang [119] that

$$
f(x_{k,d}|X_{-(k,d)}, Y, \Xi) \propto f(x_{k,d}|x_{k-1,d}, \Xi) f(x_{k+1,d}|x_{k,d}, \Xi) \mathbb{P}(Y_{k,d}|x_{i,d}, \Xi) \tag{C.30}
$$

The right-hand side of Equation (C.30) is calculated by substituting Equations (C.25)-(C.26). The Metropolis-Hastings algorithm can be applied to generate the samples of $x_{k,d}$ using

Equation (C.30); see Chapter 6 of Gamerman [43].

Letting J denote the the total number of samples of the Gibbs sampler, the steps of implementing the Gibbs Sampler are given as follows:

1. Initialize the initial sample of (Ξ, X) at $(\Xi^{(0)}, X^{(0)})$.

2. Given $(\Xi^{(j)}, X^{(j)})$, simulate $(\Xi^{(j+1)}, X^{(j+1)})$ as follows: For $j = 0, 1, \ldots, J-1$,

   (a) Draw a sample of $a^{(j+1)}$ from the Gaussian posterior conditional on the values of $((\sigma^{(j)})^2, \Theta^{(j)}, X^{(j)}, Y)$ given by Equation (C.27).

   (b) Draw a sample of $(\sigma^{(j+1)})^2$ from the inverse Gamma posterior conditional on $(a^{(j)}, \Theta^{(j)}, X^{(j)}, Y)$ given by Equation (C.28).

   (c) Draw a sample of θ_i^{j+1} from the Gamma distribution conditional on the values of $(a^{(j)}, (\sigma^{(j)})^2, \Theta_{-i}^{(j)}, X^{(j)}, Y)$ given by Equation (C.29).

   (d) Draw a sample of $x_{k,d}^{(j+1)}$, from the posterior distribution conditional on the value of $(\Xi^{(j)}, X_{-(i,d)}^{(j)}, Y)$ by Equation (C.30) with the Metropolis-Hastings algorithm.

There is no specific rule to determine the number of samplers J. We draw $J = 2,000,000$ samples and trace the plots of the samples of the unknown parameters Ξ to ensure that the samples have converged. We discard the first half of the samples and use the mean of the second half of the samples as the estimates of the unknown parameters Ξ. In addition, we use various initial values of the parameters Ξ and do not observe the dependency of the estimates on the initial value. The estimates of the unknown parameters Ξ are given in Table C.1. The parameters a and σ characterizes the dynamics of the underlying diffusion process $x(t)$ that rules the arrival rate process $\lambda(t)$[3]. The parameters $\Theta = (\theta_1, \ldots, \theta_{10})$ capture the hour-of-day effect.

In the discrete event simulation in Section 5.6, we use the parameters a and σ estimated from the dataset but ignore the hour-of-day effect. Instead, we scale the diffusion process

3. Note that $\mathbb{E}[x(t)|x(0)] = x(0)e^{-at} + (1 - e^{-at})$. Note that $1/a \approx 115$ time units, which is equivalent to 19 minutes. Thus, the pikes induced by the variation of $x(t)$ lasts at the order of 20-30 minutes.

Parameters	Mean	S.D.
a	8.73e-3	4.27e-4
σ	1.82e-2	3.29e-5
θ_1	0.488	0.97e-2
θ_2	0.712	1.46e-2
θ_3	0.754	1.53e-2
θ_4	0.768	1.60e-2
θ_5	0.752	1.51e-2
θ_6	0.663	1.37e-2
θ_7	0.591	1.22e-2
θ_8	0.581	1.17e-2
θ_9	0.612	1.23e-2
θ_{10}	0.623	1.27e-2

Table C.1: The estimates of the parameters a, σ and Θ.

$x(t)$ by the average arrival rate over the peak hours to obtain the arrival rate. We make this assumption because we assume that the number of agents is a constant over the peak hours, which ignores the hour-of-day effect in the staffing level. To be specific, we simulate the system and vary the number of agents using the arrival processes with the estimated parameters a and σ^2, the empirical service time and abandonment time distributions. We compare the simulated average waiting time and fraction of abandoning customers with the data and pick the number of agents.

REFERENCES

[1] Afèche, P. and V. Sarhangian. 2015. Rational abandonment from priority queues: Equilibrium strategy and pricing implications. *Working paper.*

[2] Aksin, Z., M. Armony and V. Mehrotra. 2007. The modern call-center: A multi-disciplinary perspective on operations management research. *Production Oper. Management* **16**, 665-688.

[3] Aksin, Z., B. Ata, S. Emadi and C. Su. 2013. Structural estimation of callers's delay sensitivity in call centers. *Management Sci.* **59**(12), 2727 - 2746.

[4] Aksin, Z., B. Ata, S. Emadi and C. Su. 2017. Impact of delay announcements in call centers: An empirical approach. *Oper. Res.* **65**(1), 242-265.

[5] Aliprantis, C.D. and K.C. Border. 2007. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, New York.

[6] Anderson, S.P., A. de Palma, J. Thisse. 1992. *Discrete Choice Theory of Product Differentiation.* The MIT Press, Cambridge, MA.

[7] Apostol, T.M. 1969. *Calculus: Volume 2*. Wiley, Hoboken, NJ.

[8] Armony M. and C. Maglaras. 2004a. On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Oper. Res.* **52**(2), 271-292.

[9] Armony M. and C. Maglaras. 2004b. Contact centers with a call-back option and real-time delay information. *Oper. Res.* **52**(4), 527-545.

[10] Armony, M., N. Shimkin and W. Whitt. 2009. The impact of delay announcements in many-server queues with abandonment. *Oper. Res.* **50**(1) 66-81.

[11] Assaf D. and M. Haviv. 1990. Reneging from processor sharing systems and random queues. *Math. Oper. Res.* **15**(1), 129-138.

[12] Ata, B. 2006. Dynamic control of a multiclass queue with thin arrival streams. *Oper. Res.* **54**(5), 876-892.

[13] Ata, B., Y. Ding and S. Zenios. 2016a. Donor-dependent scoring schemes for kidney allocation. *Working paper.*

[14] Ata, B. and J. Friedewald. 2017. Organ transplantation. *Working paper.*

[15] Ata, B., P. Glynn and X. Peng. 2016b. An equilibrium analysis of a discrete-time Markovian queue with endogenous abandonments. *Queueing Sys.*. Forthcoming.

[16] Ata, B., O. Islegen and S. Duran. 2016. An analysis of time-based pricing in electricity supply chains. *Working Paper.*

[17] Ata, B., T. Olsen. Near-optimal dynamic leadtime quotation and scheduling under convex-concave customer delay costs. *Oper. Res.* **57**(3) 753-768.

[18] Ata, B. and X. Peng. 2017. *An equilibrium analysis of a multiclass queue with endogenous abandonments in the conventional heavy traffic regime. Oper. Res.* Forthcoming.

[19] Ata, B. and S. Shneorson. 2006. Dynamic control of an M/M/1 service System with adjustable arrival and service rates. *Management Sci.* **52**(11), 1778-1791.

[20] Ata, B., A. Skaro and S. Tayur. 2017. OrganJet: Overcoming geographical disparities in access to deceased donor kidneys in the United States. *Management Sci.* Forthcoming.

[21] Ata, B. and M.H. Tongarlak. 2013. On scheduling a multiclass queue with abandonments under general delay costs. *Queueing Sys.* **74**(1), 65-104.

[22] Baccelli, F. and G. Hebuterne. 1981. On queues with impatient customers. F. Kylstra, ed. *Performance '81.* North Holland, Amsterdam, The Netherlands, 159-179.

[23] Baccelli, F., P. Boyer, and G. Hebuterne. 1984. Single-server queues with impatient customers. *Advance in Applied Probability* **16** 887-905.

[24] Bassamboo, A, J.M. Harrison and A. Zeevi. 2006. Design and control of a large call center: Asymptotic analysis of an LP-based method. *Oper. Res.* **54**(3), 419-435.

[25] Bertsimas, D, X. Doan. 2010. Robust and data-driven approaches to call centers. *Eur. J. Oper. Res.* **207**(2), 1072-1085.

[26] Besbes, O., C. Maglaras. 2009. Revenue optimization for a make-to-order queue in an uncertain market environment. *Oper. Res.* **57**(6),1438-1450.

[27] Bhulai S. and G. Koole. 2003. A queueing model for call blending in call centers. *IEEE Trans. Automatic Control* **48**(8),1434-1438.

[28] Billingsley, P. 1999. *Convergence of Probability Measures*, 2nd ed. John Wiley and Sons, New York.

[29] Boxma, O., D. Perry, and W. Stadje. 2011. The M/G/1+G queue revisited. *Queueing Sys.* **67**, 207-220.

[30] Bramson, M. 1998. State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Sys.* **30**, 89-140.

[31] Brandt A. and M. Brandt. 1999. On the $M(n)/M(n)/s$ queue with impatient calls. *Performance Evaluation* **35**, 1-18.

[32] Bremaud, P. 1981. *Point processes and queues: Martingale dynamics.* Springer, New York, NY.

[33] Brown, D.B., J. Smith and P. Sun. 2010. Information relaxations and duality in stochastic dynamic programs. *Oper. Res.* **58**(4) 785-801.

[34] ContactBabel. 2016. *US contact center decision makers' guide, 9th edition.* http://www.contactbabel.com/reports.cfm

[35] Coppel, W. A. 1965. *Stability and Asymptotic Behavior of Differential Equations*. Heath Mathematical Monographs, Boston.

[36] Czornik, A. 2005. On the generalized spectral subradius. *Linear Algebra and its Applications* **407**, 242-248.

[37] Cudina, M., K. Ramanan. 2011. Asymptotically optimal controls for time-inhomogeneous networks. *SIAM Journal on Control and Optimization* **49**(2) 611-645.

[38] Dai, J.G., S. He and T. Tezcan. 2010. Many-server diffusion limits for $G/Ph/n + GI$ queues. *The Annals of Applied Probability* **20**(5) 1854-1890.

[39] Del Moral, P. and L. Miclo. 2006. Self-Interacting Markov Chains. *Stochastic Analysis and Applications* **24**, 615-660.

[40] Deslauriers A., P. L'Ecuyer, J. Pichitlamken, A. Ingolfsson, A.N. Avramidis. 2007. Markov chain models of a telephone call center with call blending. *Comput. Oper. Res.* **34**(6),1616?1645.

[41] Z. Feldman, A. Mandelbaum, W. A. Massey, and W. Whitt. 2008. Staffing of time-varying queues to achieve time-stable performance. *Management Sci.* **54**(2), 324-338.

[42] Finch, P.D. 1960 Deterministic customer impatience in the queueing system GI/M/1. *Biometrika* **47**, 45-52.

[43] Gamerman, D. 2006. *Markov chain Monte Carlo : stochastic simulation for Bayesian inference.* Taylor & Francis.

[44] Gans, N., G. Koole, and A. Mandelbaum. 2003. Telephone call centers: Tutorial, review and research prospects. *Manufacturing Service Oper. Management* **5**, 73-141.

[45] Gans N. and Y.P. Zhou. 2003. A call-routing problem with service-level constraints. *Oper. Res.* **51**(2), 255?271.

[46] Gans, N., H. Shen, Y. Zhou, N. Korolev, A. McCord and H. Ristock 2015. Parametric forecasting and stochastic programming models for call-center workforce scheduling. *Manufacturing Service Oper. Management* **17**(4), 571-588.

[47] Gao, L. and K. Kuzu. 2016. To wait or not to wait: The theory and practice of ticket queues. *Working paper.*

[48] Gavish, B. and P. J. Schweitzer. 1977. The Markovian queue with bounded waiting time. *Management Sci.* **23**,1349-1357.

[49] Glynn, P.W. 2014. Perspectives on Traffic Modeling. *INFORMS Markov Lecture.* https://www.informs.org/Community/APS/Markov-Lecture-Slides

[50] Glynn, P., L.J. Hong and X. Zhang. 2014. Modeling call center arrivals: A tale of three timescales. *Working paper.*

[51] Grass D. and C.M. Harris. 1974. *Fundamentals of queueing theory.* Wiley.

[52] Green, L. V. and P. J. Kolesar. 1991. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Sci.* **37**, 84-97.

[53] Green, L. V., P. J. Kolesar and W. Whitt. 2007. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Oper. Management* **16**(1), 13-39.

[54] Gurvits, L. 1995. Stability of discrete linear inclusion. *Linear Algebra and its Applications* **231**, 47-85.

[55] Hampshire, R.C., W.A. Massey. 2010. Dynamic optimization with applications to dynamic rate queues. *TUTORIALS in Operations Research, INFORMS Society* 210247.

[56] Harrison, J.M., A. Zeevi. 2005. A method for staffing large call centers based on stochastic fluid models. *Manufacturing Service Oper. Management* **7**(1), 20-36.

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

[58] Hassin, R. 2016. *Rational queueing.* Taylor and Francis Group, Boca Raton, FL.

[59] Hassin, R. and M. Haviv. 1995. Equilibrium strategies for queues with impatient customers. *Oper. Res. Lett.* **17**(1), 41-45.

[60] Hassin, R. and M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems.* Kluwer Academic Publishers.

[61] Haviv, M. and Y. Ritov. 2001. Homogeneous customers renege from invisible queues at random times under deteriorating waiting conditions. *Queueing Sys.* **38**, 495-508.

[62] He, B., Y. Liu, and W. Whitt. 2016. Stabilizing performance in nonstationary queues with non-Poisson arrivals.*Probability in the Engineering and Informational Sciences* **30**, 593-621.

[63] Hirsch, M.W., S. Smale and R. Devaney. 2003. *Differential Equations, Dynamical Systems, and an Introduction to Chaos*, 2nd edition. Academic Press, San Diego.

[64] Ibrahim, R., H. Ye, P. L'Ecuyer and H. Shen. 2015. Modeling and forecasting call center arrivals: A literature survey and a case study. *International Journal of Forecasting* **32**, 865-874.

[65] Jeanblanc, M., M. Yor and M. Chensney. 2009. *Mathematical methods for financial markets.* Springer, New York, NY.

[66] Jennings, O., A. Mandelbaum, W. Massey and W. Whitt. 1996. Server staffing to meet time-varying demand. *Management Sci.* **42**, 1383-1394.

[67] Jennings, O. and J. Pender. 2016. Comparisons of standard and ticket queues. *Queueing Sys.* **84**, 145-202.

[68] Kelly, F.P. and C.N. Laws. 1993. Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling. *Queueing Sys.* **13**, 47-86.

[69] Kim, J. and A.R. Ward. 2013. Dynamic scheduling of a $GI/GI/1 + GI$ queue with multiple customer classes. *Queueing Sys.* **75**, 339-384.

[70] Kim, S., P. Vel, W. Whitt and W.C. Cha. 2015. Poisson and non-Poisson properties in appointment-generated arrival processes: the case of an endocrinology clinic. *Oper. Res. Letters*, **43**, 247-253

[71] Kim, S., W. Whitt and W.C. Cha. 2017. A data-driven model of an appointment-generated arrival processes at an outpatient clinic. *Working Paper.*

[72] Kim, S. and W. Whitt. 2014. Choosing arrival process models for service systems: Tests of a nonhomogeneous Poisson process. *Naval Research Logistics* **61**(1), 66-90.

[73] Kim, S. and W. Whitt. 2014. Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes?. *Manufacturing Service Oper. Management* **16**(3), 464-480.

[74] Kocaga, Y.L., M. Armony and A.R. Ward. 2015. Staffing call centers with uncertain arrival rates and co-sourcing. *Production and Operations Management* **24**(7), 1101-1117.

[75] Kocaga, Y.L. and A. Ward. 2010. Admission control for a multi-server queue with abandonment. *Queueing Systems* **65**(3), 275-323.

[76] Kong, Q. 2014. *A Short Course in Ordinary Differential Equations.* Springer, New York.

[77] Kuzu, K., Susan H. Xu and L. Gao. 2017. To wait or not to wait: The theory and practice of ticket queues. *Working Paper.*

[78] Legros, B., O. Jouini and G. Koole. 2016. Optimal scheduling in call centers with a callback option. *Performance Evaluation* **95**, 1-40.

[79] Lemmens, B. and N. Nussbaum. 2012. *Nonlinear Perron-Frobenius Theory*. Cambridge University Press, Cambridge, UK.

[80] Lewis, M.E., H. Ayhan, and R.D. Foley. 1999. Bias optimality in a queue with admission control. *Probability in the Engineering and Informational Sciences* **13**(3), 309-327.

[81] Lewis, M.E., H. Ayhan, and R.D. Foley. 2002. Bias optimal admission policies for a nonstationary multiclass queueing system. *Journal of Applied Probability* **39**(1), 20-37.

[82] Liao S., G. Koole, C. van Delft and O. Jouini. 2012. Staffing a call center with uncertain non-stationary arrival rate and flexibility. *OR Spectum* **34**(3), 691-721.

[83] Liu, Y. and W. Whitt. 2012. Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper. Res.* **60**(6),1551-1564.

[88] Maglaras C., J. Yao and A. Zeevi. 2017. Observational learning in queues with abandonments. *Working paper.*

[85] Mandelbaum, A. and P. Momčilović. 2012. A model for rational abandonments from invisible queues. *Math. Oper. Res.*. **37**(1), 41-65.

[86] Mandelbaum, A. and N. Shimkin. 2000. A model for rational abandonments from invisible queues. *Queueing Sys.* **36**, 141-173.

[87] Maman, S. 2009. Uncertainty in the demand for service: The case of call centers and emergency departments. Master thesis, Technion-Israel Institute of Technology, Haifa.

[88] Maglaras, C., J. Yao and A. Zeevi. 2015. Observational learning in queues with abandonments. *Working paper*. Columbia University, New York, NY.

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

[90] Örmeci, E.L, A. Burnetas and J. van der Wal. 2001. Admission policies for a two class loss system. *Stochastic Models* **17**(4) 513-539.

[91] Pang, G. and O. Perry. 2015. A logarithmic safety staffing rule for contact centers with call blending. *Management Sci.* **61**(1), 73-91.

[92] Protasov, V., R.M. Jungers,V.D. Blondel. 2010. Joint spectral characteristics of matrices: a conic programming approach. *SIAM J. Matrix Anal. Appl.* **31**(4), 2146-2162.

[93] Reed, J. and A.R. Ward. 2008. Approximating the GI/GI/1+GI queue with a nonlinear drift diffusion: Hazard rate scaling in heavy traffic, *Math. Oper. Res.* **33**(3), 606-644.

[94] Reiman, M. I. 1982. The heavy traffic diffusion approximation for Sojourn times in Jackson networks. R. L. Disney, T. J. Ott, eds. *Applied Probability of Computer Science, The Interface, II.* Birhauser, Boston, 409 - 422.

[95] Rubino, M. and B. Ata. 2009. Dynamic control of a make-to-order, parallel-server system with cancellations. *Oper. Res.* **57**(1), 94-108.

[96] Rudin, W. 1976. *Principles of Mathematical Analysis.* McGraw-Hill.

[97] Rust, J. 1987. Optimal replacement of GMC bus engines: An empirical model of haroldzurcher. *Econometrica.* **55**(5), 999-1033.

[98] Sanchez, D.A. 1992. *Ordinary Differential Equations and Stability Theory: An Introduction.* Dover Publications.

[99] Shi, P., M. Chou, J. G. Dai, D. Ding, and J. Sim. 2016. Models and insights for hospital inpatient operations: Time-dependent ED boarding time. *Management Sci.* **62**(1), 1-28.

[100] Shimkin, N. and A. Mandelbaum. 2004. Rational abandonment from tele-queues: Nonlinear waiting costs with heterogeneous preferences. *Queueing Sys.* **47**, 117-146.

[101] Spencer J., M. Sudan and K. Xu. 2014. Queueing with future information. *The Annals of Applied Probability* **24**(5), 2091-2142.

[102] Stanford, R.E. 1979. Reneging phenomena in single channel queues. *Math. Oper. Res.* **4**, 162-178.

[103] Stidham, S. 1985. Optimal control of admission to a queueing system. *IEEE Trans. Automat. Contr.* **30**(8), 705-713.

[104] Stidham, S. 2002. Analysis, design, and control of queueing systems. *Oper. Res.* **50**, 197-216.

[105] Stokey, N.L. and R.E. Lucas. 1989. *Recursive Methods in Economic Dynamics.* Harvard University Press, Cambridge, MA.

[106] Verhulst, F. 2006. *Nonlinear Differential Equations and Dynamical Systems.* Springer, New York.

[107] Ward, A.R. 2011. Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO, single class models, *Surveys in Math. Oper. Res. and Management Sci.* **16**(1), 1-14.

[108] Ward, A.R. and P.W. Glynn. 2005. A diffusion approximation for a $GI/GI/1$ queue with balking or reneging. *Queueing Sys.* **43**, 371-400.

[109] Ward, A. and S. Kumar. 2008. Asymptotically optimal admission control of a queue with impatient customers, *Mathematics of Oper. Res.* **33**(1), 167-202.

[110] Chia-Li Wang. 2016. On Socially Optimal Queue Length. *Management Sci.* **62**(3), 899-903.

[111] Whitt. W. 1999. Using different response-time requirements to smooth time-varying demand for service. *Oper. Res. Letters* **24**(1-2), 1-10.

[112] Whitt, W. 2002. *Stochastic Process Limits*. Springer, New York.

[113] Whitt, W. 2016. Queues with Time-Varying Arrival Rates: A Bibliography. *Working paper.*

[114] Whitt, W. and J. Zhao. 2017. Many-server loss models with non-Poisson time-varying arrivals. *Naval Research Logistics.* Forthcoming.

[115] Xu, K. and C. Chan. 2016. Using future information to reduce waiting times in the emergency department via diversion. *Manufacturing Service Oper. Management* **18**(3), 314-331.

[116] Yoon, S. and M. Lewis. 2004. Optimal Pricing and Admission Control in a Queueing System with Periodically Varying Parameters. *Queueing Systems* **47**(3), 177-199.

[117] Zayas-Cabán G. and M.E. Lewis. 2016. Admission control in a two Class loss system with periodically varying parameters and abandonments. *Working paper.*

[118] Zeidler, E. 1998. *Nonlinear Functional analysis and its applications: I: Fixed-point theorems.* Springer, New York.

[119] Zhang, X. 2013. A Bayesian approach for modeling and analysis of call center arrivals. *Proceedings of the 2013 Winter Simulation Conference.*

[120] Zhang, X., J. Hong and J. Zhang. 2014. Scaling and modeling of call center arrivals. *Proceedings of the 2014 Winter Simulation Conference.*

[121] Zohar, E., A. Mandelbaum and N. Shimkin. 2002. Adaptive behavior of impatient customers in tele queues: Theory and empirical support. *Management Sci.* **48**(4), 566-583.