

THE UNIVERSITY OF CHICAGO

MANY-SERVER QUEUEING MODELS WITH APPLICATIONS TO MODERN
SERVICE OPERATIONS MANAGEMENT

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

YUEYANG ZHONG

CHICAGO, ILLINOIS

DECEMBER 2023

Copyright © 2023 by Yueyang Zhong
All Rights Reserved

To my family, with a special tribute to my beloved grandfather.

TABLE OF CONTENTS

LIST OF FIGURES	viii
ACKNOWLEDGMENTS	x
ABSTRACT	xii
0 INTRODUCTION	1
1 BEHAVIOR-AWARE QUEUEING: THE FINITE-BUFFER SETTING WITH MANY STRATEGIC SERVERS	9
1.1 Introduction	9
1.1.1 Contributions of This Chapter	10
1.1.2 Literature Review	13
1.1.3 The $M/M/N/k$ Queue with Non-strategic Servers	15
1.2 The Strategic Server Queueing Model	17
1.3 The $M/M/N/k$ System	21
1.3.1 The First-Order Condition and Candidate Equilibria	21
1.3.2 Numerical Examples	23
1.4 Exact Analysis for Specialized Systems	25
1.4.1 The $M/M/N/N$ Loss System	25
1.4.2 The $M/M/1/k$ Single-Server System	28
1.5 Asymptotic Analysis	32
1.5.1 Limiting FOC	32
1.5.2 Properties of Limiting Equilibria	35
1.5.3 Prelimit Convergence and Impact of System Size	41
1.6 Looking Ahead: Implications for Behavior-Aware Queueing Models	45
1.6.1 Connecting to Empirical Work	46
1.6.2 Optimal Design and Control	48
1.6.3 Strategic Arrivals and Strategic Servers	49
1.6.4 Utility Function Generalization	51
2 LEARNING TO SCHEDULE IN MULTICLASS MANY-SERVER QUEUES WITH ABANDONMENT	54
2.1 Introduction	54
2.1.1 Contributions of This Chapter	56
2.1.2 Literature Review	58
2.1.3 Organization	60
2.1.4 Notation	60
2.2 Problem Setting	61
2.2.1 The Multiclass $GI/GI/N+GI$ Queue	61
2.2.2 Regret Performance Metric	65
2.2.3 Benchmark Policy Justification	68

2.2.4	Technical Assumptions	68
2.3	Regret Lower Bound	70
2.4	The Proposed Learn-then-Schedule Policy	73
2.5	Proof of Regret Upper Bound	77
2.5.1	Regret Decomposition and Some Bounds	77
2.5.2	Bound on Estimation Error	80
2.5.3	Balancing the Regret Terms	81
2.6	Numerical Experiments	83
2.6.1	The LTS Policy Performance	83
2.6.2	Performance Robustness	86
2.7	Asymptotic Optimality of the Benchmark $a\mu$ -Rule	87
2.7.1	The Static Scheduling Problem	87
2.7.2	The Asymptotic Optimality Result	89
2.8	Concluding Remarks	91
3	ASYMPTOTICALLY OPTIMAL IDLING IN THE $GI/GI/N+GI$ QUEUE	94
3.1	Introduction	94
3.2	The Model and Admissible Policy Class	96
3.2.1	The Model	97
3.2.2	The Admissible Policy Class	98
3.3	The Control Problem	101
3.4	The Fluid Control Problem	102
3.5	The Proposed Policy π_*^N	104
3.6	Asymptotic Optimality of π_*^N	106
3.7	Preliminary Results	108
3.7.1	Stationary Distributions of the N -Server Queue	108
3.7.2	A Fluid Limit Theorem	109
3.7.3	Properties of Stationary Fluid Model Solutions	110
3.8	Proofs of Main Results (Theorems 8 and 9)	111
3.9	Extension: Holding Costs	116
3.10	Conclusion	122
4	CONCLUSION	123
APPENDIX A	APPENDIX FOR CHAPTER 1	127
A.1	Preliminaries	127
A.1.1	Proof of Lemma 18	128
A.1.2	Proof of Lemma 19	128
A.1.3	Proof of Lemma 20	130
A.2	Proofs from Section 1.1.3	131
A.2.1	Preliminaries	131
A.2.2	Proof of Lemma 1	134
A.2.3	Proof of Lemma 2	135
A.3	Proofs from Section 1.2	139

A.3.1	Proof of Proposition 1	139
A.3.2	Proof of Lemma 3	139
A.4	Proofs from Section 1.3	141
A.4.1	Preliminaries	141
A.4.2	Proof of Theorem 1	164
A.5	Proofs from Section 1.4.1	169
A.5.1	Preliminaries	169
A.5.2	Proof of Theorem 2	175
A.5.3	Proof of Proposition 2	177
A.5.4	Proof of Proposition 3	178
A.6	Proofs From Section 1.4.2	182
A.6.1	Proof of Lemma 4	182
A.6.2	Proof of Proposition 4	183
A.6.3	Proof of Theorem 3	183
A.7	Proofs From Section 1.5	186
A.7.1	Preliminaries A: Asymptotic Properties of Erlang Formulae Under Linear Staffing	186
A.7.2	Preliminaries B: Limiting Idle Time and Derivative of Idle Time	198
A.7.3	Preliminaries C: Properties of the Limiting FOC	218
A.7.4	Preliminaries D: Auxiliary Definitions	220
A.7.5	Proof of Lemma 5	232
A.7.6	Proof of Lemma 6	233
A.7.7	Technical Details for Footnote 7	246
A.7.8	Proof of Proposition 5	248
A.7.9	Proof of Theorem 4	249
A.7.10	Proof of Proposition 6	258
A.7.11	Proof of Proposition 7	261
A.7.12	Proof of Proposition 8	275
A.7.13	Proofs of Theorems 5A and 5B	278
A.8	Proofs from Section 1.6	296
A.8.1	Proof of Proposition 9	296
APPENDIX B	APPENDIX FOR CHAPTER 2	299
B.1	Proofs from Section 2.3	299
B.1.1	Preliminaries	299
B.1.2	Proof of Lemma 7	300
B.1.3	Proof of Lemma 8	301
B.1.4	Proof of Theorem 6	306
B.1.5	Proof of Proposition 10 (i)	312
B.1.6	Proof of Proposition 10 (ii)	315
B.2	Proofs from Section 2.4	332
B.2.1	Proof of Theorem 7	332
B.3	Proofs from Section 2.5	332

B.3.1	Proof of Proposition 11	332
B.3.2	Proof of Lemma 9	333
B.3.3	Proof of Lemma 10	340
B.3.4	Proof of Lemma 11	343
B.3.5	Proof of Proposition 12	352
B.4	Proofs from Section 2.6	352
B.4.1	Justification of the Non-Idling Assumption in the Multiclass $M/M/N+M$ Queue	352
B.5	Proofs from Section 2.7	354
B.5.1	Preliminaries	354
B.5.2	Proof of Proposition 13	355
APPENDIX C APPENDIX FOR CHAPTER 3		360
C.1	The Fluid Model for γ	360
C.2	Proofs of Lemmas	362
C.2.1	Proof of Lemma 12	363
C.2.2	Proof of Lemma 13	363
C.2.3	Proof of Remark 13	366
C.2.4	Proof of Lemma 14	366
C.2.5	Proof of Lemma 16	367
C.2.6	Proof of Lemma 17	367
REFERENCES		370

LIST OF FIGURES

1	Interchange of Limits Diagram	8
1.1	Properties of equilibria, when servers are paid p per job completion, value idleness at $v = 10$ per unit time, and incur effort-cost $c(\mu) = \mu^3 + \mu$ per unit time, for $k = 3N$ and two values of the arrival rate λ	11
1.2	The tagged server's utility $U(\mu_1, \mu^{*\dagger})$ when $\lambda = 2.3$, $N = 2$, $p = 0$, $v = 1$, and $c(\mu) = \frac{3}{32}\mu^2$ for different values of k , where $\mu^{*\dagger}$ solves (1.6).	23
1.3	Server equilibria (solid lines) and the associated server utilities (dashed lines) as a function of k , when $N = 2$, $p = 0$, $v = 1$, and $c(\mu) = \frac{3}{32}\mu^2$	24
1.4	Server equilibria as a function of the staffing level N , when $\lambda = 20$, $v = 10$, and $c(\mu) = \mu^{1.5} + 0.1\mu$, for three values of p . The blue dots represent equilibria for integer values of N . μ_{\max}^* is given by Proposition 3.	26
1.5	Equilibria (solid) and local maxima that are not equilibria (dotted) when $v = 10$ and $c(\mu) = \mu^2 + 0.1\mu$, for two values of p . The $\mu^*(k)$ values in the lower portion of panel (b) are separated for illustration.	30
1.6	Existence of limiting equilibria, when $v = 10$ and $c(\mu) = \mu^3 + \mu$	36
1.7	Limiting equilibria as a function of the staffing parameter, when $v = 10$ and $c(\mu) = \mu^3 + \mu$. $p^\dagger(v) = 3.33$ and $p^\ddagger(v) = 4.58$	39
1.8	Limiting equilibria as a function of payment, when $v = 10$ and $c(\mu) = \mu^3 + \mu$. $\bar{a}(0, v) = 0.545$ and $\bar{a}(p^\dagger(v), v) = 0.882$	40
1.9	Behavior of prelimit equilibria under different payments and scalings of the staffing rule, when $k^\lambda = 3N^\lambda$, $v = 10$, and $c(\mu) = \mu^3 + \mu$	43
1.10	Behavior of prelimit equilibria under different scalings of the system size, when $p = 0$, $v = 10$, and $c(\mu) = \mu^2 + 0.1\mu$. The first row corresponds to Theorem 5B(b)(i) and the second to Theorem 5A(b). The horizontal blue and red dot-dashed lines correspond to the limiting equilibria (when they exist). The jaggedness is due to the discrete nature of N^λ and k^λ	44
1.11	Performance metrics as a function of the system size, when $N = 2$, $p = 0$, $v = 1$, and $c(\mu) = \frac{3}{32}\mu^2$, shown only when $\mu^*(\lambda, k, N, p, v)$ exists.	49
2.1	The multiclass $GI/GI/N+GI$ queue with abandonment cost	62
2.2	A graphic representation of the state space measures for a given class $j \in [J]$	64
2.3	The suboptimality gap of the benchmark policy $\pi_{a\mu}$ with respect to π_T^* and π^*	69
2.4	Illustration of the LTS policy, $\pi_{LTS}(\tau)$, that employs the LTS algorithm defined by Algorithm 1.	74
2.5	Comparison of the cumulative abandonment costs under the optimal MDP policy, the benchmark $a\mu$ -rule, and the proposed LTS policy, with 95% confidence intervals. In this figure, $\lambda_1 = 2N$, $\lambda_2 = 4N$, $\mu_1 = \mu_2 = 4$, $\theta_1 = 1$, $\theta_2 = 4$, $a_1 = 1$, $\theta_2 = 0.4$	85

2.6	Regret of the LTS policy in the $M/GI/8+GI$ queue, for different service time and patience time distributions, with 95% confidence intervals. In this figure, $\mu_1 = \mu_2 = 4$, $\theta_1 = 1$, $\theta_2 = 4$, $a_1 = 1$, $\theta_2 = 0.4$. In panel (a), the patience time distributions for class 1 and class 2 are Erlang-2(2) and Erlang-2(0.5). In panel (b), the service time distributions for class 1 and class 2 are both Erlang-2(0.5). Note: LN denotes the log-normel distribution.	87
A.1	Birth-death process for the $M/M/1/k$ system	132
A.2	Birth-death process for the $M/M/N/k$ system	132
A.3	Equivalent birth-death process for the $M/M/N/k$ system	133
A.4	Illustration of the left-hand side (solid blue curve) and the right-hand side (solid black and solid red curves) of the limiting FOC (A.109), suppose $p \leq c'(0)$.	218
A.5	Illustration of $\mu^\dagger(a, p)$, $p^\dagger(v)$, $p^\dagger(v)$ and $\bar{a}(p, v)$. Note that $v = v^\dagger(\bar{a}(p, v), p)$ in (I) and (II).	223
A.6	Illustration of the right-hand side (solid black and solid red curves) of the limiting FOC (A.139).	251
A.7	Illustration of the left-hand side (solid blue curve) and the right-hand side (solid black and solid red curves) of the limiting FOC (A.139) when $p < \frac{v}{2a}$ and $0 \leq p < \min\{c'(a), p^\dagger(v)\}$.	253
A.8	Illustration of the left-hand side (solid blue curve) and the right-hand side (solid black and solid red curves) of the limiting FOC (A.139) when $p \geq \frac{v}{2a}$ and $p > c'(a)$.	255
A.9	Illustration of the left-hand side (solid blue curve) and the right-hand side (solid black and solid red curves) of the limiting FOC (A.139) when $p < \frac{v}{2a}$ and $p > c'(a)$.	256
A.10	Illustration of the left-hand side (solid blue curve) and the right-hand side (solid black and solid red curves) of the limiting FOC (A.139) when $c'(0) < p < p^\dagger(v)$ and $a = (c')^{-1}(p)$.	257
A.11	Illustration of the left-hand side (solid blue curve) and the right-hand side (solid black and solid red curves) of the limiting FOC (A.139) when $p \geq p^\dagger(v)$ and $a \geq (c')^{-1}(p)$.	258
A.12	Illustration of the left-hand side (solid blue curve) and the right-hand side (solid black and solid red curves) of the limiting FOC (A.146) when underloaded equilibria exist.	263
A.13	Existence of the solutions to the prelimit FOC (A.173) in Theorem 5B.	283

ACKNOWLEDGMENTS

First and foremost, I owe my deepest gratitude to my esteemed advisor, Professor Amy Ward, whose unwavering support, meticulous guidance, and invaluable investment of time have profoundly shaped the essence of this thesis. Her exceptional research ethos, characterized by a rare blend of rigor, care, and passion, has left an indelible mark on my academic journey.

I am deeply indebted to my dissertation committee members: Professor John Birge, Professor Raga Gopalakrishnan, and Professor Ozan Candogan. My sincere gratitude goes to Professor John Birge for his sharp intellect, insightful guidance, and unwavering kindness, which have not only shaped my research interests but also contributed significantly to my personal growth. His advice and anecdotes have served as a wellspring of inspiration and guidance for me. I am deeply grateful to Professor Raga Gopalakrishnan for his incredibly generous investment of time in discussing technical details with me. His scholarly commitment, dedicated attitude, remarkable intelligence, and unwavering support have continually inspired me, fostering a bond that extends beyond the academic realm. A special note of gratitude is extended to Professor Ozan Candogan, whose research contributions and personal character were pivotal in my decision to pursue my PhD at Chicago Booth. Our interactions have consistently left me inspired and motivated. His infectious passion and insightful feedback have had a lasting impact on my academic pursuits. I am especially grateful for the collaborative work on Chapter 3 with Professor Amber Puha, whose passion for mathematics has been truly enlightening. Our collaboration has not only deepened my understanding of the subject but also cultivated a profound appreciation for the intricate beauty and depth within mathematical research.

In addition to my esteemed committee members, I consider myself truly fortunate to have had the opportunity to learn from many exceptional professors at Chicago Booth. My heartfelt gratitude extends to Professor Baris Ata, Professor Dan Adelman, Professor Rene Caldentey, Professor Varun Gupta, Professor Linwei Xin, and Professor Yuan Zhong for their

invaluable guidance and constant encouragement throughout various stages of my academic journey. Their contributions have profoundly enriched my PhD studies. I am also immensely thankful to Professor Levi DeValve, Professor Rad Niazadeh, and Professor Christopher Ryan for their meticulous guidance and unwavering support during my job search.

I am also grateful for mentors and collaborators outside Chicago Booth, especially Yee-Man Bergstrom, for her insightful career guidance. I would also like to extend my thanks to the dedicated staff members of the PhD office, including Malaina Brown, Cynthia Hillman, Kimberly Mayer, Amity James, for creating a supportive community.

My sincere thanks go to my fellow PhD classmates at Chicago Booth and the dear friends I have made along the way, including Mohammad Reza Aminian, Lisa Hillas, Lisa Li, Fabricio Previgliano, Gulin Tuzcuoglu, Naz Yetimoglu, and Alex Zhao, among many others. Their unwavering assistance and constant support have made my academic journey not only fulfilling but also enjoyable. I reserve special recognition for my best friend, Gorkem Unlu, for her unconditional support and encouragement throughout this journey. I am also profoundly appreciative of my office mate, Cong Zhang, whose patience, encouragement, and kindness have been my daily pillars of strength and bright spots.

Last but not least, my heartfelt gratitude goes to my family, whose unwavering support from afar has been my driving force and a constant source of inspiration. My mother, Wenshuo Chen, and my father, Yingmin Zhong, have instilled in me a courageous and upright research attitude, along with invaluable life wisdom. My uncle, aunt and cousin, Wenqi Chen, Jun Ye, and Yiyang Chen, have consistently been my ardent cheerleaders. My grandmother, Meifen Xu, has taught me the values of diligence, integrity, and perseverance in research. Lastly, I extend my deepest love and gratitude to my grandparents, Meijuan Wu, and Yongchang Chen, who have nurtured and accompanied me for over 20 years, guiding me through my entire academic journey. I cannot fathom who I am today, both physically and mentally, my personality and character, without their dedicated care and guidance.

ABSTRACT

Service system design is often informed by queueing theory, which helps system managers understand the impact of managerial design decisions on system performance. Traditional queueing theory assumes that servers are inanimate entities and system characteristics are perfectly known. However, these assumptions do not hold generally in modern service systems, where human servers exhibit strategic behavior and system characteristics may not be fully accessible. This thesis accounts for the complexities of human decision-making and the uncertainties of operational environments by integrating human server behavior and statistical learning into classical many-server queueing models, providing a framework for the analysis and optimization of behavior-aware and prediction-driven modern service systems.

In Chapter 1, we develop a game-theoretic model to investigate how human server work speed is affected by managerial decisions concerning (i) how many servers to staff and how much to pay them, and (ii) whether and when to turn away customers. We do this in the context of a finite-buffer many-server Markovian queue, where each server selfishly chooses her work speed in order to maximize an expected utility that captures an inherent trade-off between payment, idleness, and effort cost. Then, the work speeds emerge as the Nash equilibrium to a noncooperative game. We establish results on equilibrium existence and uniqueness, and demonstrate non-monotonic behavior. These results indicate that the commonly accepted rule of thumb that reducing workload decreases customer waiting time can be flawed due to servers adapting their work speeds in response to managerial incentives.

Chapter 2 studies a learning variant of a canonical scheduling problem (that is to decide which customer to serve when a server newly becomes available) in a multiclass many-server general queue with abandonment, when system characteristics (that is, distributional and parameter information regarding the inter-arrival, service, and patience times) are unknown and may be learned, and abandonments are costly. The scheduling question is of fundamental importance because scheduling determines which customer classes have longer wait times,

and, therefore, more abandonments. The difficulty is that even when system characteristics are known, characterizing an optimal scheduling policy appears intractable because the state space is very complex. Fortunately, the simple $a\mu$ -rule (that prioritizes classes for service in accordance with their cost of abandonment times service rate) solves an associated static scheduling problem that ignores system variability, and is asymptotically optimal (under certain conditions) for large systems that do not have sufficient capacity to serve all customers. Then, we only need to learn the static priority ranking of the classes given by the $a\mu$ -rule, that is based on first-order means, resulting in a significantly simplified learning problem. We propose a Learn-then-Schedule policy that first learns the unknown service rates and then schedules according to the empirical $a\mu$ -rule, which we show achieves the smallest achievable regret relative to the $a\mu$ -rule, that is of order $\log T$ (where T is the system run-time).

In Chapter 3, we delve into a control problem in the context of a single-class many-server general queue with abandonment. The objective is to strike a balance between operational costs (specifically, abandonment and holding costs) and human server utilization costs (stemming from fatigue). The control question revolves around determining when an available server should commence serving the next customer and when they should take a break. Our analysis of this control problem for large systems motivates that non-idling service disciplines are not in general optimal. This finding aligns with the understanding that overburdening employees can have adverse affect on their well-being and hinder organization growth. To address this, we propose an admission control policy designed to ensure that servers have sufficient idle time. We show that this policy is asymptotically optimal as time as well as the arrival rate and number of servers grow large.

CHAPTER 0

INTRODUCTION

The modern economy is increasingly driven by services. In 2021, services contributed to over 75% of the U.S. GDP, spanning sectors such as retail, healthcare, transportation, finance, and government.¹ As the demand for services continues to rise, a growing number of customers contend for limited resources, resulting in longer wait times. In fact, studies indicate that Americans collectively spend a staggering 37 million hours each year waiting for services.² While it is unrealistic to completely eliminate wait times, our vision is to significantly reduce them through more efficient system design.

Service system design is often informed by queueing theory, which helps system managers understand the impact of design decisions on system performance. However, traditional queueing theory often makes simplifying assumptions that do not generally hold in modern service systems. For example, it assumes that servers are inanimate entities and that system characteristics are perfectly known. In reality, modern service systems involve human servers who exhibit strategic behavior, and acquiring complete knowledge of system characteristics *a priori* may be unattainable. Consequently, service system design based on traditional queueing theory may overlook the impact of human server behavior and unknown system characteristics, leading to unintended outcomes.

The overarching goal of this thesis is to establish novel managerial insights into the optimal design of modern service systems by integrating human server behavior and unknown system characteristics into queueing theory. We particularly focus on many-server queueing models, where incoming jobs or customers may be served by one of multiple servers. This integration is a nontrivial challenge. Firstly, although empirical research has extensively explored various facets of individual server behavior in queueing systems, there is a limited

1. <https://data.worldbank.org/indicator/NV.SRV.TOTL.ZS?end=2021&locations=US&start=1997>

2. <https://waitwhile.com/assets/pdf/waiting-in-line-consumer-survey.pdf>

body of theoretical literature that accounts for rational, self-interested server behavior in queueing models. Secondly, the advancements in statistical learning tools offer opportunities for crafting accurate predictions. However, the intersection of queueing theory and statistical learning is still in its infancy. It is precisely these gaps that this thesis aims to fill, taking the first steps towards developing new theory and formal models for the analysis and optimization of modern service systems.

Specifically, we develop a comprehensive understanding of system performance, enabling us to identify (near)-optimal system design for managerial objectives such as welfare maximization and economic cost minimization. We envision smart modern service systems adopting behavior-aware and prediction-driven designs to enhance their operational and financial performance. The system design considerations explored in this thesis encompass a wide range of systemic and monetary policies, including admission control, queue discipline, customer routing policy, server staffing policy, pricing of services, and server payment schemes. The theoretical findings presented in this thesis have the potential to yield invaluable managerial insights and actionable recommendations for real-world modern service systems across various sectors, including private enterprises, public services, and non-profit organizations.

This thesis is organized into three main chapters to tackle the challenges posed by modern service systems from distinct perspectives. Below, we provide an overview of each chapter.

Chapter 1: Marrying Queueing with Human Behavior

Traditional queueing theory assumes that servers work at constant (possibly heterogeneous) rates, or speeds. That is reasonable in computer science and manufacturing contexts. However, the servers in service systems are people, and, in contrast to machines, their work speed can be influenced by the incentives created by design decisions. Motivated by a wealth of empirical research that documents servers speeding up or slowing down behavior, it is imperative for system managers to contemplate:

How is human server behavior affected by service system design?

This research question is of paramount importance since the rule of thumb based on non-strategic queueing models may not carry over to behavior-aware queueing settings.

In Chapter 1, we develop an analytical model to investigate how server work speed is affected by managerial decisions concerning (i) how many servers to staff and how much to pay them, and (ii) whether and when to turn away customers in the context of many-server queues with finite or infinite buffers (specifically, an $M/M/N/k$ queue with $k \in \mathbb{Z}_+ \cup \{\infty\}$) in which the work speeds emerge as Nash equilibrium to a noncooperative game. Each server selfishly chooses how fast to work in order to maximize her expected utility, which captures the inherent trade-off between payment, idleness, and cost of effort.

We show that a symmetric equilibrium always exists in a loss system ($N = k$) and provide conditions for equilibrium existence in a single-server system ($N = 1$). For the general $M/M/N/k$ system, we provide a sufficient condition for the existence of a solution to the first-order condition and bounds on such a solution; however, showing that it is an equilibrium is challenging due to the existence of multiple local maxima in the utility function. Nevertheless, in an asymptotic regime in which the arrival rate and number of servers become large, the utility function becomes concave, allowing us to characterize underloaded, critically loaded, and overloaded equilibria. We find that strategic servers may exhibit speedup and/or slowdown behavior in response to (i) staffing level (i.e., how many servers to staff), and (ii) payment (i.e., how much money to pay servers). Such behavior may lead to surprising system performance that is not possible in the traditional, non-strategic setting. For example, the system load can increase both when the arrival rate increases and when the arrival rate decreases, which suggests that the commonly accepted rule of thumb that reducing workload decreases customer waiting can be flawed.

Our model provides a tool for managers to trade off fixed and variable costs of staffing given a service level target. In order to predictably control system performance (e.g., lost

demand, customer wait times, server burnout, etc.), the system manager should jointly consider staffing and payment to ensure equilibrium existence and optimize equilibrium behavior. In particular, the system manager must either staff enough servers or pay them enough; otherwise, when servers are not paid enough, increasing workload beyond a tipping point may result in a sharp drop in system performance due to server “rebellion”. This research constitutes key foundational building blocks to advance the analysis and optimization of behavior-aware queueing models where both customers and servers are strategic and customers’ decisions endogenously induce a finite buffer. Beyond its immediate scope, this research bridges the gap between empirical and theoretical literature in service science, stimulating the development of innovative behavioral queueing models and fostering a robust research ecosystem that integrates theory and practice.

Chapter 1 assumes that the system and entities’ characteristics are fully known, a premise that often does not align with real-world scenarios. This concern propels the motivation for Chapter 2, where we explore the intersection of queueing theory and statistical learning.

Chapter 2: Marrying Queueing with Statistical Learning

The system characteristics of modern service systems are not always fully known and may need to be learned. A lack of (distributional and/or parameter) information may lead to system inefficiency, as effective system design requires a comprehensive understanding of these characteristics. Therefore, it is crucial to develop approaches that utilize observable information, obtained through the choices of decision-makers, to reveal hidden system characteristics for optimal system design. To this end, we seek to answer the following question:

How can statistical learning be integrated into service system design?

With the growing availability of data and rapid advancements in data-related fields such as artificial intelligence and machine learning, this research question holds significant promise

and practical applicability. Nonetheless, a notable challenge lies in the intricate interplay between learning and queueing. Queue congestion hampers efficient exploration in learning, while inefficient learning exacerbates queue congestion.

In Chapter 2, we study a learning variant of a canonical scheduling problem in a multiclass many-server queue with abandonment (specifically, the multiclass $GI/GI/N+GI$ queue), when the system characteristics (that is, distributional and parameter information regarding the inter-arrival, service, and patience times) are unknown, and abandonments are costly. The scheduling question is to decide which customer to serve when a server newly becomes available. This operational question is of fundamental importance because scheduling determines which customer classes have longer wait times (relative to their patience when waiting for service), and, therefore, more abandonments. However, even though there is much work on scheduling in queueing systems, there is comparatively less work on scheduling in queueing systems when the system characteristics are unknown.

Our objective is to determine a scheduling policy that minimizes regret, which is the difference in expected abandonment cost between a proposed policy, that does not have knowledge of system characteristics, and a benchmark policy, that has full knowledge of system characteristics. The difficulty is that exact analysis appears intractable even when the system characteristics are known because the state space is very complex. In order for the system to be Markovian, we must track: (i) the time elapsed since the last arrival for each class; (ii) the amount of time each customer in service has been in service; and (iii) the amount of time each customer in queue has spent waiting. Fortunately, the simple $a\mu$ -rule (that prioritizes classes for service in accordance with their cost of abandonment times service rate) solves an associated static scheduling problem that ignores system variability, and is asymptotically optimal for large systems that do not have sufficient capacity to serve all customers when either the service time distributions or patience time distributions are exponential. Then, our task is to learn the service rates well enough to determine the static

priority ranking of the classes given by the $a\mu$ -rule. We propose a policy that first learns and then schedules following a Learn-then-Schedule (LTS) algorithm that we develop. The algorithm is composed of a learning phase, during which empirical estimates of the service rates are formed, and an exploitation phase, during which an empirical $a\mu$ -rule based on those estimates is applied. We show that the LTS policy achieves the smallest achievable regret relative to the $a\mu$ -rule, that is of order $\log T$ (where T is the system run-time).

This research contributes to the stream of literature on integrating statistical learning into queueing theory. Notably, it is the first to consider customer abandonment in the most general setting, namely, the multiclass $GI/GI/N+GI$ queue. Leveraging large-system asymptotic approximation, we uncover a simple structure of the problem solution, which is a static priority policy based on first-order means. As a consequence, the learning problem becomes significantly simplified. We believe that the use of asymptotic analysis to simplify learning problems with complicated state space has broad applicability. It is our hope that this research serves as an inspiration for the adoption of this approach in diverse applications.

One open challenge in Chapter 2 is the asymptotic optimality of the $a\mu$ -rule when neither patience times nor service times follow exponential distributions. The underlying difficulty is associated with the absence of results that provide sufficient conditions for fluid model solutions to converge to fluid model invariant states in the time infinity limit. Addressing this technical gap requires fundamental advancements, despite ongoing efforts by numerous scholars over the years. To tackle this technical challenge, Chapter 3 takes a slight step back, focusing on a simplified single-class $GI/GI/N+GI$ queue.

Chapter 3: Marrying Human Behavior with Economic Costs

Chapter 3 extends the analysis in Chapter 1 to understand how human server behavior impacts service system performance and influences system design decisions by integrating server effort cost, arising from fatigue, into the economic cost structure. In parallel with

the scheduling problem studied in Chapter 2, this chapter addresses an analogous optimization problem within a single-class context. It operates under the assumption of complete knowledge regarding system characteristics, rendering learning processes unnecessary.

The control question pertains to determining when a newly available server should commence serving the next customer and when it should remain idle for a period. Too much idleness may lead to customer abandonment and excessive waiting, whereas too little idleness can increase server utilization costs. Therefore, the objective of the control problem is to trade off the long-run average operational costs (specifically, abandonment costs, and also, as an extension, holding costs) with server utilization costs. To solve the control problem, we consider a large-system many-server asymptotic regime in which the arrival rate and number of servers grow large. The solution to an associated fluid control problem motivates that non-idling service disciplines are not in general optimal, unless some arrivals are turned away. We propose an admission control policy designed to ensure that servers have sufficient idle time, which we show is asymptotically optimal. This policy strikes a balance between optimizing the operational efficiency of the service system and safeguarding the well-being of human servers, which is of profound importance for long-term growth of service operations.

In addition to its insightful managerial implications, this research holds paramount technical importance, offering a systemic program or guidebook for leveraging fluid approximation as a powerful tool to establish the asymptotic optimality of specific policies. These results are typically supported by a set of convergence results, as shown in Figure 1 for the asymptotic optimality result in this chapter. The interchangeability of the fluid limit (by letting the arrival rates and number of servers grow large) and the time limit (by letting time grow large) justifies regarding fluid model invariant states as first-order approximation to the stationary distributions of the stochastic system. By first taking the fluid limit and then the time limit, we navigate through the steps (1) and (2) in Figure 1. Step (1) shows the convergence of fluid-scaled stochastic systems to a fluid model solution, serving as a

functional law of large numbers approximation to the stochastic system. In (2), we establish the convergence of fluid model solutions to an invariant state as time becomes large. Alternatively, by reversing the order of limits—first taking the time limit and then the fluid limit—we traverse steps (3) followed by (4) in Figure 1. Step (3) is supported by the existence of a stationary distribution in stochastic systems, and subsequently, in (4), we prove the convergence of fluid-scaled stochastic systems in stationarity to an invariant state of the fluid model. Crucially, step (4) relies on the long-time behavior of fluid model solution in (2), exactly the technical challenge confronted in Chapter 2 in the multiclass setting.

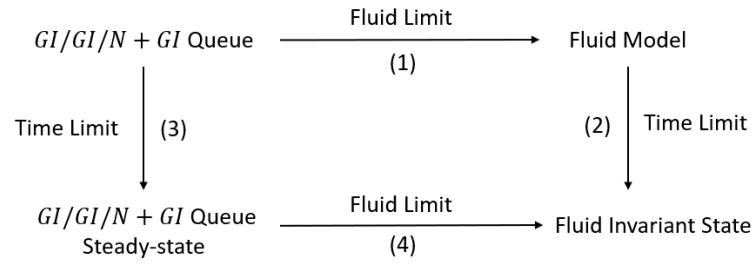


Figure 1: Interchange of Limits Diagram

CHAPTER 1

BEHAVIOR-AWARE QUEUEING: THE FINITE-BUFFER SETTING WITH MANY STRATEGIC SERVERS

1.1 Introduction

Americans spend 37 billion hours each year waiting for service, for instance, at coffee shops, public transit stops, airports, in traffic, and on customer service calls (Stone, 2012). According to traditional queueing theory, the ability to flatten peak demand for any of those services (by, for example, shifting demand to different times or to substitute services) should result in decreased waiting. However, empirical evidence shows that employees may serve customers faster or slower in response to congestion, affecting waiting. For example, one very early data-based study of traffic delay at the George Washington Bridge connecting Manhattan and New Jersey (see Figure 7 in Edie (1954)) shows that the average time a vehicle spends at a toll booth decreases as the arrival rate of vehicles increases. One of the field observations that explains such behavior is that toll collectors generally spend less time chatting with drivers when congestion is high. As a result, toll collector capacity increases, potentially resulting in the seemingly magical ability to simultaneously increase demand and decrease waiting. Later empirical studies have further substantiated this relationship between congestion and service speed (see Table 2 in Kc and Terwiesch (2009), Table 2 in DeHoratius et al. (2020), Figures 4.2 and 4.3 in Lu (2013), and the survey paper Delasay et al. (2019)).

Traditional queueing theory (see, e.g., Shortle et al. (2018)) assumes servers work at exogenous rates. This is appropriate when servers are machines. However, in service systems, which accounted for over 70% of the Gross Domestic Product in the United States in 2019/Q4 (Bureau of Economic Analysis, 2020), servers are often human, and humans respond to incentives. Workload creates one such incentive, and the review paper by Delasay et al. (2019) provides a framework for understanding reasons for servers working faster or slower

in response to workload. The design decision of whether or not to turn away customers (e.g., when faced with high congestion) impacts workload. The question of how many servers to staff produces another incentive, because working collectively to reduce congestion can affect the service rate. And, of course, payment also impacts the service rate. *Our focus in this chapter is to understand the effect of managerial decisions concerning customer admission, staffing, and payment on the server decision of how fast to work.* When customer admission decisions are implemented via a threshold policy on the number of waiting customers, the system operates as a finite-buffer queue.

We study many-server queueing systems with finite or infinite buffers (specifically, an $M/M/N/k$ model with $k \in \mathbb{Z}_+ \cup \{\infty\}$), except we do not assume fixed service rates that are exogenously given. Instead, we assume that each server selfishly chooses how fast to serve in order to maximize her expected utility. The three components driving a server's utility are a piece-rate payment, the amount of idle time, and the cost of effort expended. Working faster can lead to more idle time and higher payment; however, working faster also incurs a higher cognitive cost (more effort). Our utility function captures this inherent trade-off between payment, idleness, and cost of effort, and allows for each component to be weighted as desired.

1.1.1 Contributions of This Chapter

Service system performance measures such as the probability that a customer must wait for service, and, if so, their waiting time, depend, among other factors, on how fast the servers work. In a service system that is staffed with self-interested human servers who strategically choose their service rates, we seek to understand when one or more equilibrium service rates exist, characterize them when they do, and predict their sensitivity to changes in system parameters. The first step of equilibrium analysis involves solving a first-order condition on the servers' utility function. Towards this, we provide a sufficient condition for the existence

of a solution to the first-order condition (Theorem 1). The issue is that the first-order condition may admit multiple solutions, and more than one of those solutions can be a local maximum with nonnegative utility (Figure 1.2). As a result, an exact equilibrium analysis becomes challenging. Figure 1.1 illustrates the complex equilibrium behavior numerically in the (staffing, payment)-parameter space, for two values of the arrival rate.

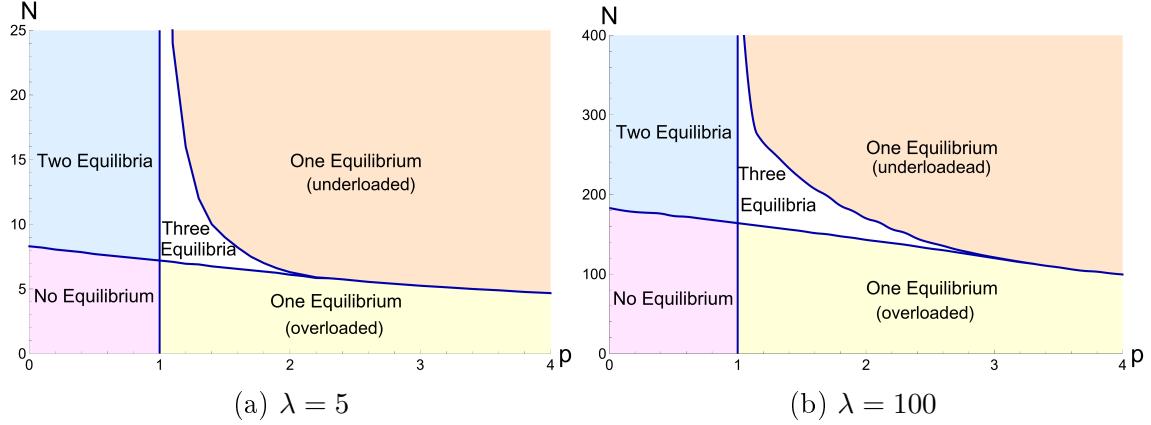


Figure 1.1: Properties of equilibria, when servers are paid p per job completion, value idleness at $v = 10$ per unit time, and incur effort-cost $c(\mu) = \mu^3 + \mu$ per unit time, for $k = 3N$ and two values of the arrival rate λ .

The difficulty of characterizing equilibria motivates us to first analyze two specialized finite-buffer systems, enabling us to study the effects of N and k separately: (a) the loss system ($N = k$), where there are no waiting customers, and (b) the single-server system ($N = 1$), where there is no competition between servers. While solutions to the first-order condition are always equilibria in the loss system (Theorem 2), that is not the case in the single-server system (Proposition 4, Theorem 3, and Figure 1.5). In the loss system, we are also able to characterize the maximum equilibrium service rate (Proposition 3), and examine a variant of our utility function that captures diminishing returns of idle time (Proposition 9).

In order to derive results for more general systems, with both many servers and buffers to hold waiting customers, we undertake an asymptotic analysis in which we allow the customer arrival rate to become large. In that asymptotic regime, a solution to the first-order condition is an equilibrium (Proposition 5) for all large enough arrival rates. We

derive a limiting first-order condition, establish convergence results to show that solutions to the limiting first-order condition approximate prelimit equilibria (Theorems 5A-5B and Figure 1.9), provide numerical examples to show the convergence rate (Figure 1.10), and characterize the maximum limiting equilibrium service rate (Proposition 8).

Throughout, we investigate the following key attributes of equilibrium behavior.

- *Uniqueness:* Multiple equilibria are possible only when there is more than one server, suggesting that competition between servers is a key driver of this phenomenon. In such a setting, multiple equilibria are observed when the effective load on a server (determined by the number of customers that eventually reach a server) is neither too large nor too small (e.g., Figures 1.3 and 1.4). Intuitively, this is because servers have more flexibility in choosing to either work faster and gain utility through increased idle time and payment, or work slower and gain utility through reduced effort-cost. Fortunately, when multiple equilibria exist, servers obtain a higher utility when they work at a faster service rate (Propositions 2 and 6).
- *System Load:* Equilibria that emerge can result in an underloaded system, meaning there would be enough capacity to handle all arriving customers in an infinite-buffer ($k = \infty$) setting; an overloaded system, meaning the number-in-system process in an infinite-buffer setting would grow unboundedly; or a critically loaded system, meaning the capacity and arrival rate are equal. An overloaded or critically loaded equilibrium may be undesirable because the number of customers turned away due to a full buffer could be very large. Our asymptotic analysis allows us to characterize when solutions to the limiting first-order condition are underloaded, overloaded, or critically loaded (Theorem 4 and Figure 1.6).
- *Monotonicity:* Non-monotonic behavior of equilibrium service rates as a function of the arrival rate, buffer size, number of servers, or payment has important consequences

when thinking about system design decisions. In a traditional queueing system with fixed, exogenous service rates, increasing either the buffer size or the number of servers results in turning away fewer customers. In contrast, when strategic servers exhibit non-monotonic behavior in equilibrium service rates (as shown numerically, e.g., Figures 1.3, 1.4, 1.5, and 1.7; and analytically in Proposition 7), the impact of increasing the buffer size or number of servers on performance metrics such as the system load or expected wait time is unclear (Figure 1.11). If servers work at the same rate or faster, less customers will be turned away, but if servers slow down, it is possible that even more customers will be turned away. Interestingly, even when equilibrium service rates are monotonic, the direction can be surprising; for example, increasing payment can cause an equilibrium service rate to decrease (Figure 1.8).

We demonstrate how our work can constitute key foundational building blocks to advance the analysis of behavior-aware queueing models: (1) by using its predictions to formulate empirically testable hypotheses, (2) in the context of optimal design and control, (3) by studying the interplay between strategic arrivals and strategic servers, and (4) by generalizing the utility function.

1.1.2 Literature Review

Our setting is one with nondiscretionary service requests, that is, tasks like collecting money at toll booths, checking out customers at cash registers, providing bank services at teller stations, and implementing COVID-19 drive-through testing at hospitals, in which subjective judgement cannot be used to reduce the number of tasks that must be completed to fulfill a request. We can therefore set aside the trade-off between speed and quality inherent in server decision-making in many service settings, the effects of which are modeled in Hopp et al. (2007) and empirically shown in Tan and Netessine (2014) and Shen et al. (2021).

There is a large body of empirical research supporting the dependence of service rate on

system load, for which we refer the reader to the survey paper Delasay et al. (2019). One complication is that system load also depends on service rate, meaning the dependence is bidirectional, which must be carefully accounted for in empirical studies. In comparison, an analytical model provides a structured mapping for the dependence, the results of which can be used to provide context for empirically testable hypotheses (which we briefly discuss in Section 1.6.1). We model the dependence via a queueing game that captures how system design decisions affect the willingness of servers to settle into a sustained service rate; i.e., an equilibrium endogenously determined in the steady state. The papers that model transient server speedup and slowdown effects (see, e.g., Chan et al. (2014); Dong et al. (2015); Delasay et al. (2016) and Do et al. (2018)) use exogenous, state-dependent Markovian models rather than the game-theoretic analysis used in this chapter.

Recent empirical research has shown that queue design has a nontrivial impact on service system performance in the nondiscretionary task setting. For example, queue visibility can impact how much effort servers are willing to put in to process orders (Rosokha and Wei, 2020). As another example, Song et al. (2015), Shunko et al. (2018), Wang and Zhou (2018), and Wang et al. (2022) all show that each server having its own dedicated queue can result in shorter wait times for customers as compared to pooling customers into a shared queue, which validates a conjecture in Rothkopf and Rech (1987). This is in stark contrast to the long-standing, conventional queueing wisdom that pooling is beneficial (in homogeneous customer and server environments), when service rates are exogenous (see, e.g., Shortle et al. (2018)). The analytical work in Armony et al. (2021) uses a game-theoretic queueing model to explain the aforementioned observed empirical phenomenon. This chapter is written in a similar spirit, except that we are more focused on predicting how changes in system load and payment affect the service rate.

The service rates endogenously chosen by servers in our model are the solution to a queueing game. The large literature on queueing games is surveyed in Hassin and Haviv

(2003) and, more recently, in Hassin (2016) and Allon and Kremer (2018). Much of that literature assumes exogenous service rates and focuses on incorporating customer behavior into queueing models. Some exceptions are Kalai et al. (1992), Gilbert and Weng (1998), Cachon and Harker (2002), Christ and Avi-Itzhak (2002), Debo et al. (2008), and Geng et al. (2015b), but, in contrast to our model, the maximum number of servers in all these papers is two. The spirit of our asymptotic analysis is most similar to Ibrahim (2018), Dong and Ibrahim (2020), Gopalakrishnan et al. (2016a), and Zhan and Ward (2019), all of which use large-system asymptotic analyses to study many-server queues in which servers have some decision-making power. The difference between the first two papers and this one is that a server's decision concerns whether or not to show up for work rather than how fast to work once present. The last paper uses payment to influence how hard the server works. The most closely related paper is the third one, which models the server decision identically when there is no payment, but assumes an infinite buffer; as a result, when faced with too much demand, servers rebel and refuse to work, rather than receive payment to continue working in overloaded conditions. (More specifically, overloaded equilibria do not arise in Gopalakrishnan et al. (2016a).)

1.1.3 The $M/M/N/k$ Queue with Non-strategic Servers

In a non-strategic $M/M/N/k$ queueing system, customers arrive to a system with N servers ($N \in \mathbb{Z}_+$) and system size $k \geq N$ ($k \in \mathbb{Z}_+ \cup \{\infty\}$) (equivalently, buffer size $k - N \geq 0$) according to a Poisson process with rate $\lambda > 0$. Customers that arrive to find all N servers busy choose to join the queue and customers that arrive to find $k - N$ customers waiting in the queue are lost. Waiting customers are served according to the first-come-first-served discipline. Servers only idle when no customers are waiting. The time required to serve each customer is independent and identically distributed according to an Exponential distribution with a mean of one time unit when the server works at rate one. When servers work

at heterogeneous rates, we denote the service rate of server i by $\mu_i > 0$, and the service rate vector by $\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)$. We denote by $I_i(\boldsymbol{\mu}; \lambda, k, N)$ and $B_i(\boldsymbol{\mu}; \lambda, k, N) = 1 - I_i(\boldsymbol{\mu}; \lambda, k, N)$ the long-run fraction of time that server i is idle and busy (henceforth referred to simply as “idle time” and “busy time”), given the service rate vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)$.

Definition 1. *In an $M/M/N/k$ queue with arrival rate λ , a homogeneous service rate μ is underloaded if $\lambda/N\mu < 1$, critically loaded if $\lambda/N\mu = 1$, and overloaded if $\lambda/N\mu > 1$.*

When more than one server is idle, we assume that the idle server that is assigned to the next customer is determined by an idle-time-order-based routing policy (that uses only the order in which the servers last became idle).¹ Examples of idle-time-order-based routing policies include uniformly random routing and Longest/Shortest Idle Server First.

Lemma 1. *In an $M/M/N/k$ queue, all idle-time-order-based policies result in the same steady-state probabilities.*

Lemma 1 allows us to simplify our analysis without loss of generality by assuming that the routing policy is uniformly random. Such a result has been shown in Theorem 9 in Gopalakrishnan et al. (2016a) for an infinite-buffer system and in Theorem 1 in Haji and Ross (2015) for a loss system.

Our analysis requires working with only a “mildly heterogeneous” $M/M/N/k$ system, in which a “tagged” server (say, server 1) works at rate μ_1 and all the other servers work at rate $\mu > 0$. We denote the idle time of the tagged server in such a system by $I(\mu_1, \mu; \lambda, k, N)$. When $k = \infty$, we define $I(\mu_1, \mu; \lambda, \infty, N) := \lim_{k \rightarrow \infty} I(\mu_1, \mu; \lambda, k, N)$.

Lemma 2 (Expression for Idle Time²). *In an $M/M/N/k$ queue ($k \in \mathbb{Z}_+$) where one server*

1. The steady-state analysis of $M/M/N/k$ systems with routing policies that depend on the service rates (e.g., Fastest Server First, Slowest Server First) is extremely challenging, even for $N = 3$ in the nonstrategic setting (Mokaddis et al., 1998). While it is certainly interesting to explore the impact of such routing policies as an additional aspect of system design, this is beyond the scope of this chapter.

2. When $k \rightarrow \infty$, Equation (1.1) can be seen to be identical to the expression for the server idle time in an infinite-buffer system, given in Theorem 1 in Gopalakrishnan et al. (2016a), provided the stability condition $\lambda < (N - 1)\mu + \mu_1$ is satisfied.

operates at rate $\mu_1 > 0$ and the other $N - 1$ servers operate at rate $\mu > 0$, the steady-state probability that the server operating at rate μ_1 is idle is given by

$$I(\mu_1, \mu; \lambda, k, N) = \left(1 + \rho \frac{\mu}{\mu_1} \left(\frac{1 - ErlC(N, \rho)}{N - \rho} + \left(1 - \left(\frac{\rho}{N - (1 - \frac{\mu_1}{\mu})} \right)^{k-N} \right) \frac{ErlC(N, \rho)}{(N - \rho) - (1 - \frac{\mu_1}{\mu})} \right) \right)^{-1}, \quad (1.1)$$

where $\rho = \frac{\lambda}{\mu}$, and $ErlC(N, \rho)$ denotes the Erlang C formula, given by

$$ErlC(N, \rho) = \left(\frac{\rho^N}{N!} \frac{N}{N - \rho} \right) \Bigg/ \left(\sum_{j=0}^{N-1} \frac{\rho^j}{j!} + \frac{\rho^N}{N!} \frac{N}{N - \rho} \right).$$

When $\lambda \geq (N - 1)\mu + \mu_1$, $I(\mu_1, \mu; \lambda, \infty, N) = 0$.

1.2 The Strategic Server Queueing Model

The setting we focus on is an $M/M/N/k$ system in which servers are players in a noncooperative game. Servers are paid based on individual service volume at p per service completion, and value their idle time at rate v per unit time. The function $c : [0, \infty) \rightarrow [0, \infty)$ captures the cost of effort per unit time required to serve at a certain rate, and is strictly increasing and strictly convex, with $c(0) = 0$. Each server $i \in \{1, \dots, N\}$ chooses her service rate $\mu_i \in (0, \infty)$ to maximize the utility function

$$U_i(\boldsymbol{\mu}; \lambda, k, N, p, v) := p \cdot \mu_i B_i(\boldsymbol{\mu}; \lambda, k, N) + v \cdot I_i(\boldsymbol{\mu}; \lambda, k, N) - c(\mu_i). \quad (1.2)$$

The utility function (1.2) captures the trade-off between payment, idleness, and effort that is a first-order determinant of the server experience. When $p = 0$, and $v = 1$, (1.2) is exactly the utility function that appears in Gopalakrishnan et al. (2016a) and Armony et al. (2021), which is consistent with a volunteer or fixed-wage service system. When $p > 0$, there is a piece-rate payment, as in (Kalai et al., 1992; Gilbert and Weng, 1998; Christ and Avi-Itzhak, 2002; Zhan and Ward, 2019).

The difficulty in decision-making for any server i is that although she has complete control over her service rate, unless $N = 1$, she cannot maximize her own utility in (1.2) without considering the choices of the other servers. This is because the utility of server i is determined by all components of the vector μ through the functions B_i and I_i . For the remainder of this chapter, we assume $N > 1$, unless explicitly specified.

The servers want to choose service rates $\mu = (\mu_1, \dots, \mu_N)$ that satisfy

$$U_i(\mu; \lambda, k, N, p, v) = \max_{\mu_i > 0} U_i(\mu_1, \dots, \mu_{i-1}, \mu_i, \mu_{i+1}, \dots, \mu_N; \lambda, k, N, p, v), \quad \forall i \in \{1, \dots, N\},$$

and, in particular, constitute a (pure) Nash equilibrium of the game. The focus of this chapter is on characterizing *symmetric* Nash equilibria $\mu^*(\lambda, k, N, p, v)$, given by

$$\mu^* \in \arg \max_{\mu_1 > 0} U(\mu_1, \mu^*; \lambda, k, N, p, v), \quad (1.3)$$

where

$$U(\mu_1, \mu; \lambda, k, N, p, v) := p\mu_1 + (v - p\mu_1) \cdot I(\mu_1, \mu; \lambda, k, N) - c(\mu_1) \quad (1.4)$$

denotes the utility of Server 1 when working at rate $\mu_1 > 0$, with all the other servers working at rate $\mu > 0$, in a “mildly heterogeneous” $M/M/N/k$ system. Because the servers in our model are homogeneous, in the sense that they share the same strategic behavior, their utility functions are symmetric; therefore, specializing to Server 1 in (1.3) is without loss of generality and the focus on *symmetric* Nash equilibria is natural. Throughout, the terms “equilibrium service rate,” “server equilibrium,” and simply “equilibrium” all refer to a symmetric Nash equilibrium service rate. We expose or suppress the dependence of the utility function, the idle time, and the equilibrium service rate on the parameters λ, k, N, p , and v as appropriate, for ease of exposition.

We do not impose an explicit individual rationality condition, because any equilibrium service rate, defined by (1.3), must be individually rational.

Proposition 1 (Individual Rationality). *In an $M/M/N/k$ system, any equilibrium service rate $\mu^* > 0$ must satisfy $U(\mu^*, \mu^*) \geq 0$.*

The implicit assumption is that the working environment is such that the servers do find showing up to work worthwhile; the question remaining is how much effort the servers will put in during their workday. The normalization $c(0) = 0$ is for convenience. For incorporating an outside option with strictly positive utility $M > 0$, the normalization would be $c(0) = -M$. Additionally, individual rationality implies an implicit upper bound on any equilibrium.

Lemma 3. *In an $M/M/N/k$ system, any equilibrium service rate μ^* satisfies $\mu^* \leq \mu_{\max}^*(p, v)$, where*

$$\mu_{\max}^*(p, v) = \begin{cases} \mu_0, & c'(0) < p \text{ and } c\left(\frac{v}{p}\right) < v \\ c^{-1}(v), & c'(0) < p \text{ and } c\left(\frac{v}{p}\right) \geq v, \\ \min\left\{\frac{v}{p}, c^{-1}(v)\right\}, & c'(0) \geq p \end{cases}$$

where μ_0 is the unique solution for $\mu > 0$ to $c(\mu) = p\mu$ when $c'(0) < p$.

The strategic server model presented in this section is fundamentally different than its non-strategic counterpart, because the system load may not change monotonically as a function of λ or N . The system load in the non-strategic queue, for a fixed and exogenous service rate μ , is $\lambda/N\mu$, which is increasing in λ and decreasing in N . In contrast, determining if the system load in the strategic queue, $\lambda/N\mu^*(\lambda, k, N)$, is increasing or decreasing in λ or N requires understanding how changes in those parameters affect the equilibrium service rate $\mu^*(\lambda, k, N)$. Moreover, the system load in the non-strategic setting does not depend on the system size k , whereas the system load in the strategic setting does.

The above observation has important consequences for system design because there is a generally recognized trade-off between the cost of reducing system load from, for example, having to turn away customers or hire more staff, and realizing better performance, as

measured, for example, by smaller expected wait time. In the non-strategic setting, hiring more staff increases N , which simultaneously decreases system load and decreases expected wait time. In contrast, in the strategic setting, an increase in N may or may not result in a decrease in system load, and may or may not result in a decrease in expected wait time. As a result, in the strategic setting, there is potential to improve performance and reduce costs by *decreasing* N . However, this is not possible to know without understanding the behavior of $\mu^*(\lambda, k, N)$, which is the focus of this chapter.

The questions we answer regarding the behavior of $\mu^*(\lambda, k, N)$ are:

- (Q1) *When does an equilibrium exist? When it exists, is it unique? If not unique, which equilibrium do the servers prefer?* If either (i) a unique equilibrium exists, or (ii) multiple equilibria and a justifiable equilibrium selection criterion exist, then system performance can be predicted using traditional queueing formulae.
- (Q2) *What circumstances result in an underloaded ($\lambda < N\mu^*$), critically loaded ($\lambda = N\mu^*$), or overloaded ($\lambda > N\mu^*$) system at equilibrium? How do equilibrium service rates behave as a function of the system design parameters k , N , and p ?* The most basic system design question concerns whether or not the capacity is sufficient to serve the arriving customers. Answering more sophisticated questions requires knowing if and when servers will speed up or slow down in response to changes in system parameters.
- (Q3) *What is the maximum possible equilibrium service rate?* When the equilibrium is close to its maximum,

$$\mu_{\max}^*(p, v) := \sup \{ \mu^*(\lambda, k, N, p, v) : (1.3) \text{ holds for some } \lambda, k, N \}, \quad (1.5)$$

system design changes focused on k and N are unlikely to be worthwhile. Therefore, the maximum possible equilibrium acts as a comparison benchmark for the system manager to decide whether or not investing resources to contemplate such changes is warranted.

Answering the above questions requires first working with (1.3) to characterize server equilibria, which involves (i) solving the corresponding first-order condition, and (ii) filtering out solutions that are not global maxima.

1.3 The $M/M/N/k$ System

We start with the first-order condition and its solutions in Section 1.3.1, and then we present some numerical examples of the general $M/M/N/k$ system in Section 1.3.2.

1.3.1 The First-Order Condition and Candidate Equilibria

We analyze the first-order condition (FOC) for an interior maximum μ_1 in (1.3). By definition, from (1.3), $\mu^* > 0$ is a symmetric equilibrium if and only if $U(\mu_1, \mu)$ attains an interior global maximum at $\mu_1 = \mu$. Therefore, any equilibrium service rate μ^* must solve the *symmetric* FOC:

$$\frac{\partial U(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} = 0 \Leftrightarrow c'(\mu) = p(1 - I(\mu, \mu)) + (v - p\mu) \frac{\partial I(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu}. \quad (1.6)$$

Solutions to the FOC (1.6) may or may not be equilibria, depending on whether or not they are global maxima of the corresponding utility functions. Therefore, we define solutions to the FOC separately from equilibria.

Definition 2. *Any solution for $\mu \in (0, \infty)$ to the symmetric FOC (1.6) is a candidate symmetric equilibrium, denoted by μ^* .*

The right-hand side of (1.6), which can be evaluated using Lemma 2, is complicated, rendering its analysis challenging. Still, as a first step, we can provide existence criteria for solutions to (1.6).

Theorem 1 (Sufficient Condition for Candidate Equilibria).³ In an $M/M/N/k$ system, if $c(\mu)$ satisfies

$$\lim_{\mu \downarrow 0} \frac{c'(\mu) - p}{\mu^{k-N}} < v \frac{k}{N^2} \left(\frac{N}{\lambda} \right)^{k-N+1}, \quad (1.7)$$

then the FOC (1.6) admits at least one solution $\mu^{\star?} \in (0, \infty)$. Furthermore, an entire function $c(\mu)$ satisfies (1.7) if and only if it is of the form

$$c(\mu) = c_E \mu^q \cdot h(\mu), \quad (1.8)$$

where $c_E \in \mathbb{R}$, $c_E \neq 0$, $q \in \mathbb{Z}_+$; h is an entire function with $h(0) = 1$; and, either (a) $q = 1$ and $c_E < p + \frac{v}{\lambda} \mathbb{1}\{k = N\}$, or (b) $q \geq 2$.

Remark 1. The first condition of (1.7) is meaningful only when $\lim_{\mu \downarrow 0} c'(\mu) - p = 0$ or $k = N$ because, when $k > N$, the inequality trivially holds when $\lim_{\mu \downarrow 0} c'(\mu) - p < 0$, and does not hold when $\lim_{\mu \downarrow 0} c'(\mu) - p > 0$. Interpreting the behavior of $c'(\mu)$ in the neighborhood of 0 as an indicator of a server's inherent resistance to being disturbed from an idle state, Theorem 1 provides that a candidate symmetric equilibrium exists, as long as the piece-rate payment p is large enough to compensate for, if not overcome, this resistance.

In the rest of this chapter, for the numerical examples, we focus on a special family of cost functions, motivated by the structural characterization (1.8) of Theorem 1:

$$c(\mu) = c_E \mu^q + d_E \mu, \text{ for some } q > 1, c_E > 0, \text{ and } d_E \geq 0. \quad (1.9)$$

This family of functions is obtained when the entire function (1.8) is $h(\mu) = 1 + \frac{c_E}{d_E} \mu^{q-1}$.

Much of our analysis needs the following assumption (implying $q \geq 2$ for (1.9)).

Assumption 1. $c'''(\mu)$ exists and is nonnegative for all $\mu > 0$.

3. If we consider a more general effort-cost function $c(\mu)$ that is continuous and differentiable for all $\mu \in [0, \infty)$, then an extra condition, $\lim_{\mu \uparrow \infty} \mu^2 c'(\mu) + p \mu \lambda (1 + \frac{1}{N}) > v \frac{\lambda}{N} - \frac{p}{2} \lambda^2$, is needed in (1.7), and then the form in (1.8) can be generalized to $c(\mu) = b_E + c_E \mu^q \cdot h(\mu)$, where $b_E \in \mathbb{R}$.

Although Theorem 1 provides a sufficient condition for a candidate equilibrium to exist, it is silent on whether or not a candidate equilibrium is an equilibrium. A candidate equilibrium $\mu^{*?}$ is an equilibrium if and only if its corresponding utility function $U(\mu_1, \mu^{*?})$ attains a global maximum at $\mu_1 = \mu^{*?}$. However, that is difficult to determine because $U(\mu_1, \mu^{*?})$ may have multiple local maxima, as illustrated in Figure 1.2 (similar to the behavior in Figure 5 in Netessine and Shumsky (2005)). In Figures 1.2a and 1.2b, $\mu^{*?}$ that solves (1.6) is a global maximizer of the function $U(\mu_1, \mu^{*?})$. However, in Figure 1.2c, $\mu^{*?}$ that solves (1.6) is a local, but not a global maximizer, and, therefore, is not an equilibrium. Because of this phenomenon, characterizing conditions for a candidate equilibrium $\mu^{*?}$ to be a global maximizer of its corresponding utility function $U(\mu_1, \mu^{*?})$ is hard. In the next subsection, we expand the numerical setting of Figure 1.2 by including additional values of λ and k in order to illustrate complex equilibrium characteristics that underscore the challenge of exact equilibrium analysis.

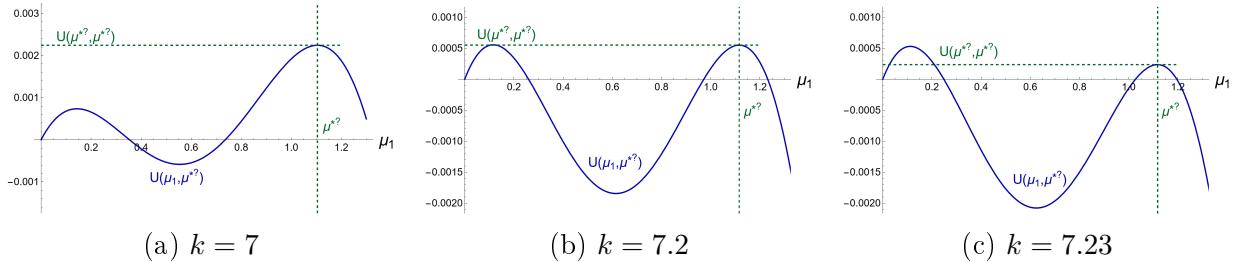


Figure 1.2: The tagged server's utility $U(\mu_1, \mu^{*?})$ when $\lambda = 2.3$, $N = 2$, $p = 0$, $v = 1$, and $c(\mu) = \frac{3}{32}\mu^2$ for different values of k , where $\mu^{*?}$ solves (1.6).

1.3.2 Numerical Examples

Figure 1.3 shows the equilibrium service rates and the corresponding server utilities in an $M/M/2/k$ system, when the piece-rate payment and the server's valuation for idleness are specialized to $p = 0$ and $v = 1$. Figure 1.3 highlights some of the challenges that arise in answering questions (Q1)-(Q3).

Equilibrium Existence and Uniqueness: There may be a unique equilibrium (Figure 1.3a)

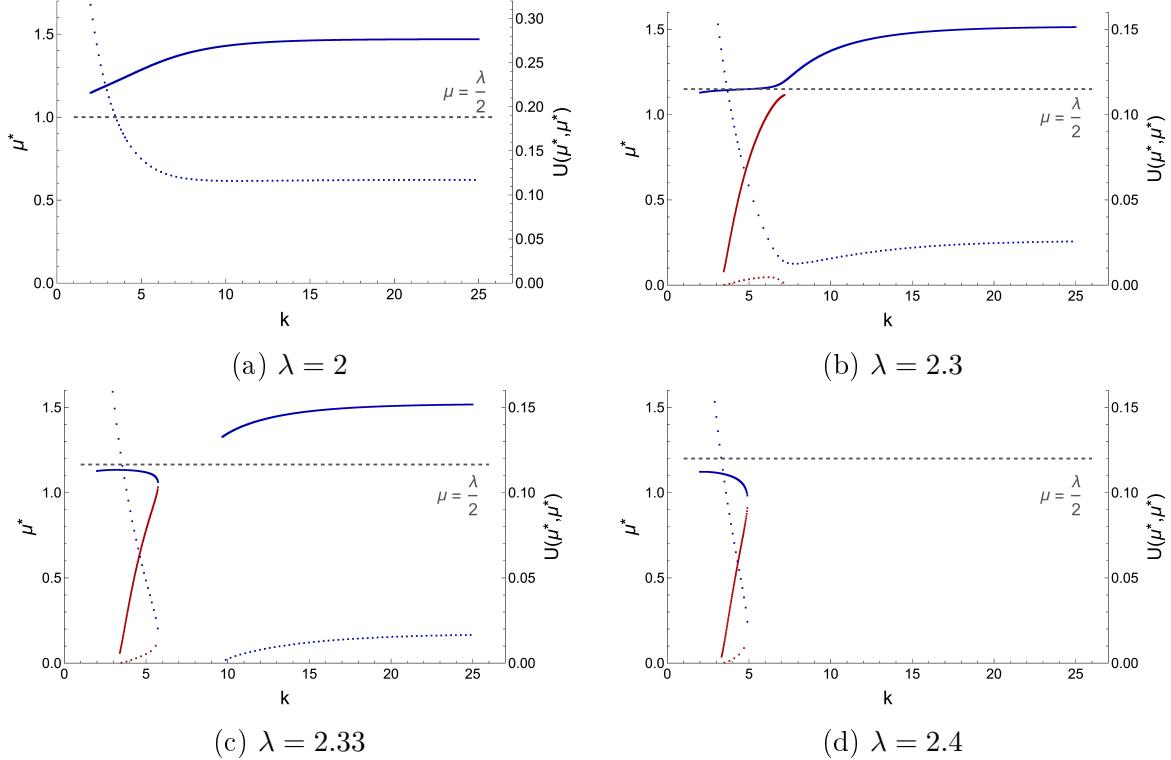


Figure 1.3: Server equilibria (solid lines) and the associated server utilities (dashed lines) as a function of k , when $N = 2$, $p = 0$, $v = 1$, and $c(\mu) = \frac{3}{32}\mu^2$.

for all k , and Figures 1.3b and 1.3c for larger k), multiple equilibria (Figures 1.3b-1.3d for smaller k), or no equilibria (Figure 1.3c for “medium-sized” k and Figure 1.3d for “medium-sized” and larger k). Multiple equilibria can arise when the servers gain sufficient utility either by working faster to increase their idle time or by working slower to save their effort-cost. When there are multiple equilibria, the faster equilibrium also results in higher utility, which is also the equilibrium preferred by the system manager. The discontinuity of equilibrium behavior in Figure 1.3c is surprising, and underscores the complexity of equilibrium behavior and the need for better understanding.

Equilibrium Behavior: As λ increases, server equilibria transition from being all underloaded (Figure 1.3a), to all overloaded (Figure 1.3d), as can be seen from the dashed gray lines in Figure 1.3. For smaller λ (Figures 1.3a and 1.3b), the equilibria are always increasing in k , whereas for larger λ (Figures 1.3c and 1.3d), the larger equilibrium is not monotonic in k .

Maximum Equilibrium: Figure 1.3 suggests that the equilibrium service rate has a finite upper bound corresponding to the μ_{\max}^* defined in (1.5), as we vary k . Any larger value of k results in equilibria being close to μ_{\max}^* , which is smaller than the upper bound of 5.33 in Lemma 3 (Figure 1.3a-1.3c). We expect to see such behavior as we vary other system parameters, namely λ and N . This suggests that there will always be a maximum possible equilibrium service rate, and there may be a multitude of parameters under which the servers will work close to their maximum.

By virtue of the aforementioned complicated equilibrium behaviors, exact analysis for a general $M/M/N/k$ system, to the extent that it is tractable, appears unlikely to produce any useful insights. Thus, we focus on exact analysis for specialized systems (Section 1.4) and asymptotic analysis for general systems (Section 1.5).

1.4 Exact Analysis for Specialized Systems

The specialized finite-buffer systems we focus on are the loss system (Section 1.4.1) and the single-server system (Section 1.4.2), two systems that are well-studied in the non-strategic setting (Kelly (1991), Cohen (2012)). These allow us to study the effects of competition and the effects of the buffer size separately, because the loss system has more than one server but no buffer ($k = N > 1$) whereas the single-server system has positive buffer ($k > N = 1$).

1.4.1 The $M/M/N/N$ Loss System

The $M/M/N/N$ loss system is an extreme case of an $M/M/N/k$ system, in which $k = N$; that is, there is no buffer to hold waiting customers. An arriving customer immediately goes into service if there is an idle server, but is lost if all servers are busy.

Theorem 2 (Characterizing Equilibria). *In an $M/M/N/N$ loss system, $\mu^* > 0$ is an*

equilibrium service rate if and only if it satisfies the FOC (1.6), which can be written as

$$c'(\mu) = \left(p(1 - I(\mu, \mu; \lambda, N)) + \frac{v}{\mu} I(\mu, \mu; \lambda, N) \right) (1 - I(\mu, \mu; \lambda, N)). \quad (1.10)$$

Furthermore, a solution to (1.10) exists if $c'(0) < p + v/\lambda$.

Remark 2. The sufficient condition (1.7) of Theorem 1 is the least restrictive for the loss system, wherein it simplifies to $c'(0) < p + v/\lambda$, consistent with Theorem 2. Any necessary condition must be close to this sufficient condition, as explained later by Remark 3.

Theorem 2 implies that every solution $\mu^* > 0$ to the FOC is an equilibrium, that is, μ^* is a global maximizer of $U(\mu_1, \mu^*)$. As revealed in its proof, this is because, in a loss system, the utility function $U(\mu_1, \mu)$ is strictly concave in μ_1 , for all $\mu_1 > 0$ and $\mu > 0$.

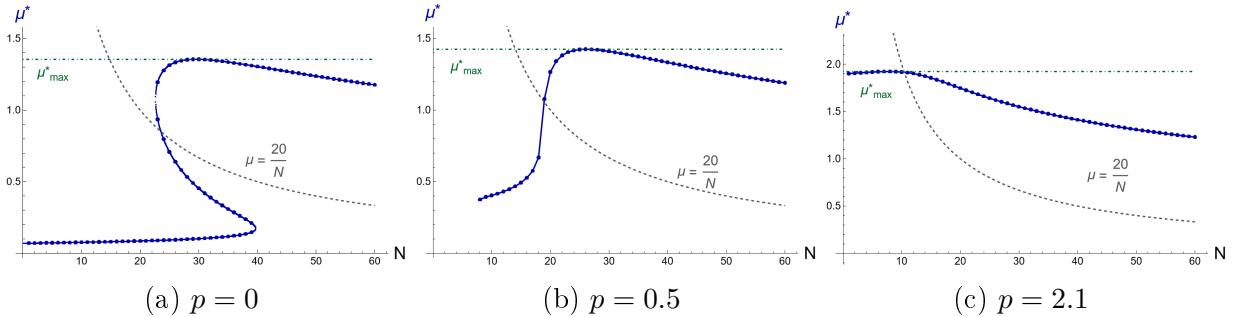


Figure 1.4: Server equilibria as a function of the staffing level N , when $\lambda = 20$, $v = 10$, and $c(\mu) = \mu^{1.5} + 0.1\mu$, for three values of p . The blue dots represent equilibria for integer values of N . μ_{\max}^* is given by Proposition 3.

Figure 1.4 shows all equilibria for various values of N , for three values of p . Theorem 2 provides a condition for the existence of equilibrium in the loss system. However, that equilibrium may be neither unique (Figure 1.4a) nor monotonic (Figures 1.4a and 1.4b). Fortunately, we can rigorously establish that larger the equilibrium, larger the utility, a property that was demonstrated numerically in Figure 1.3 for a general system discussed in Section 1.3.2.

Proposition 2 (Equilibrium Selection). *In an $M/M/N/N$ loss system, given λ , N , p and v , if μ_1^* , μ_2^* are two distinct equilibria with $\mu_1^* > \mu_2^*$, then, $U(\mu_1^*, \mu_1^*) > U(\mu_2^*, \mu_2^*)$.*

The maximum equilibrium service rates μ_{\max}^* in Figure 1.4a and 1.4b are much smaller than 4.5, which is obtained using the upper bound given in Lemma 3. This is also observed numerically in Section 1.3.2. We can provide a rigorous characterization of μ_{\max}^* .

Proposition 3 (Maximum Equilibrium and Idle Time). *In an $M/M/N/N$ loss system, under Assumption 1, $\mu_{\max}^*(p, v)$ is strictly increasing in p . Let $p^\ddagger(v)$ be the unique solution for $p > c'(0)$ to $c'(\frac{v}{2p}) = p$. If $0 \leq p < p^\ddagger(v)$, then*

$$\mu_{\max}^* c'(\mu_{\max}^*) = \frac{v^2}{4(v - p\mu_{\max}^*)} \quad \text{and} \quad I(\mu_{\max}^*, \mu_{\max}^*) = \frac{v - 2p\mu_{\max}^*}{2v - 2p\mu_{\max}^*}.$$

If $p \geq p^\ddagger(v)$, then

$$c'(\mu_{\max}^*) = p \quad \text{and} \quad I(\mu_{\max}^*, \mu_{\max}^*) = 0.$$

Proposition 3 implies that a larger payment can contribute to a larger maximum equilibrium service rate. Moreover, for a fixed small p , when servers work at the maximum equilibrium service rate, they also enjoy a significant amount of idle time. On the other hand, for a fixed sufficient large p , servers could be incentivized to speed up with no idleness.

Example 1. For the family of cost functions specified by (1.9), when $q = 1$, we can compute a closed-form expression for μ_{\max}^* , from Proposition 3. If $0 \leq p < p^\ddagger(v) = c_E + d_E$, then

$$\mu_{\max}^* c'(\mu_{\max}^*) = \frac{v^2}{4(v - p\mu_{\max}^*)} \quad \Rightarrow \quad \mu_{\max}^* = \frac{v/2(c_E + d_E)}{1 + \sqrt{1 - p/(c_E + d_E)}}.$$

It is easy to see that μ_{\max}^* is strictly increasing in p , as predicted by Proposition 3.

Figure 1.4 also serves to illustrate Proposition 3. As N increases, the equilibrium service rate in Figure 1.4b ($p = 0.5$) first increases, attains a maximum ($\mu_{\max}^* \approx 1.424$), and then

decreases. The same is true in Figure 1.4a ($p = 0$) for the largest equilibrium service rate ($\mu_{\max}^* \approx 1.354$). It is clear that the value of μ_{\max}^* when $p = 0.5$ is greater than that when $p = 0$, which is consistent with Proposition 3.

Additionally, as mentioned in the paragraph before Proposition 2, equilibria are not always unique (Figure 1.4a). Comparing Figure 1.4a with Figure 1.4b, we observe that a sufficiently large payment can help to bring about uniqueness of equilibrium in the system. This is because servers who work slowly at the smaller equilibrium in Figure 1.4a are motivated to speed up under payment incentive, resulting in uniqueness of equilibrium for all N in Figure 1.4b.

1.4.2 The $M/M/1/k$ Single-Server System

The $M/M/1/k$ ($k \in \mathbb{Z}_+ \cup \{\infty\}$) system allows us to investigate the effects of the buffer size in isolation from the effects of competition. The reasons equilibria may not exist in the $M/M/1/k$ system, and their characteristics when they do, are building blocks that can help us understand the more complex equilibrium behavior exhibited by the $M/M/2/k$ system (in Section 1.3.2).

Since the general model and associated notation introduced in Section 1.2 implicitly assumed that there were at least two servers in the system, we first redefine the key quantities of interest for an $M/M/1/k$ system. The server's idle time, obtained by substituting $N = 1$ in (1.1), is given by

$$I(\mu; \lambda, k) = \left(1 + \sum_{i=1}^k \left(\frac{\lambda}{\mu} \right)^i \right)^{-1} = \begin{cases} \frac{1 - (\lambda/\mu)}{1 - (\lambda/\mu)^{k+1}}, & \lambda \neq \mu, \\ \frac{1}{k+1}, & \lambda = \mu. \end{cases} \quad (1.11)$$

Since there is only one server, the server's utility function, from (1.2), is

$$U(\mu; \lambda, k, p, v) = (v - p\mu)I(\mu; \lambda, k) + p\mu - c(\mu). \quad (1.12)$$

An equilibrium when there is only one server is a service rate that maximizes this utility function. Therefore, the equivalent of (1.3), for an $M/M/1/k$ system, is $\mu^*(\lambda, k, p, v) \in \arg \max_{\mu > 0} U(\mu; \lambda, k, p, v)$, and the FOC (1.6) evaluates to

$$(c'(\mu) - p)\mu \sum_{i=0}^k \left(\frac{\lambda}{\mu}\right)^i + p\mu = (v - p\mu) \frac{\sum_{i=0}^k i \left(\frac{\lambda}{\mu}\right)^i}{\sum_{i=0}^k \left(\frac{\lambda}{\mu}\right)^i}. \quad (1.13)$$

We continue using the term “equilibrium” for consistency.

The discussion concerning individual rationality from Section 1.2 is also applicable to $M/M/1/k$ system, that is, it can be shown that any equilibrium $\mu^* > 0$ must satisfy the individual rationality condition, $U(\mu^*) \geq 0$.

Unlike in the loss system, $U(\mu)$ is not necessarily concave in μ in the single-server system. Therefore, not all solutions to the FOC (1.13) need be local maxima, let alone equilibria. Still, a useful consequence of there being only one server is that an equilibrium $\mu^* > 0$ is guaranteed to exist as long as $U(\mu) \geq 0$ for some $\mu > 0$. (This is not true in the $M/M/N/k$ system where local maxima with nonnegative utilities may not be equilibria, as in Figure 1.2.)

Lemma 4. *If $U(\mu; \lambda, k, p, v) \geq 0$ for some $\mu > 0$, then an equilibrium exists.*

This observation, along with the monotonicity of $I(\mu; \lambda, k)$ (namely, $I(\mu; \lambda, k)$ is strictly decreasing in k) suggests that equilibrium existence in an $M/M/1/k$ system is a property that is monotonic in k for small enough p .

Proposition 4 (Monotonicity of Equilibrium Existence). *When $p \leq c'(0)$, if there exists an equilibrium $\mu^*(k) > 0$ for an $M/M/1/k$ system for some $k \in \mathbb{Z}_+ \cup \{\infty\}$, then, there exists an equilibrium $\mu^*(k') > 0$ for any $M/M/1/k'$ system with $k' \leq k$.*

Proposition 4 establishes a threshold structure for the existence of equilibria in the single-server system, which simplifies the equilibrium analysis. As our next result shows, an equilibrium exists for all k only under certain conditions.

Theorem 3 (Existence of Equilibrium). *In an $M/M/1/k$ system, given $p \geq 0$ and $v > 0$, the following hold:*

- (a) *If $p \leq c'(0) - v/\lambda$, $\mu^*(k) > 0$ does not exist for any k ;*
- (b) *If $c'(0) - v/\lambda < p \leq c'(0)$, $\mu^*(k) > 0$ exists for some k . However, $\mu^*(k) > 0$ exists for all k if and only if $v > \lambda c'(\lambda)$ and the unique solution for μ to $\mu^2 c'(\mu) = v\lambda$ also satisfies $\mu c'(\mu) + c(\mu) \leq v + p\lambda$;*
- (c) *If $p > c'(0)$, $\mu^*(k) > 0$ exists for all k .*

Theorem 3 suggests that an equilibrium is more likely to exist for all k when either the piecerate payment or servers' valuation for idleness is large enough. In particular, when servers' valuation for idleness is large enough (i.e., $v > \lambda c'(0)$), there always exists an equilibrium for some k regardless of the piecerate payment. When the piecerate payment is large enough (i.e., $p > c'(0)$), there always exists an equilibrium for all k regardless of how servers value their idle time.

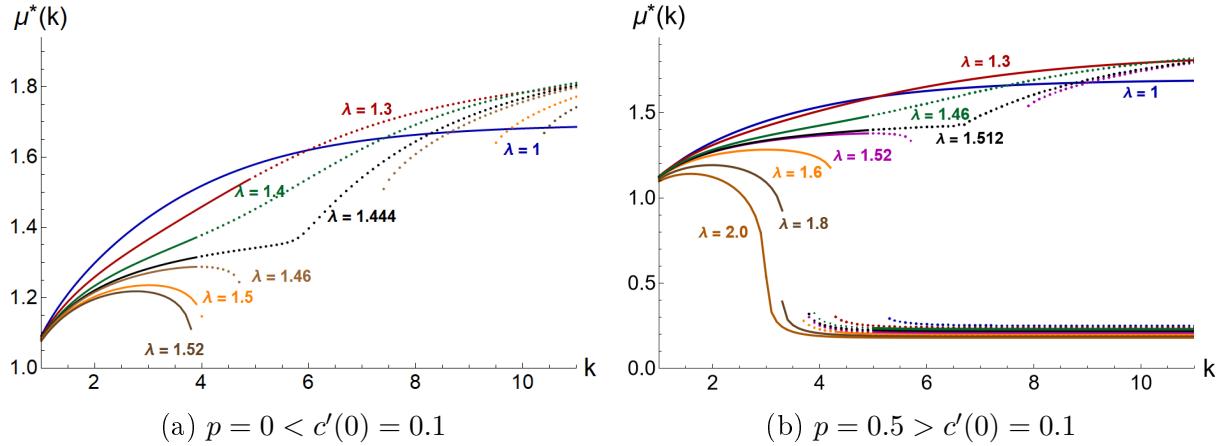


Figure 1.5: Equilibria (solid) and local maxima that are not equilibria (dotted) when $v = 10$ and $c(\mu) = \mu^2 + 0.1\mu$, for two values of p . The $\mu^*(k)$ values in the lower portion of panel (b) are separated for illustration.

Figure 1.5 shows, for different values of λ , server equilibria as a function of k . Under a fixed effort-cost function, we consider two different values of p ; namely $p < c'(0)$ in Figure 1.5a

and $p > c'(0)$ in Figure 1.5b. In contrast to the $M/M/2/k$ system (Figure 1.3), in which competitive effects are present, multiple equilibria never occur in the single-server system in Figure 1.5. Still, equilibria may be non-monotonic, and their existence can be discontinuous (as in Figure 1.3).

Figure 1.5a shows that for larger values of λ , equilibria cease to exist even when local maxima exist, and that local maxima cease to exist when their behavior becomes non-monotonic (at $\lambda = 1.444$). Equilibria cease to exist when the server's utility at the local maximum becomes negative, causing her to “rebel” against having a larger buffer. The non-monotonicity occurs because the payment is not enough to compensate for the diminishing returns (in both throughput and idle time) and overcome the increasing losses in effort-cost from working faster.

Figure 1.5b is similar to Figure 1.5a when the y -axis is restricted to values above 1, except that the larger payment “tempers” the non-monotonic behavior. For instance, when $\lambda = 1.46$, local maxima are discontinuous and non-monotonic in Figure 1.5a, but continuous and monotonic in Figure 1.5b.

In comparison to Figure 1.5a, the difference in Figure 1.5b is the possibility of a discontinuous drop in equilibria as k increases.⁴ This is because the payment is large enough to induce a second, smaller local maximum (which quickly converges to $(c')^{-1}(p) = 0.2$) with positive utility that “softens” the aforementioned rebellion by getting the server to switch to it before her utility from the larger local maximum fully vanishes. As a consequence, an equilibrium exists for all k , regardless of λ , as predicted by Theorem 3 (c).

4. The discontinuous drop becomes continuous as λ increases beyond 1.8 in Figure 1.5b. This can be explained as follows. For smaller λ , there exist two local maxima for some values of k , which are “disconnected” from each other. This leaves the server with no choice but to “jump” discontinuously to the smaller local maximum if and when it begins to dominate the larger local maximum. For larger λ (beyond 1.8 in Figure 1.5b), there exists a unique, continuous local maximum for all k , which is also the equilibrium.

1.5 Asymptotic Analysis

Developing criteria for when an equilibrium will exist, much less characterizing its value, is challenging, as demonstrated in Section 1.3. In this section, we develop a many-server approximation to the FOC (1.6), which we use to predict equilibrium behavior. Section 1.5.1 evaluates a many-server limit of the idle time, which allows us to write a limiting version of (1.6) and to establish concavity of the utility function. Section 1.5.2 analyzes the aforementioned limiting FOC to determine when solutions exist and to understand their properties. Section 1.5.3 provides convergence results and numerical figures supporting using the limiting FOC's solutions to study equilibrium behavior.

1.5.1 Limiting FOC

We consider a sequence of systems indexed by the arrival rate λ , and let λ become large. Our convention when we wish to refer to any process or quantity associated with a system having arrival rate λ is to superscript the appropriate symbol by λ . Specifically,

- N^λ denotes the staffing level and $k^\lambda \geq N^\lambda$ the system size;
- $\mu^{*,\lambda}$ denotes an equilibrium service rate.

Our approach is to first study a limiting version of the FOC (1.6), when μ is fixed (a reasonable approximation for large enough λ). The insights obtained from this analysis guide our subsequent steps towards characterizing prelimit equilibria.

Developing a limiting version of the FOC (1.6) requires understanding how its right-hand side behaves as a function of λ .

Lemma 5 (Limiting Idle Time and Derivative of Idle Time). *Fix $\mu > 0$. If $N^\lambda = \frac{1}{a}\lambda + o(\lambda)$ for $a > 0$, then*

$$\lim_{\lambda \rightarrow \infty} I^\lambda(\mu, \mu) = \left[1 - \frac{a}{\mu}\right]^+ \quad \text{and} \quad \lim_{\lambda \rightarrow \infty} \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} = \frac{a[\mu - a]^+}{\mu^3}.$$

Otherwise, if $N^\lambda = f(\lambda) + o(f(\lambda))$ for $f(\lambda) \in o(\lambda) \cap \omega(1)$ or $f(\lambda) \in \omega(\lambda)$,⁵ then the first limit above is either 0 or 1, and the second limit above is 0.

Lemma 5 motivates⁶ focusing on the linear relationship

$$N^\lambda = \frac{1}{a}\lambda + o(\lambda). \quad (1.14)$$

Then, the limiting idle time in Lemma 5 (except for when $a = \mu$) is nonzero if and only if the system is underloaded ($\lambda < N^\lambda\mu$) for all large enough λ . The rate at which the system size k^λ grows (which is at least linear because $k^\lambda \geq N^\lambda$) does not matter, and can only be a second-order determinant of system performance. This is not concerning, given that the principle of conservation of mass, under the lens of a first-order analysis of a finite-buffer queue that ignores variability, dictates that the percentage of time that a server is idle must be $\left(\frac{N^\lambda\mu - \lambda}{N^\lambda\mu}\right)^+ = \left(1 - \frac{\lambda}{N^\lambda\mu}\right)^+$, which is not affected by the system size k^λ .

Under the linear staffing in (1.14), for fixed $\mu > 0$, taking the limit of the utility function $U^\lambda(\mu, \mu; \lambda, k^\lambda, N^\lambda, p, v)$ as $\lambda \rightarrow \infty$ yields

$$U^\infty(\mu, \mu; a, p, v) := p\mu + (v - p\mu) \left[1 - \frac{a}{\mu}\right]^+ - c(\mu). \quad (1.15)$$

Taking the limit in the FOC (1.6) as $\lambda \rightarrow \infty$ yields $c'(\mu) = p \left(1 - \left[1 - \frac{a}{\mu}\right]^+\right) + (v -$

5. We use the o and ω notations to denote the limiting behavior of functions. Formally, for any two real-valued functions $f(x)$, $g(x)$ that take nonzero values for sufficiently large x , we say that $f(x) \in o(g(x))$ (equivalently, $g(x) \in \omega(f(x))$) if $\lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} = 0$. In other words, f is dominated by g asymptotically (equivalently, g dominates f asymptotically).

6. When $N^\lambda = f(\lambda) + o(f(\lambda))$, the limiting FOC evaluates to $c'(\mu) = p$, resulting in the only possible solution to the limiting FOC being $\mu = (c')^{-1}(p)$, which is not very interesting. When $N^\lambda = f(\lambda) + \omega(\lambda)$, the limiting FOC evaluates to $c'(\mu) = 0$, resulting in the only solution to the limiting FOC being $\mu = 0$, which precludes having a nonzero equilibrium in the limit.

$p\mu) \frac{a[\mu-a]^+}{\mu^3}$, which, after algebra, is equivalent to

$$c'(\mu) = p \left(1 - \left[1 - \frac{a^2}{\mu^2} \right]^+ \right) + v \frac{a}{\mu^2} \left[1 - \frac{a}{\mu} \right]^+. \quad (1.16)$$

Therefore, we expect that a solution to the FOC (1.6) will, for large enough λ , be very close to a solution to (1.16).

Given that the growth rate of the system size k^λ does not affect (1.16), we expect the exact results for the loss system in Section 1.4.1 to have parallels in the many-server asymptotic regime. Recall, from Theorem 2, that in a loss system, finding an equilibrium service rate is equivalent to solving the corresponding FOC. The same holds true here, because the utility function (1.4), under the linear staffing rule (1.14), is strictly concave, for all large enough λ .

Lemma 6 (Concavity of Utility). *Under the staffing rule (1.14), except when $\mu = a$ and $\lim_{\lambda \rightarrow \infty} N^\lambda - \frac{\lambda}{a}$ is finite and strictly less than 1, for all large enough λ , the second partial derivative of $U^\lambda(\mu_1, \mu)$ (see (1.4)) with respect to μ_1 is strictly negative for all $\mu_1 > 0$ and $\mu > 0$.*

Proposition 5 (Existence of Prelimit Equilibria). *Under the staffing rule (1.14), except when $\mu = a$ and $\lim_{\lambda \rightarrow \infty} N^\lambda - \frac{\lambda}{a}$ is finite and strictly less than 1, for all large enough λ , $\mu^{*,\lambda} > 0$ is an equilibrium if and only if it satisfies the FOC (1.6).*

Lemma 6 and Proposition 5 are remarkable, given the behavior observed in Figure 1.2 in Section 1.3 that prevented us from being able to establish more generally that a (prelimit) candidate equilibrium is an equilibrium. Consequently, in the remainder of this section, our analysis deals directly with equilibria, without having to distinguish them from candidate equilibria (except when dealing with critically loaded equilibria, where we must exclude a small and insignificant subset of staffing rules for which $\mu = a$ and $\lim_{\lambda \rightarrow \infty} N^\lambda - \frac{\lambda}{a}$ is finite and strictly less than 1).⁷

7. If $\mu = a$ and $\lim_{\lambda \rightarrow \infty} N^\lambda - \frac{\lambda}{a} =: x \in (-\infty, 1)$, then, there exist linear staffing rules for which

1.5.2 Properties of Limiting Equilibria

Our first set of results shed light on the existence and characteristics of limiting equilibria.

We begin with an analysis of the different types of limiting equilibria that could emerge as solutions to the limiting FOC (1.16). Let the tuple (n_u, n_c, n_o) represent the number of underloaded, critically loaded, and overloaded limiting equilibria.

Definition 3. *Under linear staffing (1.14), letting $\lambda \rightarrow \infty$ in Definition 1, a limiting equilibrium service rate μ is underloaded if $\mu > a$, critically loaded if $\mu = a$, and overloaded if $\mu < a$.*

The right-hand side of the limiting FOC (1.16) incorporates a payment term and an idleness term. When $\mu \leq a$, the payment term becomes p , and the idleness term becomes zero, implying that a server in an overloaded or critically loaded limiting equilibrium has no idle time to enjoy. As a result, an overloaded or critically loaded limiting equilibrium is uniquely determined by $c'(\mu) = p$. On the other hand, when $\mu > a$, the payment term becomes $p\frac{a^2}{\mu^2}$, which is decreasing in the equilibrium service rate, and the idleness term becomes $v\frac{a}{\mu^2}(1 - \frac{a}{\mu})$, which is not a monotonic function of μ , therefore possibly admitting multiple underloaded limiting equilibria. We build on this intuition to precisely characterize (n_u, n_c, n_o) in the next result.

Theorem 4 (Existence and Characterization of Limiting Equilibria). *Fix $v > 0$. Let $p^\dagger(v)$ be the unique solution for $p > c'(0)$ to $p(c')^{-1}(p) + \frac{1}{2}((c')^{-1}(p))^2 c''((c')^{-1}(p)) = \frac{v}{2}$. Under Assumption 1, the following hold:*

- (a) [Areas 2,3 in Figure 1.6] If $p \leq c'(0)$, then $n_c = n_o = 0$; and there exists a unique $\bar{a}(p, v) > 0$ such that $n_u = 0, 1$, or 2 according to whether $a > \bar{a}(p, v)$, $a = \bar{a}(p, v)$, or

$\lim_{\lambda \rightarrow \infty} \left| \frac{\partial I^\lambda(\mu_1, a)}{\partial \mu_1} \right| = \infty$ (which would, depending on the relationship between the parameters a , p , and v , result in $\lim_{\lambda \rightarrow \infty} \frac{\partial U^\lambda(\mu_1, a)}{\partial \mu_1} = \infty$) for exactly one $\mu_1 > 0$ (namely, $\mu_1 = a(1 - x)$); for the technical details, see Section A.7.7 in the appendix to this chapter. Moreover, this exclusion is relevant only to Theorem 5B (c), which deals with the possible existence of a critically loaded equilibrium for all large enough λ .

$a < \bar{a}(p, v)$, respectively.

(b) [Areas 1,4,5 in Figure 1.6] If $p > c'(0)$,

(i) $n_o = 1$ if $a > (c')^{-1}(p)$; $n_o = 0$ otherwise;

(ii) $n_c = 1$ if $a = (c')^{-1}(p)$; $n_c = 0$ otherwise;

(iii) If $p < p^\dagger(v)$, there exists a unique $\bar{a}(p, v) > 0$ such that $n_u \geq 1$ if and only if $a \leq \bar{a}(p, v)$. In particular, when $a \leq \bar{a}(p, v)$, $n_u = 2$ if $(c')^{-1}(p) < a < \bar{a}(p, v)$ and $n_u = 1$ otherwise.

(iv) If $p \geq p^\dagger(v)$, $n_u = 1$ if $0 < a < (c')^{-1}(p)$; $n_u = 0$, otherwise.

Remark 3. The sufficient condition for equilibrium existence in Theorem 2 for a loss system, $p > c'(0) - v/\lambda$, becomes $p > c'(0)$ in our asymptotic regime. Theorem 4 shows that this is both necessary and sufficient to guarantee equilibrium existence.

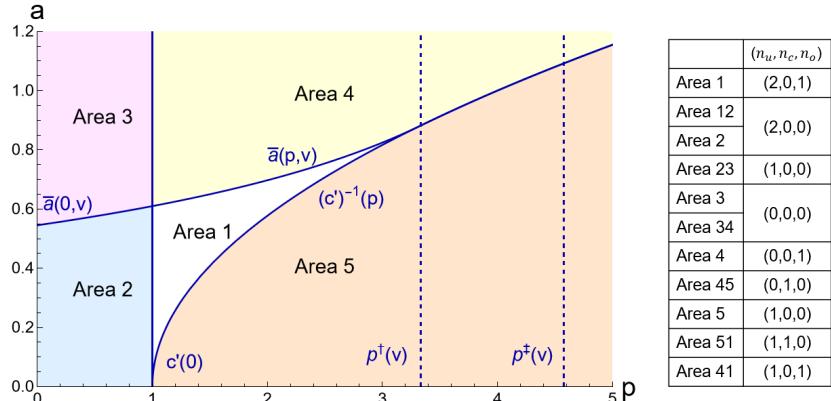


Figure 1.6: Existence of limiting equilibria, when $v = 10$ and $c(\mu) = \mu^3 + \mu$.

Theorem 4 establishes that $c'(0)$ is the infimum piece-rate payment that can overcome the non-existence of limiting equilibria (Area 3 in Figure 1.6) when the staffing is too small (i.e., $a > \bar{a}(p, v)$, recalling (1.14)), by incentivizing the servers to give up all their idle time. Recalling Remark 1, $c'(0)$ is interpreted as a measure of a server's inherent resistance to moving from an idle state to a working state, and $p > c'(0)$ would, according to Theorem 1,

ensure that the piece-rate payment is large enough to overcome this resistance and ensure equilibrium existence. This implies that the system manager *must* use payment to ensure the existence of equilibrium for *all* staffing levels. A more stringent condition, $p \geq p^\dagger(v)$ (visually, $p^\dagger(v)$ is the point where Areas 1, 4, and 5 meet in Figure 1.6), ensures the existence of a *unique* equilibrium for *all* staffing levels.

The equilibria in Figure 1.6 are all underloaded when the staffing level is larger ($a < \bar{a}(p, v)$) and payment is either small enough ($p \leq c'(0)$) or large enough ($p > c'(a)$), as shown in Areas 2 and 5, respectively, and all overloaded when the staffing level is smaller ($a > \bar{a}(p, v)$) and payment exceeds $c'(0)$, as shown in Area 4. This means that the system manager must consider the impact of staffing jointly with payment to guarantee that the servers will settle into either an underloaded equilibrium (Areas 2 and 5) or an overloaded equilibrium (Area 4). When it comes to Area 1, where both underloaded and overloaded equilibria exist, the system manager can either avoid it altogether by carefully controlling the payment, or, as is true in the loss system (recall Proposition 2), rely on the propensity of the servers to settle into the fastest equilibrium (which is guaranteed to be underloaded). A critically loaded equilibrium emerges only along boundaries between regions in Figure 1.6, namely in Areas 45 and 51, and is realized when an equilibrium transitions continuously from being overloaded in one region to being underloaded in another region and vice versa. Fortunately, when multiple equilibria exist, the fastest equilibrium induces the highest utility, as observed in the loss system (Proposition 2).

Proposition 6 (Equilibrium Selection). *If μ_1^* , μ_2^* are two distinct limiting equilibria with $\mu_1^* > \mu_2^*$, then, $U^\infty(\mu_1^*, \mu_1^*) > U^\infty(\mu_2^*, \mu_2^*)$.*

The behavior of limiting equilibria is complex and depends on the (a, p) -parameter space (defined in Figure 1.6) as well as their type (underloaded versus overloaded). Our next result, for which we provide intuition and numerical illustration immediately after, characterizes the monotonicity (or lack thereof) of limiting equilibria, when they exist (recall Theorem 4).

Proposition 7 ((Non)Monotonicity of Limiting Equilibria). *Fix $v > 0$. Let $p^\dagger(v)$ and $p^\ddagger(v)$ be as defined in Theorem 4 and Proposition 3, respectively, which satisfy $p^\ddagger(v) > p^\dagger(v) > c'(0)$. Under Assumption 1, the following hold:*

- (a) [Overloaded: Areas 1,4,41 in Figure 1.6] Let $0 < \mu_o^*(a, p; v) < a$ be the overloaded equilibrium, when it exists.
 - (i) $\mu_o^*(a, p; v)$ is strictly increasing in p .
 - (ii) $\mu_o^*(a, p; v)$ does not depend on a .
- (b) [Underloaded: Areas 1,12,2,23,41,5,51 in Figure 1.6] Let $\mu_1^*(a, p; v) > a$ be the (larger) underloaded equilibrium, and $\mu_2^*(a, p; v) > a$ be the smaller underloaded equilibrium, when they exist.
 - (i) $\mu_1^*(a, p; v)$ is strictly increasing in p , and $\mu_2^*(a, p; v)$ is strictly decreasing in p .
 - (ii) $\mu_2^*(a, p; v)$ is strictly increasing in a .
 - (iii) (1) If $0 \leq p < p^\dagger(v)$, $\mu_1^*(a, p; v)$ is strictly increasing in a for $a \in (0, a^\dagger(p; v))$, and strictly decreasing in a for $a > a^\dagger(p; v)$, where $a^\dagger(p; v)$ is the unique solution for $a \in (0, \frac{v}{2p})$ to $c'(\frac{v}{p+\frac{v}{2a}}) = \frac{av}{2} \left(\frac{p}{v} + \frac{1}{2a} \right)^2$.

(2) If $p \geq p^\dagger(v)$, $\mu_1^*(a, p; v)$ is strictly increasing in a .

The behavior of overloaded equilibria is simpler than that of underloaded equilibria. This is because, in overloaded equilibria, all servers are busy 100% of the time, i.e., the fraction of time they are idle is zero in the limit, thereby isolating each server and eliminating competitive effects. Mathematically, the limiting FOC (1.16) simplifies to $p = c'(\mu)$ when $\mu < a$, which implies that an overloaded equilibrium exists if and only if $c'(0) < p < c'(a)$, and is unique, given by $\mu^* = (c')^{-1}(p)$. This is reminiscent of the behavior observed numerically in the single-server system, where, as λ grows, the overloaded equilibrium is eventually determined by $p = c'(\mu)$ for all k ; see Figure 1.5b. As a result, it is not affected

by changes in the staffing (Proposition 7(a)(ii), illustrated in Figures 1.7b-1.7d), and is increasing in p (Proposition 7(a)(i), illustrated in Figure 1.8).

Underloaded equilibria may or may not be monotonically increasing in response to changes in the staffing level (Proposition 7(b)(ii)–(iii)), which is reminiscent of the behavior observed numerically in the loss system; see Figure 1.4. Loosely speaking, when the payment is small, non-monotonicity in the larger underloaded equilibrium occurs when the increase in effort overwhelms the increase in the other two components of the utility (recalling that c is convex increasing), an effect not present for the smaller underloaded equilibrium (Figures 1.7a-1.7b). A larger payment ($p \geq p^\dagger(v)$) can not only help temper this effect, but also ensure equilibrium uniqueness (Figure 1.7c). Eventually, a large enough payment ($p \geq p^\ddagger(v)$) ensures monotonicity by encouraging the servers to keep increasing their service rate until all their idle time is lost and the equilibrium becomes overloaded (Figure 1.7d).

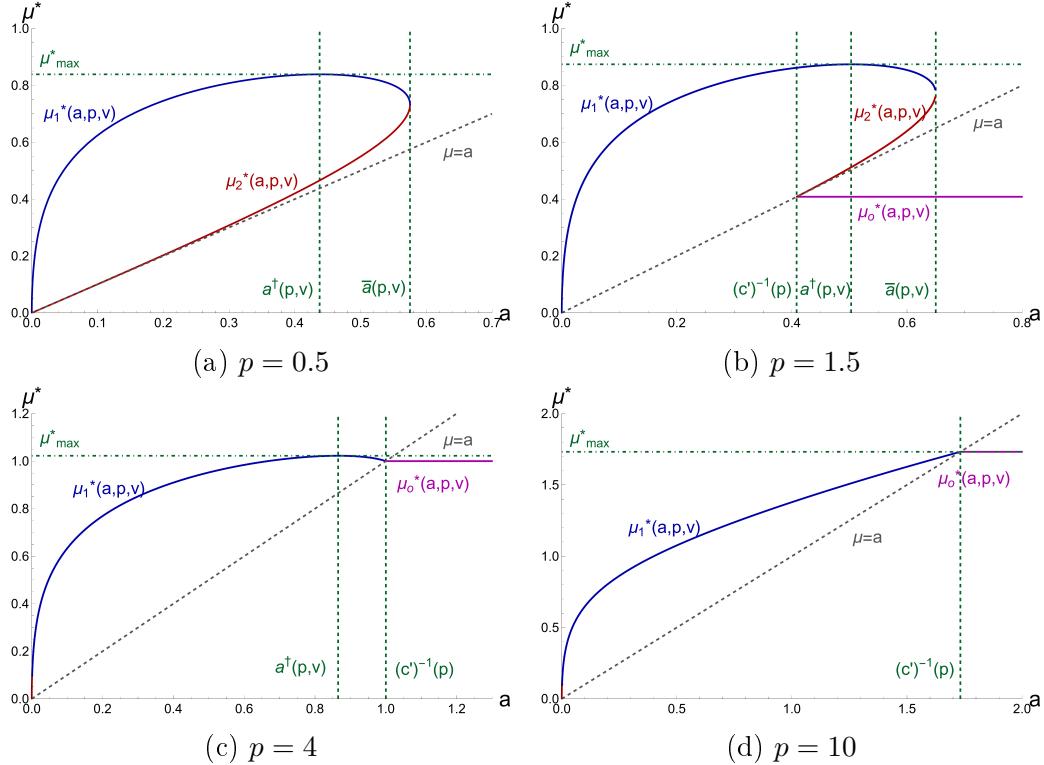


Figure 1.7: Limiting equilibria as a function of the staffing parameter, when $v = 10$ and $c(\mu) = \mu^3 + \mu$. $p^\dagger(v) = 3.33$ and $p^\ddagger(v) = 4.58$.

When the staffing parameter is fixed, a higher payment motivates the servers who choose the faster underloaded equilibrium to work faster; however, surprisingly, the servers that choose the slower underloaded equilibrium (when it exists) work slower when p is larger (Proposition 7(b)(i), which is illustrated in Figure 1.8). This is because the effect of the payment on the equilibrium service rate is nonlinear, meaning that the marginal change in busy time and idleness may differ significantly at different service rates.

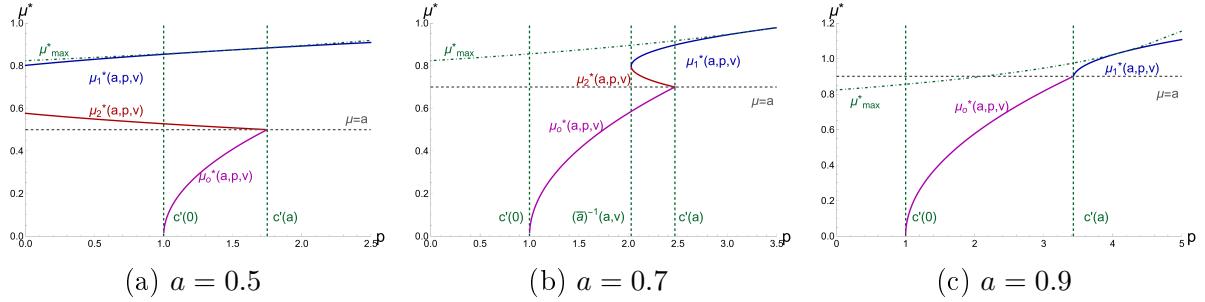


Figure 1.8: Limiting equilibria as a function of payment, when $v = 10$ and $c(\mu) = \mu^3 + \mu$.
 $\bar{a}(0, v) = 0.545$ and $\bar{a}(p^\dagger(v), v) = 0.882$.

Finally, from Proposition 7(a)(iii)(1), the maximum limiting equilibrium,

$$\mu_{\max}^*(p, v) := \sup \{\mu^*(a, p, v) : (1.16) \text{ holds for some } a\}, \quad (1.17)$$

can be characterized exactly. (The definition in (1.17) is analogous to that in (1.5) for finite systems.)

Proposition 8 (Maximum Limiting Equilibrium). *Under Assumption 1,*

$$\mu_{\max}^*(p, v) = \begin{cases} \left(\frac{p}{v} + \frac{1}{2a^\dagger(p; v)} \right)^{-1}, & 0 \leq p < p^\dagger(v) \\ (c')^{-1}(p), & p \geq p^\dagger(v) \end{cases}, \quad (1.18)$$

where $a^\dagger(p; v)$ is the unique solution for $a \in (0, \frac{v}{2p}]$ to $c' \left(\frac{v}{p + \frac{v}{2a}} \right) = \frac{av}{2} \left(\frac{p}{v} + \frac{1}{2a} \right)^2$. Moreover, (1.18) satisfies the equations characterizing μ_{\max}^* exactly for the loss system in Proposition 3.

1.5.3 Prelimit Convergence and Impact of System Size

We expect to leverage limiting equilibria to predict equilibrium behavior in the prelimit system. In particular, we would like to use Theorem 4 (illustrated by Figure 1.6) to predict the existence of solutions to the FOC (1.6) that are underloaded, critically loaded, and overloaded, recalling from Proposition 5 that the solutions to (1.6) are equilibria for all large enough λ .

For each λ , let the tuple $(n_u^\lambda, n_c^\lambda, n_o^\lambda)$ represent the number of underloaded, critically loaded and overloaded equilibria in the system with arrival rate λ .

Theorem 5A (Nonexistence of Prelimit Equilibria). *Under the staffing rule (1.14), except when $\mu = a$ and $\lim_{\lambda \rightarrow \infty} N^\lambda - \frac{\lambda}{a}$ is finite and strictly less than 1, and under Assumption 1, the following holds for all large enough λ .*

- (a) [Areas 2,23,3,5 in Figure 1.6] If $n_c = n_o = 0$ and $p \neq c'(0)$, then $n_o^\lambda = 0$.
- (b) [Areas 3,4,34 in Figure 1.6] If $n_u = n_c = 0$, then $n_u^\lambda = 0$.
- (c) [Everywhere except Areas 45,51 in Figure 1.6] If $n_c = 0$, then $n_c^\lambda = 0$.
- (d) [Areas 23,41 in Figure 1.6] If $n_u = 1$, $p < c'(a)$, and $\lim_{\lambda \rightarrow \infty} N^\lambda - \frac{\lambda}{a} < 0$, then $n_u^\lambda = 0$.

Theorem 5B (Existence and Convergence of Prelimit Equilibria). *Under the staffing rule (1.14), except when $\mu = a$ and $\lim_{\lambda \rightarrow \infty} N^\lambda - \frac{\lambda}{a}$ is finite and strictly less than 1, and under Assumption 1, the following holds for all large enough λ .*

- (a) [Areas 1,4,41 in Figure 1.6] If $n_o = 1$, then $n_o^\lambda \geq 1$.
- (b) (i) [Areas 1,12,2 in Figure 1.6] If $n_u = 2$, then $n_u^\lambda \geq 2$.
- (ii) [Areas 5,51 in Figure 1.6] If $n_u = 1$ and $p \geq c'(a)$,⁸ then $n_u^\lambda \geq 1$.

8. Note that when $a = (c')^{-1}(p)$ (Area 51 in Figure 1.6), there also exists a limiting critically loaded candidate equilibrium, but its prelimit behavior is an open question and is not discussed here.

(iii) [Areas 23,41 in Figure 1.6] If $n_u = 1$ and $p < c'(a)$, then $n_u^\lambda \geq 2$ if $N^\lambda - \frac{\lambda}{a} \geq 0$.

(c) [Area 45 in Figure 1.6] If $n_c = 1$, then $n_u^\lambda + n_c^\lambda + n_o^\lambda \geq 1$.

Furthermore, from continuity, if $\mu^* > 0$ is a limiting equilibrium, then for any $\epsilon > 0$, there exists $\lambda(\epsilon)$ large enough such that for all $\lambda > \lambda(\epsilon)$, there exists a prelimit equilibrium $\mu^{*,\lambda}$ satisfying $|\mu^{*,\lambda} - \mu^*| < \epsilon$.

Remark 4. When $n_u = 1$ and $p < c'(a)$, Theorems 5A (d) and 5B (b)(iii), when taken together, almost exactly characterize the staffing rules for which underloaded equilibria exist in the prelimit. The only staffing rules that they do not address are those for which $\lim_{\lambda \rightarrow \infty} N^\lambda - \frac{\lambda}{a} = 0$ and $N^\lambda - \frac{\lambda}{a} < 0$ for all large enough λ . This gap is vanishingly small compared to that between Theorems 7 (ii)(a) and 7 (ii)(b) in Gopalakrishnan et al. (2016a), which did not address staffing rules for which $N^\lambda - \frac{\lambda}{a} \in (-3, 0)$ for all large enough λ .

Figure 1.9 shows nine possible behaviors that illustrate the convergence in Theorems 5A and 5B, corresponding to each area in Figure 1.6. Observe that, consistent with Theorem 4, equilibria either always exist or disappear after a certain arrival rate, and two equilibria may appear to merge into one as the arrival rate increases.

Not surprisingly, when λ is large, only overloaded prelimit equilibria converge to an overloaded limiting equilibrium and only underloaded prelimit equilibria converge to an underloaded limiting equilibrium. The convergence of prelimit equilibria to a critically loaded limiting equilibrium (Areas 45 and 51), however, is more nuanced, and the load characteristics of prelimit equilibria (underloaded or overloaded) are consequential to the system performance. As stated in Theorem 5B(c) and illustrated in Figure 1.9f, in Area 45, the convergence can be either from above or from below. Theorem 5B(c) stops short of characterizing the conditions that would guarantee either behavior. For Area 51, although there appear to exist some prelimit equilibria associated with the critically loaded limiting equilibrium in Figure 1.9g, Theorem 5B is silent on the matter, because the formal analysis of this case is challenging and remains open.

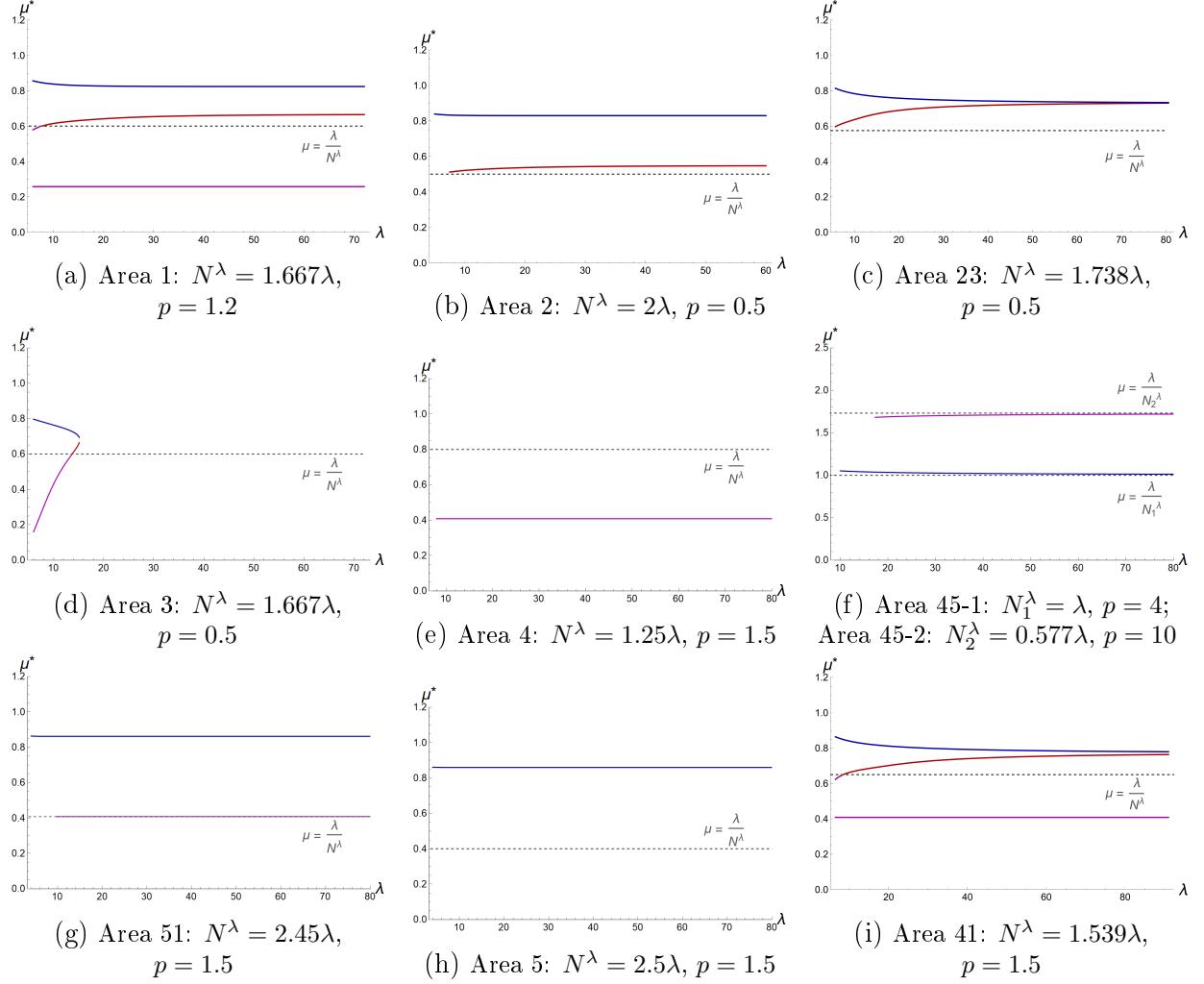


Figure 1.9: Behavior of prelimit equilibria under different payments and scalings of the staffing rule, when $k^\lambda = 3N^\lambda$, $v = 10$, and $c(\mu) = \mu^3 + \mu$.

Remark 5. The predictive power of Theorems 5A and 5B, along with Figure 1.9 (and Figure 1.10 below), suggests that the maximum prelimit equilibrium $\mu_{\max}^{*,\lambda}$ will be close to the maximum limiting equilibrium μ_{\max}^* characterized by Proposition 8, for large enough λ . However, these results do not preclude the existence of “rogue” prelimit equilibria that never get close enough to a limiting equilibrium, no matter how large the arrival rate. Still, Lemma 3 does preclude the existence of unboundedly large prelimit equilibria.

Another observation from Figure 1.9 is that the predicted equilibrium existence and load characteristics (underloaded or overloaded) appear valid even for smaller values of λ , raising

questions regarding the convergence rate, which we discuss next.

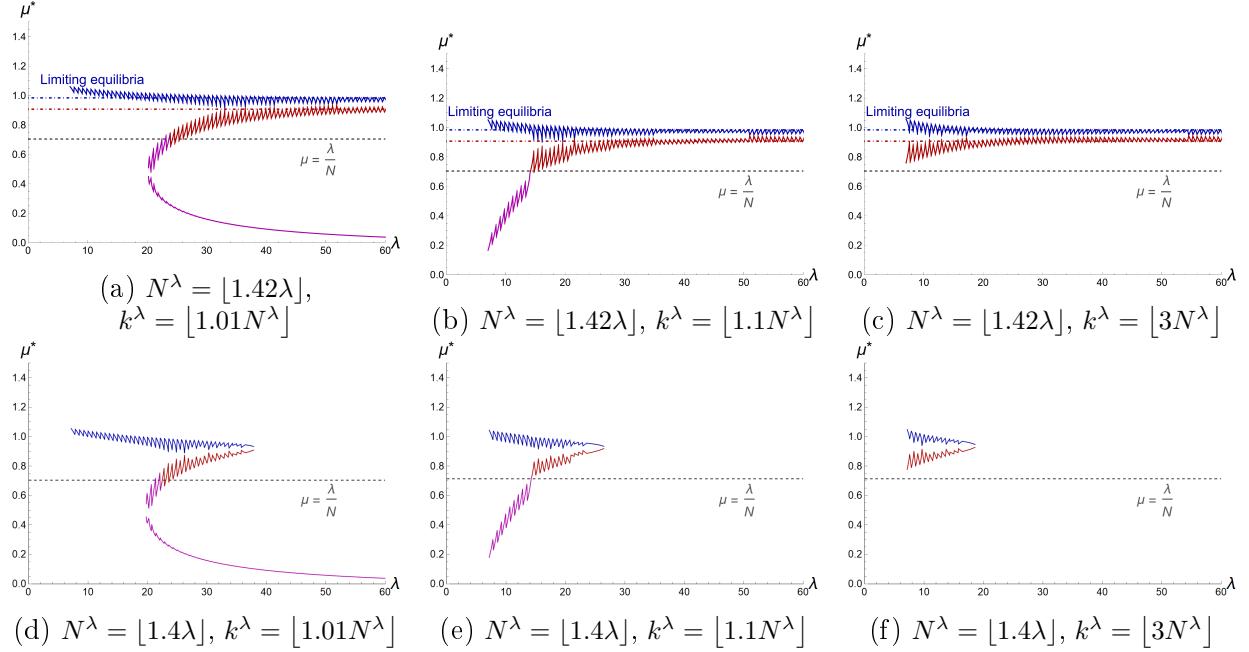


Figure 1.10: Behavior of prelimit equilibria under different scalings of the system size, when $p = 0$, $v = 10$, and $c(\mu) = \mu^2 + 0.1\mu$. The first row corresponds to Theorem 5B(b)(i) and the second to Theorem 5A(b). The horizontal blue and red dot-dashed lines correspond to the limiting equilibria (when they exist). The jaggedness is due to the discrete nature of N^λ and k^λ .

Figure 1.10 suggests that even though the growth of the system size does not influence the statements of Theorems 5A and 5B, it *does* influence the quality of the suggested equilibrium approximation, because it affects the convergence rate. In particular, the convergence rate is faster when the system size grows faster, as can be seen by comparing Figure 1.10c to Figure 1.10a or 1.10b. Then, an open question (beyond the scope of this chapter) is to formally quantify this effect. Such an analysis that captures second-order effects of the system size might also help to better understand the rich behavior of equilibria with respect to k observed numerically in the $M/M/2/k$ (Figure 1.3) and $M/M/1/k$ (Figure 1.5) systems.

Another observation is that Figures 1.10a and 1.10d each show a sequence of overloaded equilibria that appear to decay to 0 in the limit, but are not predicted by our analysis in this section; recall Remark 5, and also note that $\mu = 0$ is forbidden by Lemma 5. Characterizing

and predicting such equilibria is an important open question, since the existence of such equilibria may be consequential to system performance. For instance, recalling that servers prefer to work at the maximum equilibrium service rate, these equilibria are benign in the setting of Figure 1.10a, but in Figure 1.10d, their presence leads to a discontinuous drop from an underloaded to an overloaded equilibrium service rate at $\lambda \approx 40$ in response to a further increase in the arrival rate.

The discontinuous drop in Figure 1.10d is akin to the servers “rebelling” against having a workload high enough such that the payment does not compensate for the little idle time and the cost of effort. In Figures 1.10e and 1.10f, the larger system size intensifies the rebellion, and no equilibria exist for larger λ . This behavior parallels the behavior seen in Figure 1.5 for the single-server system.

1.6 Looking Ahead: Implications for Behavior-Aware Queueing Models

The long-term goal of research on behavior-aware queueing is to gain insight into effective ways to design, manage, and upgrade operations, given different customer and server behaviors. This chapter analyzes a game-theoretic queueing model in which the service rate decision is endogenous, and answers fundamental questions regarding existence and behavior (recall questions (Q1)-(Q3) at the end of Section 1.2). In this section, we provide directions for how to extend our work. Section 1.6.1 uses our analytical results to develop empirically testable hypotheses. Section 1.6.2 discusses system design optimization that accounts for server discretion over work speed. The resulting recommendations will be much different from those based on non-strategic queueing models with exogenous service rates, because the existence of an equilibrium and its behavior depend on design decisions. Section 1.6.3 establishes the link between our model and strategic customer models, thereby forming a basis to investigate the implications of interactions between strategic servers and strategic arrivals.

Finally, the utility function (1.2) may be generalized, which we discuss in Section 1.6.4.

1.6.1 Connecting to Empirical Work

Empirical work has documented that system load affects service rate; however, the direction of the effect is situational. When does an increase in load force the servers to learn how to perform their tasks more efficiently, and so work faster at the same effort level? Alternatively, when does an increase in load cause fatigue, leading to the servers working more slowly? Or, worse, when do servers decide “I am not paid enough for this”, and stop trying? The answer may, in part, depend on the server interactions, because the presence of many servers can potentially cause the servers to work faster (due to social incentives as evidenced in Mas and Moretti (2009) and Bandiera et al. (2010)) or slower (a phenomenon known as social loafing that is introduced in Latane et al. (1979) and comprehensively discussed in Karau and Williams (1993)). Still, even motivated servers fatigue, causing the service speed to have an inverted U-shape (as documented in Kc and Terwiesch (2009), Staats and Gino (2012) and Tan and Netessine (2019)), and may rebel against insufficient payment for overly high workload (Brodsky and Amabile (2018), Burke and Cooper (2008)).

Our model is rich enough to incorporate many of the aforementioned phenomena. The numerical examples in Figures 1.3-1.5 and 1.7 all evidence the equilibrium service rate both increasing and decreasing as load increases and payment remains fixed, and Figure 1.8 shows that increased payment can correspond to increased equilibrium service rate. Here, the term “load” is used loosely, and we associate an increase in λ or k with an increase in load and an increase in N with a decrease in load, as is true in the non-strategic setting. In reality, because μ^* is a function of λ , k , N , and p , we must perform analysis to see how changes in those parameters affect the load $\lambda/N\mu^*(\lambda, k, N, p)$.

Our examples also show servers rebelling when payment does not compensate for the effects of high workload on their utility. Figures 1.5a and 1.10e-1.10f show equilibria ceasing

to exist in response to increasing k and increasing λ , respectively, whereas Figures 1.5b and 1.10d show a discontinuous drop to a slower equilibrium. Figure 1.4a also evidences a discontinuous drop when N is decreased.

Even though analyzing the behavior of $\mu^*(\lambda, k, N, p)$ is difficult, the asymptotic analysis performed in Section 1.5 allows us to predict how changes in the parameters λ , k , N , and p affect $\mu^*(\lambda, k, N, p)$, from which we can develop empirically testable hypotheses. In particular, the convergence results Theorems 5A and 5B support using properties derived in the asymptotic setting (like Theorem 4, Proposition 6 and Proposition 7) to predict prelimit performance.

H1: (From Theorem 4) The servers do not settle into any sustained service rate if the staffing and payment are small.

H2: (From Proposition 6) There can be multiple sustainable service rates. Furthermore, when multiple sustainable service rates are present, the servers settle into the fastest one.

Motivated by H2 above, we use Proposition 7, which is illustrated in Figures 1.7 and 1.8, to develop hypotheses related to the fastest equilibrium service rate. To do that, recall that the relationship in (1.14) implies that $N \approx \lambda/a$ for fixed λ .

H3: (From Figure 1.7 (a)-(c)) When the payment is small, the servers' work speed first increases and then decreases, as the staffing level N shrinks.

H4: (From Figure 1.7 (b)-(d)) When the payment is large, once the staffing level N becomes small enough, the servers' work speed is not affected by changes to the staffing level.

H5: (From Figure 1.8) The servers' work speed increases in the payment.

The hypotheses H3 and H4 together imply that payment can mitigate the slowdown that occurs when staffing levels are reduced, an observation reminiscent of earlier discussion concerning Figures 1.4, 1.5, and 1.7 regarding payment encouraging equilibrium uniqueness

and monotonicity. Still, as observed numerically, insufficient payment can cause the servers to effectively stop working.

H6: (From Figures 1.4a, 1.5, and 1.10d-1.10f) Increasing λ or k , or decreasing N , eventually causes the servers to discontinuously reduce their work speed (possibly to zero), when payment is low.

1.6.2 Optimal Design and Control

There is a long history of using queueing models to inform system design; see, e.g., Stidham (2009). One classical problem is that of admission control, which determines whether and when to turn away arriving customers. The admission decisions do not affect the system load $\lambda/N\mu$, but do affect performance metrics such as the expected wait time, $W(\mu; \lambda, k, N)$, which is increasing in k . The trade-off is between the cost of decreasing k , which forces more customers to be turned away, and decreasing $W(\mu; \lambda, k, N)$. An optimal decision rule d often has a threshold structure, defined through a parameter $T \in \{0, 1, 2, \dots\}$ as

$$d(k) = \begin{cases} \text{admit,} & k \leq T \\ \text{turn away,} & k > T \end{cases}; \quad (1.19)$$

see, for example, Section 11.5.4 in Puterman (2014a) (which details an admission control problem in a single-server setting⁹).

Such admission control problems are not straightforward in queueing models that account for strategic server behavior. First, $\mu^*(\lambda, k, N, p, v)$ may cease to exist when the parameters change. That can likely be handled when there exists a threshold k_0 for which $\mu^*(\lambda, k, N, p, v)$ exists if and only if $k \leq k_0$, as in Figure 1.3d for the $M/M/2/k$ system; however, more care is required when such a threshold does not exist, as in Figure 1.3c, where equilibria may dis-

9. The admission control problem in Section 11.5.4 in Puterman (2014a) has holding costs instead of waiting costs. Little's law can be used to connect the two problems.

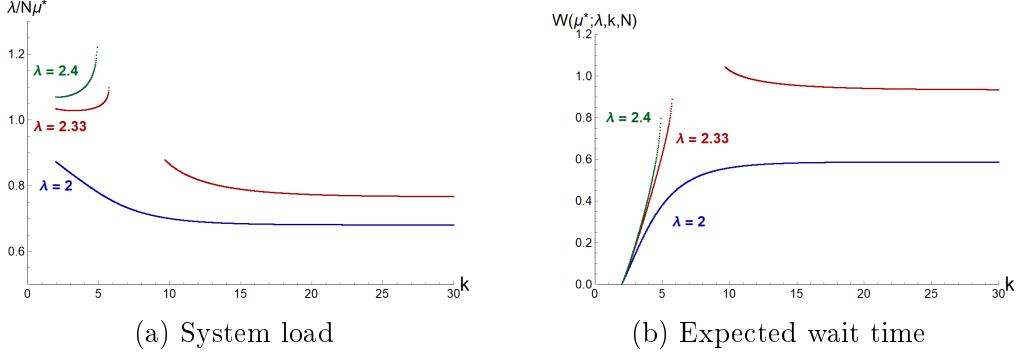


Figure 1.11: Performance metrics as a function of the system size, when $N = 2$, $p = 0$, $v = 1$, and $c(\mu) = \frac{3}{32}\mu^2$, shown only when $\mu^*(\lambda, k, N, p, v)$ exists.

appear and later reappear. Furthermore, in contrast to the traditional, non-strategic setting discussed in the previous paragraph, the effect of k on the system load $\lambda/N\mu^*(\lambda, k, N, p, v)$ is not clear, nor is the effect of k on the expected wait time, $W(\mu^*(\lambda, k, N, p, v); \lambda, k, N)$. In Figure 1.11a (which has the same parameters as Figure 1.3), the system load is decreasing in k for $\lambda = 2$, increasing in k for $\lambda = 2.4$, and non-monotonic and discontinuous in k for $\lambda = 2.33$. In Figure 1.11b, the expected wait time is increasing in k for $\lambda = 2$ and $\lambda = 2.4$, and non-monotonic and discontinuous in k for $\lambda = 2.33$. Surprisingly, the expected wait time can increase both when system load increases and when system load decreases, as can be seen by comparing the curves corresponding to $\lambda = 2.4$ and $\lambda = 2$, which is not possible in the traditional, non-strategic setting. As a consequence, intuition developed in the traditional, non-strategic setting may not carry over.

More broadly, queueing models that endogenize server behavior require revisiting system design problems. The work in this chapter provides a foundation to do so in many-server settings.

1.6.3 Strategic Arrivals and Strategic Servers

The foundational work of Naor (1969) showed that when customers strategically trade off the value of service and the cost of waiting, their equilibrium joining strategy has a threshold

form; that is, an arriving customer joins the system if and only if there are no more than $k^*(\mu)$ customers in the system. That threshold form arises in many models in which customers make strategic joining decisions (see, e.g., Knudsen (1972); Economou and Kanta (2008); Debo et al. (2012); Haviv (2014); Simhon et al. (2016)), and we refer the reader to Hassin and Haviv (2003) and Hassin (2016) for a comprehensive review. The relevance to this chapter is that a threshold type equilibrium joining strategy results in an $M/M/N/k^*(\mu)$ queueing system. Then, to capture the interaction between strategic customers and strategic servers, we define the joint equilibrium (k^*, μ^*) , in which $k^*(\mu^*)$ is the aforementioned threshold that depends on the service rate, and $\mu^* \in \arg \max_{\mu_i > 0} U_i(\mu_i, \mu^*; \lambda, k^*, N, p, v)$.

There is every reason to believe that a system with both strategic arrivals and strategic servers will behave much differently from one with only strategic arrivals or only strategic servers. This is because, the interplay of the arrival's joining decision and the servers' decision of how fast to work is not clear. Servers may work faster as k increases (as shown in Figure 1.3), but that may then induce more arrivals to join, which, in turn may then cause the servers to work slower. Such interaction adds more complexity to understanding joint equilibrium behavior. The only existing paper to study this interplay (Chung et al., 2020) establishes potentially surprising “benefit of anarchy” results in the restricted single-server setting ($N = 1$). Somewhat relevant, but in a modeling framework different from ours, is the empirical work in Altman et al. (2021) and Daw et al. (2023) showing how *individual* interactions between customers and servers affect the service rate.

The interplay between strategic arrivals and strategic servers has consequences for the system manager's ability to impose tolls on customers in order to ensure that customers acting in their selfish interest will induce a socially optimal outcome, as discussed in Naor (1969), wherein only arrivals are modeled as being strategic. The issue is that the individually optimal threshold exceeds the socially optimal one, a phenomenon that occurs in a broad class of such models (Hassin and Snitkovsky, 2020), because joining customers ignore the

negative externality that they impose on those that join after them (and must wait longer than they otherwise would have). Then, the imposition of a toll (or other joining deterrent, as discussed in Haviv and Oz (2016)) serves to lower the individually optimal joining threshold to the socially optimal one. However, when the servers are also strategic, the unpredictable behavior of the system load and expected wait time shown in Figure 1.11 suggests the need for a more complicated strategy to ensure a socially optimal outcome.

1.6.4 Utility Function Generalization

The utility function (1.2) linearly trades off the server's effort-cost with the payment and the long-run idleness fraction, where the effort-cost is increasing and convex in the service rate. Two natural generalizations of our utility model involve focusing on the modeling of effort-cost and idle time.

1.6.4.1 Effort-Cost Function

Other models of effort-cost have been used in the literature. For example, should this cost be based on μ , or should it only be incurred when the server is busy, as in Zhan and Ward (2018)? Moreover, different from our assumption that working faster becomes increasingly more taxing, from the prospect theory of Tversky and Kahneman (1992) and some behavioral experiments on learning such as Pooley and Bump (1993), the effort-cost may depend on context effects and exhibit concavity or an S-shape. Further work is needed to understand the effect of different effort-cost models on our results.

1.6.4.2 Idle Time

In general, a server's utility need not be a linear function of her idle time. For example, a server could value an increase of 1% a lot more when it corresponds to an increase from, say, 12% to 13%, than when it corresponds to an increase from, say, 92% to 93%. In order to

capture such a diminishing returns phenomenon, we could consider the utility function

$$U(\mu_1, \mu) = p \cdot \mu_1 B(\mu_1, \mu) + v \cdot I(\mu_1, \mu)^\alpha - c(\mu_1), \quad \alpha > 0, \quad (1.20)$$

which is a concave (and increasing) function of the idle time when $\alpha \leq 1$ ($\alpha = 1$ in (1.20) corresponds to the utility function (1.2)). The loss system ($k = N$) presents the simplest setting to study the resulting implications on server equilibria.

Proposition 9 (Equilibria under Generalized Utility). *In an $M/M/N/k$ system, when $k = N$, $p = 0$, and $v = 1$, under the utility function (1.20),*

(a) *For $\alpha \in (0, 1]$, $\mu^* > 0$ is an equilibrium if and only if it satisfies the corresponding FOC.*

(b) *For all $\alpha > 0$, the maximum candidate equilibrium, denoted by $\mu_{\max}^{*,?}$, is given by the unique solution to*

$$\mu c'(\mu) = \left(\frac{\alpha}{\alpha + 1} \right)^{\alpha+1}.$$

Furthermore, every server enjoys an idle time of $\frac{\alpha}{\alpha+1}$ when operating at $\mu_{\max}^{,?}$.*

The 50% idle time under μ_{\max}^* (when $p = 0$ and $v = 1$) from Proposition 3 is a result of setting $\alpha = 1$ in Proposition 9, i.e., the servers are risk-neutral. Intuitively, when the servers are risk-averse (respectively, risk-seeking), they enjoy more (respectively, less) idle time under $\mu_{\max}^{*,?}$.

Note that $\mu_{\max}^* = \mu_{\max}^{*,?}$ when $\alpha \in (0, 1]$, due to Proposition 9(a). When $\alpha > 1$, the utility function (1.20) is a strictly convex function of the idle time and $\mu_{\max}^{*,?}$ would only serve as an upper bound on the equilibrium service rate. Still, if the effort-cost function is “convex enough”, it may be possible to extend the maximum α for which the utility function (1.20) is concave beyond 1, extending Proposition 9(a) in the process.

Additionally, the underlying behavioral assumptions that servers want as much idleness

as possible may be suspect. Growing empirical research shed light on idleness aversion and the desire for purposeful busyness (Hsee et al., 2010; Wilson et al., 2014; Brodsky and Amabile, 2018; Yang and Hsee, 2019). As a result, servers' value for idleness may not be always increasing. More research is needed to understand the appropriate function to use to model how servers value idleness.

CHAPTER 2

LEARNING TO SCHEDULE IN MULTICLASS MANY-SERVER QUEUES WITH ABANDONMENT

2.1 Introduction

Scheduling problems have been extensively studied in the literature for a wide range of applications, including manufacturing networks (Pinedo (2012)), computer systems (Harchol-Balter (2013)), service systems such as call centers (Gans et al. (2003); Aksin et al. (2007)), and healthcare systems (Hopp and Lovejoy (2012)). A scheduling policy decides how server capacities (human employees or machines) are allocated over time to serve the incoming demand (customers or jobs). In a system with multiple classes of demand, the scheduling decision is critical to determine the quality of service each class receives (e.g., throughput, wait time).

Scheduling problems are often studied assuming that all the system characteristics (including distributional and parameter information) are known. However, in many applications, system characteristics are not known. There are two common ways to deal with this mismatch. The first is to design scheduling policies that are parameter agnostic; i.e., that make scheduling decisions based on the current system state, without using information on system characteristics (see Unlu and Zhong (2023) and references therein for works on parameter agnostic policies). For example, the scheduling policy that prioritizes the longest queue is such a policy. However, such policies can perform poorly compared to ones that also use information on system characteristics. A second approach is to spend effort learning about the unknown system characteristics (see Asanjarani et al. (2021) for a survey), and to then exploit the learned information to devise a scheduling policy. The work in this chapter is the first to follow the second approach in a queueing system in which customers with too long wait times may abandon the system without being served.

The focus of this chapter is to solve a scheduling problem in a multiclass many server queue with abandonment, when system characteristics are unknown, and abandonments are costly. In this queue, customers from different classes arrive with a service requirement and a patience time. The patience time determines how long the customer is willing to wait in queue before abandoning the system without being served. The unknown system characteristics are the model primitives for the multiclass $GI/GI/N+GI$ queue; i.e., the inter-arrival, service, and patience time distributions for each class. The scheduling policy decides which customer to next serve when a server becomes available. That decision is equivalent to choosing which class to next serve when customers within each class are served in a head-of-the-line (HL) fashion, with the designated HL customer in each class being the one that has waited the longest. Since the model primitives are unknown, the scheduling policy must dynamically learn the model primitives, or their summary statistics, through feedback from past scheduling decisions, and then base current scheduling decisions on the learned information. The inherent tradeoff between *exploring* different scheduling policies to learn the unknown model primitives and *exploiting* the most promising policy based on past observations parallels the exploration-exploitation tradeoff in the classical multi-armed bandit (MAB) problem.

Our objective in this chapter is to characterize the asymptotic convergence rate of *regret*, which is the difference in expected abandonment cost between a proposed policy and a benchmark policy under full knowledge of model primitives. The issue is that exact analysis appears intractable even when the model primitives are known due to the complicated state space. In order for the system to be Markovian, the state must track: (i) the time elapsed since the last arrival for each class; (ii) the amount of time each customer in service has been in service; and (iii) the amount of time each customer in queue has spent waiting. The state-space is infinite-dimensional, because there is no restriction on the number of customers that can be in queue; that is, the buffer size is infinite. The very large state space makes it

seemingly impossible to compute an exactly optimal scheduling policy, or to fully explore the state space to determine the optimal action in every system state. Fortunately, the $a\mu$ -rule (a state-independent, static priority policy that ranks classes using their class-dependent abandonment cost a and service rate μ , in the order of $a\mu$) performs well with respect to the expected abandonment cost under the stationary distribution. (The $a\mu$ -rule is exactly the well-known $c\mu$ -rule (Smith (1956)), except with the value of c modified to be the class-dependent abandonment cost instead of the class-dependent holding cost.) This motivates us to use the $a\mu$ -rule as the benchmark policy when defining regret. Then, we can consider a learning variant of the $a\mu$ -rule, which mimics the $a\mu$ -rule with empirical estimates used as surrogates for the unknown true parameters.

2.1.1 Contributions of This Chapter

Algorithm and Regret Upper Bound

We propose the Learn-then-Schedule (LTS) policy that is based on the LTS algorithm. The LTS algorithm is composed of a learning phase and an exploitation phase. During the learning phase, empirical estimates of the service rates are formed from past observations, with the goal of learning the benchmark $a\mu$ -rule. During the exploitation phase, an empirical $a\mu$ -rule based on those estimated rates is applied. We prove in Theorem 7 that the LTS policy has a regret upper bound of order $\log T$, where T is the system run-time.

One challenge in regret analysis in queueing systems is the treatment of transient system dynamics that arise at the beginning of every policy update. For this, we decompose the overall regret into the regret accumulated in the learning phase and the regret accumulated in the exploitation phase, and analyze each separately. The upper bound on the regret accumulated in the learning phase is linear in the length of the learning phase (Proposition 11), which implies that the learning phase length should be of order $\log T$. The upper bound on the regret accumulated in the exploitation phase is more complex (Proposition 12). It

depends on whether or not the correct $a\mu$ -ranking is learned in the learning phase. When the correct ranking is learned, the regret can be directly bounded (Lemma 9). When the correct ranking is not learned, the upper bound proof requires establishing performance guarantees for the empirical estimates of the service rates (Lemma 10), and the empirical $a\mu$ -rule based on those empirical estimates (Lemma 11).

Regret Lower Bound

We further show that our proposed LTS policy achieves the optimal regret rate, which is of order $\log T$. In particular, no other non-anticipating and non-preemptive policy can achieve a rate of regret that is smaller than order $\log T$ for all multiclass $GI/GI/N+GI$ queues. To provide a regret lower bound, we construct a specific problem instance, and show that the regret in that problem instance has a lower bound that is of order $\log T$. The problem instance we construct is a 2-class $D/D/1+M$ queue with inter-arrival and service times equal to one, and large abandonment rates for both classes, which converges to its stationary performance quickly (Proposition 10). When the abandonment rate of class 1 is larger (smaller) than that of class 2, the abandonment cost of class 1 is set to be correspondingly larger (smaller), making the benchmark $a\mu$ -rule the exact optimal policy (Lemma 7). We show in Theorem 6 that the regret in the aforementioned problem instance is at least of order $\log T$ under any admissible policy that eventually learns the benchmark $a\mu$ -rule; otherwise, the regret grows at least linearly over time (Lemma 8).

Static Scheduling Problem Benchmark

We use the solution to a static scheduling problem, that ignores system variability, as the benchmark policy when defining regret. The $a\mu$ -rule is our benchmark policy, because the $a\mu$ -rule solves the relevant static scheduling problem, and is asymptotically optimal for large systems that do not have sufficient capacity to serve all customers, when either the service

time distributions or patience time distributions are exponential (Proposition 13). Then, we only need to learn the static priority ranking of the classes given by the $a\mu$ -rule, that is based on first-order means, and eliminates the need to learn higher moments of the unknown distributions (which could result in higher-order regret when misestimated as shown in Ashutosh et al. (2021)). Consequently, the learning problem becomes significantly simplified. We believe that leveraging asymptotic analysis to define simpler learning problems holds promise for more applications in which the state space has an excessively large dimension, and we hope that the work in this chapter inspires more use of this approach.

2.1.2 Literature Review

The topic of combining statistical learning and optimal control has received considerable attention in operations management, and is studied in the settings of inventory control (Kunnumkal and Topaloglu, 2008; Huh and Rusmevichientong, 2014), assortment optimization (Sauré and Zeevi, 2013), network revenue management (Besbes and Zeevi, 2012; den Boer and Zwart, 2015), and large matching markets (Kalvit and Zeevi, 2022), among other application areas. Recent years have witnessed a growing body of works that combine statistical learning and optimal control in queueing systems, where the control spans pricing (Jia et al. (2022)), pricing and capacity sizing (Chen et al. (2023b,a)), scheduling (Krishnasamy et al. (2018, 2021)), and matching (Sun et al. (2018)); we refer the interested reader to Walton and Xu (2021) for a broader review of learning in queues. This chapter contributes to this stream of literature that particularly focuses on learning scheduling policies in queueing systems.

Krishnasamy et al. (2018) is the first paper that formulates the scheduling problem in multiclass single and many server queueing systems with unknown model primitives (and no abandonment) using the notion of finite-time regret. For a multiclass single server queue, the paper establishes that the empirical $c\mu$ -rule, which makes scheduling decisions using the empirical estimates of the service rates formed from past observations as surrogates for the

true means, has regret that is upper bounded by a constant that is independent of time, where regret is defined as the difference between the cumulative holding cost between the empirical $c\mu$ -rule and the $c\mu$ -rule under full parameter knowledge (which is known to be optimal). The constant regret result is possible because, in a stable system, inherent stochastic fluctuations result in seeing service completion samples from every class under the empirical $c\mu$ -rule, which eliminates the need for an explicit exploration strategy and provides for “free” exploration. In contrast, in an overloaded system with customer abandonment (i.e., one that is unstable if no customers abandon), which is the focus of this chapter, customers from lower priority classes may abandon instead of being served, meaning that service completion samples from lower priority classes may not be collected without an explicit exploration strategy.

Apart from Krishnasamy et al. (2018), there are several other relevant papers that aim to tackle the exploration-exploitation tradeoff between learning and scheduling in wireless networks (Krishnasamy et al. (2021)), load balancing (Choudhury et al. (2021)), best-channel identification (Stahlbuhk et al. (2018, 2021)), and decentralized bipartite queueing systems (Gaitonde and Tardos (2020); Sentenac et al. (2021); Freund et al. (2023)). However, none of these papers have customer abandonment.

The regret benchmark used in this chapter is an asymptotically optimal scheduling policy, when either the service time distributions or patience time distributions are exponential. Previous work has established the asymptotic optimality of $c\mu$ -type rules when either the patience time distributions are all exponential (Atar et al. (2010, 2011a, 2014)), or when the service time distributions are all exponential and the patience time distributions have non-decreasing hazard functions (Long et al. (2020)). Our contribution is to observe that when abandonment costs are the only system costs, then the $a\mu$ -rule is asymptotically optimal under more general conditions.

This chapter is also related to “scheduling with testing” problems (Sun et al. (2018); Levi et al. (2019)), where the optimal policy combines testing the jobs up to a certain

time and serving the jobs based on empirical estimates; “permutation” or “learning to rank” problems (Liu (2011); Fogel et al. (2015)), where the goal is to find an optimal order of items based on the observed information about individual items; and “single machine scheduling” problems (Smith (1956); Lee and Vojnovic (2021)), which ask to minimize the total weighted completion time for a fixed number of jobs, with weights c and processing times $1/\mu$. Finally, the learning of unknown model primitives in queues can also be addressed through a robust optimization approach (Bren and Saghafian (2019)).

2.1.3 Organization

The remainder of this chapter is organized as follows. We end this section with a summary of our mathematical notation. Section 2.2 formulates the scheduling problem in the multiclass $GI/GI/N+GI$ queueing system and defines our regret performance metric against the $a\mu$ -rule. A regret lower bound is established in Section 2.3. Then, we propose a policy based on an algorithm that we develop in Section 2.4 and prove its associated regret upper bound in Section 2.5. In Section 2.6, we evaluate the numerical performance of our proposed policy and the benchmark policy. In Section 2.7, we solve the static scheduling problem and show the asymptotic optimality of our benchmark policy. Finally, Section 2.8 concludes the chapter and points out future directions.

2.1.4 Notation

The following notation will be used throughout this chapter. We denote the set of integers endowed with the discrete topology by \mathbb{Z} , the set of non-negative integers by \mathbb{Z}_+ , the set of positive integers by \mathbb{N} , the set of real numbers endowed with the Euclidean topology by \mathbb{R} , and the set of non-negative real numbers by \mathbb{R}_+ .

Given a cumulative distribution function (abbreviated c.d.f. henceforth) F defined on \mathbb{R}_+ that is absolutely continuous with respect to Lebesgue measure and having probability

density function f , the right edge of its support is given by $H = \sup\{x \in \mathbb{R}_+ : F(x) < 1\} \in (0, \infty]$, and the associated hazard function is given by $h(x) = \frac{f(x)}{1-F(x)}$ for $x \in [0, H)$. We write $\bar{F} = 1 - F$.

2.2 Problem Setting

We study a learning variant of a canonical scheduling problem in a multiclass $GI/GI/N+GI$ queue, in which the inter-arrival, service, and patience time distributions are unknown. Section 2.2.1 formally defines the queueing model on a fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Section 2.2.2 defines our regret performance metric, and Section 2.2.3 justifies the benchmark policy we use to define our regret performance metric. We end in Section 2.2.4 with some technical assumptions.

2.2.1 The Multiclass $GI/GI/N+GI$ Queue

We consider a queueing system with a pool of N homogeneous servers that have identical capabilities to serve customers from J classes, as shown in Figure 2.1. Customers from class $j \in [J] := \{1, 2, \dots, J\}$ arrive according to a renewal process $\{E_j(t) : t \geq 0\}$ that is independent of all other customer arrival processes. Customers who cannot be served immediately upon arrival are kept in infinite-buffer queues dedicated to their classes. There, they wait in first come, first served (FCFS) order, so that the head-of-the-line (HL) customer is the one that has been waiting in the system the longest. A HL scheduling policy must specify whether an available server will work or idle when customers are waiting, and, when the decision is to work, which class the server will next serve.

Each customer arrives with a service time and a patience time that may depend on the customer class. The service time is the amount of time the customer needs to spend with a server, and the patience time is the maximum amount of time a customer will wait in the system to begin service. Customers whose patience time expire abandon the system without

receiving service, a phenomenon also known as reneging in the literature. The scheduling policy determines which classes have longer wait times (relative to their patience times), and, therefore, more abandonments.

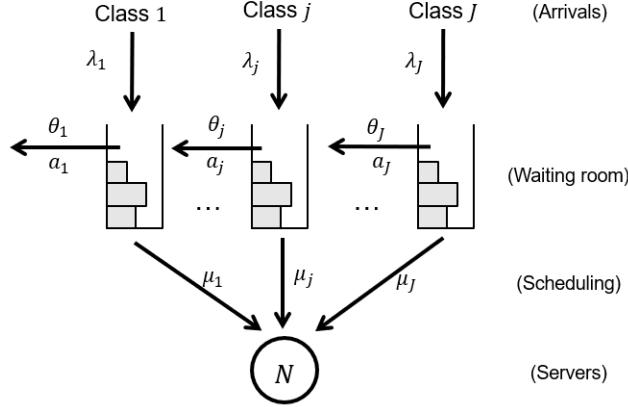


Figure 2.1: The multiclass $GI/GI/N+GI$ queue with abandonment cost

2.2.1.1 The Model Inputs

We assume that the inter-arrival, service, and patience times of each class $j \in [J]$ customer are i.i.d. sampled from cdf's G_j^a , G_j^s and G_j^r , respectively. We assume that G_j^a , G_j^s and G_j^r are absolutely continuous with density functions g_j^a , g_j^s and g_j^r respectively that have (possibly infinite) right edges of support H_j^a , H_j^s and H_j^r respectively, hazard functions h_j^a , h_j^s and h_j^r respectively, means $1/\lambda_j \in (0, \infty)$, $1/\mu_j \in (0, \infty)$ and $1/\theta_j \in (0, \infty)$ respectively, and variances $(\sigma_j^a)^2 \in (0, \infty)$, $(\sigma_j^s)^2 \in (0, \infty)$ and $(\sigma_j^r)^2 \in (0, \infty) \cup \{\infty\}$ respectively. Finally, we make the following exponential moment assumptions which ensure that certain large deviation estimates hold for the renewal processes associated with the inter-arrival and service times; that is, there exist some sufficiently small constants $\Upsilon^a > 0$ and $\Upsilon^s > 0$ such that $\Lambda_j^a(l) := \log \left(\int_0^{H_j^a} e^{lx} dG_j^a(x) \right) < \infty$ for all $l \leq \Upsilon^a$, and $\Lambda_j^s(l) := \log \left(\int_0^{H_j^s} e^{lx} dG_j^s(x) \right) < \infty$ for all $l \leq \Upsilon^s$, for all $j \in [J]$.¹ We further assume the service time distributions are sub-

1. Note that $\Lambda_j^a(l)$ and $\Lambda_j^s(l)$ are defined, with values in $(-\infty, \infty) \cup \{\infty\}$ for all values of l .

exponential² so that exponential tail bounds hold for the service time estimates; that is,
 $\log \left(\int_0^{H_j^s} e^{l(x-1/\mu_j)} dG_j^s(x) \right) \leq l^2(\sigma_j^s)^2/2$ for all $l \leq \Upsilon^s$, for all $j \in [J]$.

2.2.1.2 The State Descriptor

The state space for the multiclass $GI/GI/N+GI$ queues is exactly as described in Section 2 in Puha and Ward (2019). We repeat from that paper here. For $H \in [0, \infty]$, we let $\mathbf{M}[0, H)$ be the set of finite non-negative Borel measures on $[0, H)$, endowed with the topology of weak convergence, which is a Polish space.

The system state at time $t \geq 0$ is described as follows: for each $j \in [J]$,

- $\alpha_j(t) \in [0, H_j^a)$ is the time that has elapsed since the last class j customer arrived to the system;
- $X_j(t) \in \mathbb{Z}_+$ is the number of class j customers in the system (either in queue or being served);
- $\nu_j(t) \in \mathbf{M}[0, H_j^s)$, shown in Figure 2.2(a), encodes the length of time every class j customer in service at time t has been in service, and is also known as the age-in-service;
- $\eta_j(t) \in \mathbf{M}[0, H_j^r)$, shown in Figure 2.2(b), stores the amount of time that has passed between each class j customer's arrival time up until that customer's potential abandonment time (that is, the arrival time plus the sampled patience time), for every class j customer that arrived before time t , and without regard for whether or not that customer has entered service.³

2. Sub-exponential distributions have thinner tails than an exponential distribution, including normal distribution, exponential distribution, Weibull distribution, Pareto distribution, log-normal distribution, and gamma distribution, among others.

3. The measure η_j , $j \in [J]$, tracks class j customers that are “potentially” waiting in the queue. Customers “potentially” in the queue are those that have arrived, but whose potential abandonment time has not passed. The term potential refers to the fact that such customers may or may not have entered and/or finished service.

The measures ν_j and η_j track the evolution of unit atoms over time, where each atom is associated with a particular class j customer's time-in-service, or time-since-arrival, as shown in Figure 2.2.

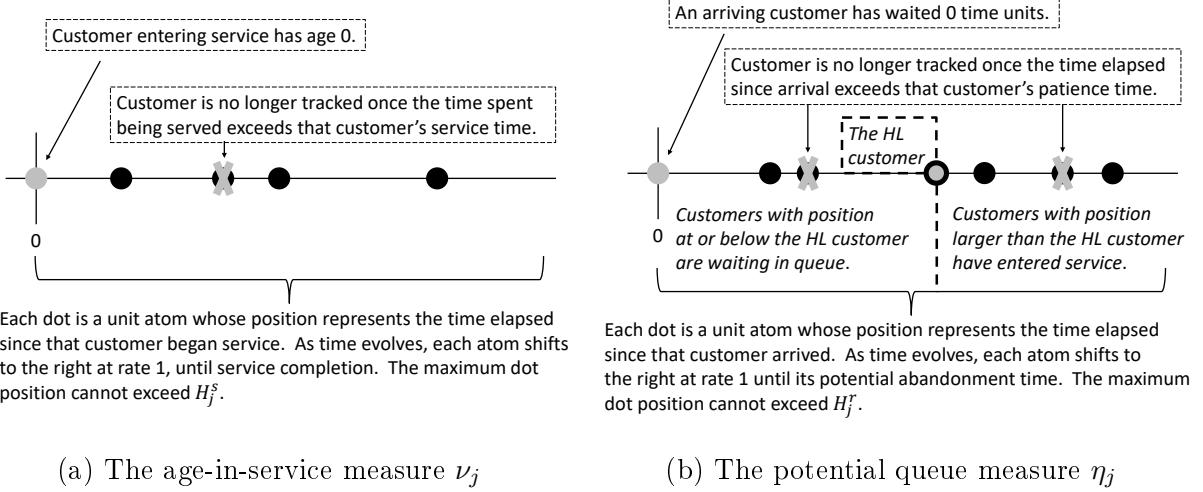


Figure 2.2: A graphic representation of the state space measures for a given class $j \in [J]$.

The state process $Y := (\alpha, X, \nu, \eta)$ is a right continuous process with left limits that takes values in

$$\mathbb{Y} := \mathbb{R}_+^J \times \mathbb{Z}_+^J \times \times_{j=1}^J \mathbf{M}[0, H_j^s) \times \times_{j=1}^J \mathbf{M}[0, H_j^r). \quad (2.1)$$

That is, for all $t \geq 0$, $Y(t) \in \mathbb{Y}$.

We do not provide the full system dynamics required to determine how the system state evolves over time, because that involves too much mathematical overhead and is not relevant to the focus of this chapter. Instead, we provide an idea of how certain processes of interest can be derived from the state, and refer the reader to Section 2 in Puha and Ward (2022) for the full system dynamics. Suppose the state at time $t \geq 0$ is $(\alpha(t), X(t), \nu(t), \eta(t))$. Then,

In particular, the potential customers waiting in queue are the HL customer and those that have waited less than the HL customer. All potential customers that have waited longer than the HL customer have entered service (and may or may not have finished service). The fact that the potential queue measure $\eta_j, j \in [J]$, is independent of the scheduling policy is helpful for analytic tractability.

the number of class $j \in [J]$ customers in service at time $t \geq 0$, is given by the total mass of $\nu_j(t)$ (i.e., count the number of dots on the x axis in Figure 2.2a); that is,

$$B_j(t) = \int_0^{H_j^s} \nu_j(t)(dx). \quad (2.2)$$

Furthermore, the number of class $j \in [J]$ customers waiting in queue at time $t \geq 0$ can be expressed in terms of the state variable $X_j(t)$ and $B_j(t)$ defined above as follows:

$$Q_j(t) = X_j(t) - B_j(t) \geq 0. \quad (2.3)$$

The state at all times $t \geq 0$ must satisfy the additional constraints that

- No more than N customers can be in service; i.e., $\sum_{j=1}^J B_j(t) \leq N$, and
- No customer in queue can have an expired patience time; i.e., $Q_j(t) \leq \int_0^{H_j^s} \eta_j(t)(dx)$.

2.2.2 Regret Performance Metric

The regret metric evaluates the performance of an admissible scheduling policy (that does not have knowledge of model primitives) against a benchmark policy (that has full knowledge of model primitives). An admissible scheduling policy is one that cannot use knowledge of future customer arrival, service, or patience times to make decisions (i.e., is non-anticipating), and also enforces that any customer taken into service must be served to completion (i.e., is non-preemptive).

Definition 4 (Admissible Policies). *The class of admissible scheduling policies, Π , consists of all non-anticipating and non-preemptive HL policies⁴.*

HL policies are common scheduling policies but are not always optimal, as can be seen from the single class $GI/GI/N+GI$ queue (Bassamboo and Randhawa (2016)).

4. A precise mathematical definition of non-anticipating and non-preemptive HL control policies for the multiclass $GI/GI/N+GI$ queue can be found in Definition 3 in Puha and Ward (2019).

Since customer abandonments are a clear indicator of customer dissatisfaction, they are costly. However, all abandonments may not be equally costly; that is, abandonments from designated VIP classes may be more costly than abandonments from non-VIP classes. To capture this, we assume that there is a class-dependent cost $a_j \in (0, \infty)$ incurred each time a class $j \in [J]$ customer abandons. Then, if $R_j(T; \pi)$ tracks the cumulative number of abandonments from class $j \in [J]$ over $[0, T]$ under scheduling policy $\pi \in \Pi$, the expected total cost over $[0, T]$ associated with scheduling policy π is

$$\mathcal{C}_T(\pi) := \mathbb{E} \left[\sum_{j=1}^J a_j R_j(T; \pi) \right]. \quad (2.4)$$

The regret performance metric measures the difference in expected total cost over a finite horizon between an admissible policy $\pi \in \Pi$ and a genie algorithm that has full knowledge of the model primitives. Let $\pi_T^* \in \Pi$ be a policy that minimizes the expected total cost over $[0, T]$; that is,

$$\mathcal{C}_T(\pi_T^*) := \inf_{\pi \in \Pi} \mathcal{C}_T(\pi).$$

Then, one strawman definition for the regret of scheduling policy $\pi \in \Pi$ at time $T \geq 0$ is

$$\mathcal{R}(T; \pi) := \mathcal{C}_T(\pi) - \mathcal{C}_T(\pi_T^*) = \mathbb{E} \left[\sum_{j=1}^J a_j (R_j(T; \pi) - R_j(T; \pi_T^*)) \right] \geq 0. \quad (2.5)$$

However, minimizing (2.5) directly is challenging, because the complicated state space prohibits us from exploring the entire state space to determine the optimal action in every system state, or providing an explicit characterization of π_T^* . Instead, we focus on the

long-run average expected cost⁵, defined as

$$\mathcal{C}(\pi) := \limsup_{T \rightarrow \infty} \frac{1}{T} \mathcal{C}_T(\pi), \quad (2.6)$$

and focus on finding a benchmark policy to replace π_T^* in (2.5) that performs well with respect to the long-run average expected cost.

The benchmark policy we use is the $a\mu$ -rule (which is the well-known $c\mu$ -rule, except with the value of c modified to be the class-dependent abandonment cost instead of the class-dependent holding cost).

Definition 5 (The $a\mu$ -Rule). *The $a\mu$ -rule, denoted by $\pi_{a\mu}$, is a non-preemptive, state-independent, static priority scheduling policy that ranks classes according to the descending order of their indices $a_j \mu_j$, $j \in [J]$, and, when a server becomes available, serves the HL customer from the class with waiting customers that has the highest index (and idles only if no customers are waiting).*

Then, we define our regret performance metric using the benchmark $a\mu$ -rule as follows.

Definition 6 (Regret). *The regret of policy $\pi \in \Pi$ at time $T \geq 0$ with respect to $\pi_{a\mu}$, defined in Definition 5, is*

$$\mathcal{R}_{a\mu}(T; \pi) := \mathbb{E} \left[\sum_{j=1}^J a_j (R_j(T; \pi) - R_j(T; \pi_{a\mu})) \right].$$

Note that the regret in Definition 6 can possibly be negative when T is small, because the benchmark policy $\pi_{a\mu}$, chosen for its long-run average performance, may not perform well in a short time period. In the remainder of the chapter, we aim to find a policy that minimizes the regret in Definition 6 as T becomes large.

5. The long-run average cost is finite because, for each class $j \in [J]$, for all $t \geq 0$, and under any $\pi \in \Pi$, $R_j(t; \pi) \leq E_j(t) + X_j(0)$ (that is, the cumulative number of abandonments is upper bounded by the cumulative number of arrivals together with the initial number of customers in the system), $\lim_{t \rightarrow \infty} \mathbb{E}[E_j(t)]/t = \lambda_j$ by the key renewal theorem, and $X_j(0) \in \mathbb{Z}_+$.

2.2.3 Benchmark Policy Justification

The graphs in Figure 2.3 demonstrate the performance of the benchmark policy $\pi_{a\mu}$ in the 2-class and 4-class $M/M/N+M$ queues (where the inter-arrival, service, and patience times are all exponentially distributed). The model parameters are set such that the system is overloaded in the sense that the arrival rates exceed the system capacity (i.e., $\lambda_1/\mu_1 + \lambda_2/\mu_2 > N$). We compare the associated cost in (2.4) under $\pi_{a\mu}$, based on 1000 independent simulation runs, to (i) that under π_T^* in (2.5) which minimizes the finite-horizon expected total cost, and (ii) that under a policy π^* which achieves the infimum of the long-run average expected cost (2.6); that is, $\mathcal{C}(\pi^*) := \inf_{\pi \in \Pi} \mathcal{C}(\pi)$. Both π_T^* and π^* can be computed numerically in this example.

We observe that $\pi_{a\mu}$ approximates the performance of π_T^* well for large T as N becomes larger. This is because, consistent with Markov decision process theory, π_T^* and π^* perform similarly for large T , and, later, in Section 2.7 (see Proposition 13 therein), we prove that π^* and $\pi_{a\mu}$ perform similarly when N is large. Note that the suboptimality gap with respect to π_T in Figure 2.3 is always positive, whereas the suboptimality gap with respect to π^* is only positive for large T (reflecting the fact that the long-run average optimal policy π^* may not perform well in the short term).

2.2.4 Technical Assumptions

The use of the benchmark policy $\pi_{a\mu}$ requires some assumptions. First, we must ensure that the rule is well-defined, which requires that all indices are distinct. (If there are two or more classes having the same index, multiple optimal scheduling policies exists.)

Assumption 2 (Index Separation). *There exists some finite positive constant Δ such that $|a_i\mu_i - a_j\mu_j| \geq \Delta$, for any $i, j \in [J]$ with $i \neq j$.*

Second, our analysis requires that the model inputs to the $GI/GI/N + GI$ queueing model are such that the following convergence rate result holds when the system operates

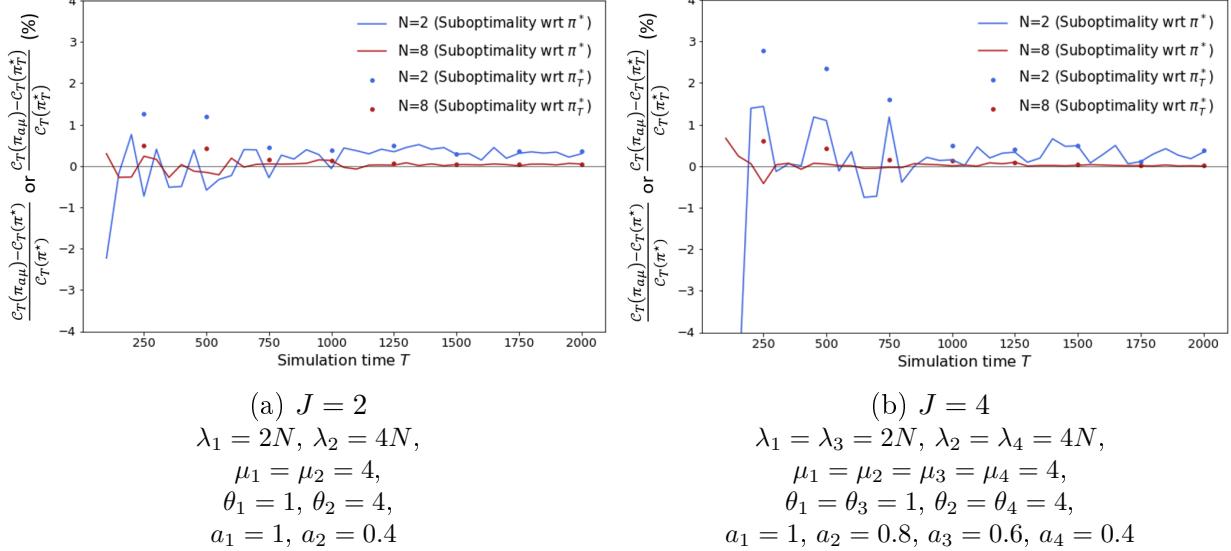


Figure 2.3: The suboptimality gap of the benchmark policy $\pi_{a\mu}$ with respect to π_T^* and π^* .

under $\pi_{a\mu}$. Let $D_j(t; \pi)$ track the cumulative number of service departures from class $j \in [J]$ over $[0, t]$ under policy $\pi \in \Pi$.

Assumption 3 (Convergence Rate). *The model inputs to the GI/GI/ $N + GI$ queue are such that when the queue operates under $\pi_{a\mu}$, the following holds:*

- (i) *The associated state process $\{Y(s; \pi_{a\mu}) : s \geq 0\}$ admits a stationary distribution;*
- (ii) *There exists some finite positive constant κ such that for any two distinct initial states $y, y' \in \mathbb{Y}$, for each $j \in [J]$, and for all $t > 0$:*

$$\left| \mathbb{E} \left[\frac{R_j(t; \pi_{a\mu}, y)}{t} \right] - \mathbb{E} \left[\frac{R_j(t; \pi_{a\mu}, y')}{t} \right] \right| \leq \frac{\kappa}{t},$$

or

$$\left| \mathbb{E} \left[\frac{D_j(t; \pi_{a\mu}, y)}{t} \right] - \mathbb{E} \left[\frac{D_j(t; \pi_{a\mu}, y')}{t} \right] \right| \leq \frac{\kappa}{t}.$$

Remark 6. Assumption 3 (i) is true by Theorem 4.2 in Atar et al. (2014). To verify Assumption 3 (ii), note that when patience times follow class-dependent exponential distributions, $\mathbb{E} \left[\frac{R_j(t)}{t} \right] = \theta_j \mathbb{E} \left[\frac{1}{t} \int_0^t (X_j(s) - B_j(s)) ds \right]$ for all $j \in [J]$, and when service times fol-

low class dependent exponential distributions, $\mathbb{E} \left[\frac{D_j(t)}{t} \right] = \mu_j \mathbb{E} \left[\frac{1}{t} \int_0^t B_j(s) ds \right]$ for all $j \in [J]$.

In either case, Assumption 3 (ii) simplifies to showing that the distance between two time-averaged state processes converge to zero at rate 1/time as time becomes large. When neither patience times nor service times follow exponential distributions, Assumption 3 (ii) is likely to be hard to verify due to the complex state descriptor (recalling Section 2.2.1.2)⁶.

2.3 Regret Lower Bound

We establish an order $\log T$ lower bound on the regret (see Definition 6) that any admissible policy (see Definition 4) can achieve over all multiclass $GI/GI/N+GI$ queues that satisfy Assumptions 2 and 3. To do this, we construct a specific problem instance in the overloaded regime, and show that the regret in that problem instance has a lower bound that is of order $\log T$.

Instance 1. *The problem instance is a 2-class $D/D/1+M$ queue, where the inter-arrival and service times for both classes are deterministic and equal to one, and the patience times for each class are exponentially distributed with rates $\theta_1 > 2.2$ and $\theta_2 > 2.2$, respectively. Moreover, the abandonment costs for each class satisfy $(a_1 - a_2)(\theta_1 - \theta_2) > 0$. Customers from both classes enter the system at integer-valued times $1, 2, 3, \dots$, and the system starts empty at time 0 (with no customers waiting or being served).*

In Problem Instance 1, the scheduling policy π_T^* in (2.5) (which minimizes the expected total cost over a finite horizon $[0, T]$ for any $T \geq 0$) is exactly the benchmark $a\mu$ -rule.

6. Assumption 3 (ii) can be expressed in terms of the state descriptor in Section 2.2.1.2. Specifically, from Section 4.1.1 in Puha and Ward (2022), the expected cumulative number of abandonments can be expressed by its martingale compensator; that is, $\mathbb{E}[R_j(t)] = \mathbb{E} \left[\int_0^t \left(\int_0^{\chi_j(s)} h_j^r(u) \eta_j(s)(du) \right) ds \right]$, $j \in [J]$, where $\chi_j(t)$ denotes the age of the HL class j customer at time t . Similarly, the expected cumulative number of service completions can be expressed by its martingale compensator; that is, $\mathbb{E}[D_j(t)] = \mathbb{E} \left[\int_0^t \left(\int_0^\infty h_j^s(u) \nu_j(s)(du) \right) ds \right]$, $j \in [J]$.

Lemma 7. *For any 2-class $D/D/1+M$ queue in Problem Instance 1, $\pi_T^* = \pi_{a\mu}$ for all $T \geq 0$.*

Problem Instance 1 is constructed such that there is an arrival from each class at every time unit, and the server can choose to serve either a customer from class 1 or from class 2. To learn the correct $a\mu$ -ranking, the server needs to strike a balance between prioritizing each class for service. This prioritization choice resembles the classic MAB problem of arm selection. In particular, prioritizing class 1 (i.e., selecting arm 1) results in more class 2 abandonments, and vice versa. There is a classic exploration-exploitation tradeoff between prioritizing each class to learn the unknown service rates, and exploiting the current empirical estimates to serve customers using an empirical $a\mu$ -rule. Then, as established in the seminal MAB paper Lai and Robbins (1985), we expect the regret lower bound to grow logarithmically over time.

We are interested in admissible polices that are eventually consistent with the benchmark $a\mu$ -rule. This is because any policy that never learns the benchmark $a\mu$ -rule (i.e., is inconsistent with the $a\mu$ -rule for all time) has regret that grows linearly over time, which is not desirable.

Assumption 4 (Consistency). *$T - \mathbb{E}[T(\pi_{a\mu})] = o(T^\alpha)$ for every $\alpha > 0$, where $T(\pi_{a\mu})$ denotes the cumulative time that the $a\mu$ -rule is applied over $[0, T]$.*

Lemma 8. *For any 2-class $D/D/1+M$ queue in Problem Instance 1, if Assumption 4 is not satisfied, then $\mathcal{R}_{a\mu}(T; \pi) = \Omega(T)$ under any $\pi \in \Pi$ (where Π is defined in Definition 4).*

Assumption 4 restricts the cumulative time that policies other than the $a\mu$ -rule are applied to $o(T^\alpha)$, which is sub-linear in T when $\alpha \in (0, 1)$. Such consistency assumptions are typically used in the MAB literature; see, for example, Lai and Robbins (1985); Burnetas and Katehakis (1996); Salomon et al. (2013).

Lemma 8 motivates analyzing consistent policies, that satisfy Assumption 4, in order to determine the minimum asymptotic regret rate.

Theorem 6. *For any 2-class $D/D/1+M$ queue in Problem Instance 1, under Assumption 4, $\mathcal{R}_{a\mu}(T; \pi) = \Omega(\log T)$ under any $\pi \in \Pi$ (where Π is defined in Definition 4).*

Theorem 6 is reminiscent of the regret lower bound results in classic MAB settings, that also find a lower bound on regret that is of order $\log T$. Similar to some recent MAB literature such as Bubeck and Cesa-Bianchi (2012); Combes et al. (2015); Perchet et al. (2016), the proof idea is to first lower bound the expected cumulative time of not applying the optimal policy (which is the benchmark $a\mu$ -rule from Lemma 7), and then translate this time-related lower bound into a lower bound on regret. However, in contrast to classic MAB problems where the cumulative count of suboptimal arm pulls directly translates to regret, our challenge lies in carefully constructing arguments to account for the continuous-time stochastic queueing dynamics when associating the cumulative time of employing a suboptimal policy to regret.

Given Theorem 6, if we can further verify that Problem Instance 1 satisfies Assumptions 2 and 3, then we can conclude that the smallest possible regret rate an admissible policy can achieve for all multiclass $GI/GI/N+GI$ queues satisfying Assumptions 2 and 3 is of order $\log T$. Assumption 2 can be readily verified in Problem Instance 1 since $a_1 \neq a_2$ and $\mu_1 = \mu_2$ ensures that $a_1\mu_1 \neq a_1\mu_2$. Assumption 3 is verified in the following result using the shift-coupling method (Thorisson (2000)).

Proposition 10. *For any 2-class $D/D/1+M$ queue in Problem Instance 1, the following hold for any $\pi \in \Pi$ (where Π is defined in Definition 4):*

(i) *The associated state process $\{Y(s; \pi) : s \geq 0\}$ admits a stationary distribution.*

(ii) *For any two distinct initial states $y, y' \in \mathbb{Y}$, for each $j \in [J]$, and for all $t > 0$:*

$$\begin{aligned} & \left| \mathbb{E} \left[\frac{R_j(t; \pi, y)}{t} \right] - \mathbb{E} \left[\frac{R_j(t; \pi, y')}{t} \right] \right| \\ & \leq 8 \left(3 + \frac{(1 - \delta_1)e\theta_1}{(e^{\theta_1} - e\theta_1)(1 - e\theta_1\delta_1)} + \frac{(1 - \delta_2)e\theta_2}{(e^{\theta_2} - e\theta_2)(1 - e\theta_2\delta_2)} \right) \frac{1}{t}, \end{aligned}$$

where δ_j is the smallest solution to $\delta_j = e^{-\theta_j(1-\delta_j)}$, for $j \in \{1, 2\}$.

Corollary 1. *For any 2-class $D/D/1+M$ queue in Problem Instance 1, Assumption 3 holds.*

2.4 The Proposed Learn-then-Schedule Policy

This section proposes a policy, that we term Learn-then-Schedule (LTS), whose associated regret upper bound has the same order as the regret lower bound established in Theorem 6 in Section 2.3. This implies that our proposed policy achieves the optimal regret rate, which is of order $\log T$, and no other admissible policy can achieve a smaller rate of regret for all multiclass $GI/GI/N+GI$ queues satisfying Assumptions 2 and 3.

The LTS policy has no a priori knowledge of model primitives (that is, the inter-arrival, service, and patience time distributions), and instead must employ an algorithm, that we also term Learn-then-Schedule (LTS), to learn the $a\mu$ -ranking. The LTS algorithm first invests effort to learn the unknown service rate μ_j for each class $j \in [J]$ (by routing customers to servers and observing the service completions), and second mimics the $a\mu$ -rule based on the learned service rates. In other words, the LTS policy is a learning-based variant of the $a\mu$ -rule that has two phases: a *learning* phase in $[0, \tau]$ for some $\tau \in [0, T]$, and an *exploitation* phase in $[\tau, T]$. During the learning phase, we form empirical estimates of the service rates, denoted by $\{\hat{\mu}_j(\tau)\}_{j \in [J]}$. During the exploitation phase, we apply the empirical $a\mu$ -rule, which uses the estimates $\{\hat{\mu}_j(\tau)\}_{j \in [J]}$ to define the class ranking (by substituting $\hat{\mu}_j(\tau)$ for μ_j in Definition 5). This procedure is explained in detail below, and is illustrated in Figure 2.4. The LTS algorithm pseudocode is given in Algorithm 1.

In the learning phase, we collect service completions to form empirical estimates of the service rate μ_j for each class $j \in [J]$. However, we note that in an overloaded system (i.e., one that is unstable if no customers abandon), customers from lower priority classes may abandon instead of getting served, meaning that service completion samples from lower priority classes may not be collected without an explicit exploration strategy (as was the

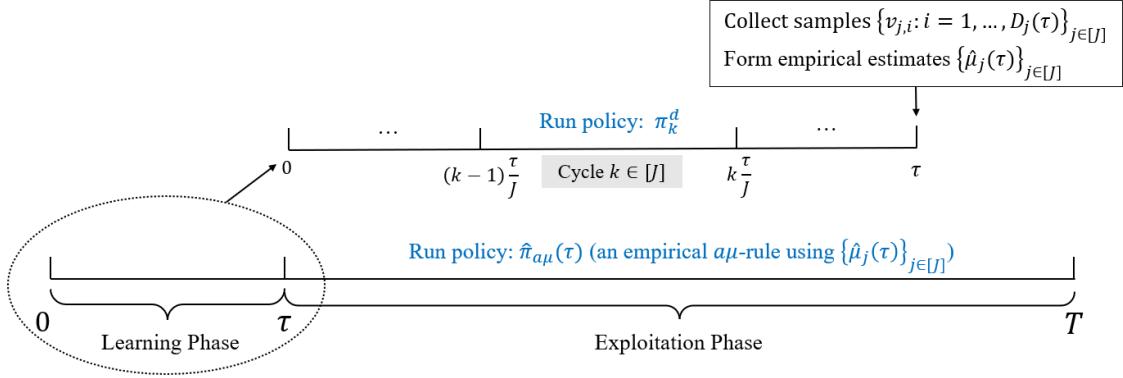


Figure 2.4: Illustration of the LTS policy, $\pi_{LTS}(\tau)$, that employs the LTS algorithm defined by Algorithm 1.

case in Problem Instance 1 in Section 2.3). Hence, we implement a strategy that takes turns serving each class during the learning phase. More specifically, we divide the learning phase, spanning time 0 to τ , into J cycles with equal length τ/J , as illustrated in the magnified segment of Figure 2.4. In each cycle $k \in [J]$, the system operates under the dedicated scheduling policy π_k^d defined below, that only serves class k customers and never serves all other classes.

Definition 7. *The non-preemptive dedicated scheduling policy, denoted by π_k^d , $k \in [J]$, only serves class k customers (either immediately upon arrival if idle servers are present, or after waiting in queue otherwise), and never serves other classes even while customers from those classes are waiting.*

Remark 7. *We choose to apply a dedicated policy in each cycle of the learning phase for analytical simplicity, because the system operates like a single-class queue within each cycle. Other policies may also be able to collect sufficiently many service completion samples from every class during the learning phase. For example, after dividing the learning phase into J equal cycles, one could implement a strategy that rotates the class priority rankings in each cycle to ensure every class has first position in the priority rankings at least once. Alternatively, one can also consider employing a randomized policy throughout the learning phase, where the class to serve is determined by tossing a J -sided dice.*

Algorithm 1 The Learn-then-Schedule (LTS) Algorithm

```

1: Input: Length of the learning phase  $\tau$ .
2: Initialization: The state at time 0,  $Y(0) = (Y_1(0), \dots, Y_J(0))$ .
   [Learning Phase]
3: for  $k \in [J]$  do
4:   while  $t \in \left[\frac{(k-1)\tau}{J}, \frac{k\tau}{J}\right]$  do
5:     Operate the system under a non-preemptive policy  $\pi_k^d$  (defined in Definition 7).
6:   end while
7:   Finish serving any customer in service at time  $\frac{k\tau}{J}$  to respect non-preemption.
8: end for
9: For each class  $j \in [J]$ , observe  $D_j(\tau)$  service completions during  $[0, \tau]$ , and their corresponding service times, denoted by  $\{v_{j,i} : i = 1, \dots, D_j(\tau)\}$ .
10: Output 1: For each  $j \in [J]$ , form an empirical estimate  $\hat{\mu}_j(\tau)$  using  $\{v_{j,i} : i = 1, \dots, D_j(\tau)\}$ .
11: Output 2: The total abandonment cost incurred in the learning phase.
   [Exploitation Phase]
12: while  $t \in [\tau, T]$  do
13:   Operate the system under a non-preemptive policy  $\hat{\pi}_{a\mu}(\tau)$  (defined in Definition 8).
14: end while
15: Output: The total abandonment cost incurred in the exploitation phase.

```

When the learning phase ends at time τ , we collect $D_j(\tau)$ service completion samples for each class $j \in [J]$, and denote their corresponding service times by $\{v_{j,i} : i = 1, \dots, D_j(\tau)\}_{j \in [J]}$. Using all these service completion samples, empirical estimates of the service rates can be constructed at time τ as follows

$$\hat{\mu}_j(\tau) = \frac{D_j(\tau)}{\sum_{i=1}^{D_j(\tau)} v_{j,i}}, \quad \forall j \in [J]. \quad (2.7)$$

Next, in the exploitation phase, we apply the empirical $a\mu$ -rule defined below.

Definition 8. *The empirical $a\mu$ -rule, denoted by $\hat{\pi}_{a\mu}(\tau)$, is as defined in Definition 5, with μ_j replaced by $\hat{\mu}_j(\tau)$ defined in (2.7).*

The LTS policy combines the policies applied in the learning and exploitation phases, as illustrated in Figure 2.4.

Definition 9. Given $\tau \in [0, T]$, the LTS policy, denoted by $\pi_{LTS}(\tau)$, is defined by Algorithm 1, which uses policy π_k^d in cycle $k \in [J]$ of the learning phase $[0, \tau]$, and uses policy $\hat{\pi}_{a\mu}(\tau)$ in the exploitation phase $[\tau, T]$, where π_k^d and $\hat{\pi}_{a\mu}(\tau)$ are as given in Definitions 7 and 8, respectively.

The LTS policy is admissible (that is, $\pi_{LTS}(\tau) \in \Pi$ where Π is as given in Definition 4) because the policies π_k^d , $k \in [J]$, and $\hat{\pi}_{a\mu}(\tau)$ are all non-anticipating and non-preemptive.

The performance of the LTS policy $\pi_{LTS}(\tau)$ heavily depends on the length of the learning phase, τ , which captures the tradeoff between exploration and exploitation. A larger value of τ implies more exploration. Recall from large deviations theory (see, e.g., Varadhan (1984); Shwartz and Weiss (1995); Deuschel and Stroock (2001)) that the tail probability of a sample mean exhibits an exponential decay rate. Then, in order to give good estimate $\hat{\mu}_j(\tau)$ for each class $j \in [J]$, we must collect at least order $\log T$ number of class j service completion samples. As a result, we expect the length of the learning phase τ to be at least of order $\log T$. We further expect that the smaller the index separation parameter Δ in Assumption 2, the more precise the estimates $\hat{\mu}_j(\tau)$, $j \in [J]$, must be, necessitating a longer learning phase. Since the abandonment cost incurred during the learning phase depends on its length, and the number of classes J whose service rates must be estimated, the overall regret upper bound must be at least of order $\log T$.

Theorem 7 (Regret Upper Bound). *Under Assumptions 2 and 3, if $\tau = \mathcal{O}\left(\frac{J \log T}{\Delta^2}\right)$, then $\mathcal{R}_{a\mu}(T; \pi_{LTS}(\tau)) = \mathcal{O}\left(\frac{J^2 \log T}{\Delta^2}\right)$.*

Note that the constants T , J and Δ appearing in Theorem 7 are known. The regret upper bound does not require knowledge of the unknown model inputs in Section 2.2.1.1.

2.5 Proof of Regret Upper Bound

In this section, we prove the regret upper bound in Theorem 7 for our proposed LTS policy, given in Definition 9 in Section 2.4. We do this by decomposing the regret $\mathcal{R}_{a\mu}(T; \pi_{LTS}(\tau))$ into that accumulated in the learning phase, denoted by

$$\mathcal{R}_{a\mu}^{Learn}(T; \pi_{LTS}(\tau)) := \mathbb{E} \left[\sum_{j=1}^J a_j (R_j(\tau; \pi_{LTS}(\tau)) - R_j(\tau; \pi_{a\mu})) \right], \quad (2.8)$$

and that accumulated in the exploitation phase, denoted by

$$\begin{aligned} & \mathcal{R}_{a\mu}^{Exploit}(T; \pi_{LTS}(\tau)) \\ &:= \mathbb{E} \left[\sum_{j=1}^J a_j \left((R_j(T; \pi_{LTS}(\tau)) - R_j(\tau; \pi_{LTS}(\tau))) - (R_j(T; \pi_{a\mu}) - R_j(\tau; \pi_{a\mu})) \right) \right]. \end{aligned} \quad (2.9)$$

Section 2.5.1 establishes an upper bound for (2.8), and develops the framework to establish an upper bound for (2.9) (which requires more effort). Section 2.5.2 provides a bound on the probability that the LTS policy has not correctly learned $\pi_{a\mu}$ at time τ , which is key to establishing an upper bound for (2.9). Finally, we choose τ in Section 2.5.3 to balance the regret accumulated in the learning and exploitation phases, which results in both τ being order $\log T$ and the regret being order $\log T$ (ignoring constants).

2.5.1 Regret Decomposition and Some Bounds

We expect the regret accumulated in the learning phase to have an upper bound that grows linearly in time, because the dedicated scheduling policies, π_k^d , $k \in [J]$, are all suboptimal. To gain intuition, note that the cumulative number of abandonments in each class is upper bounded by the cumulative number of arrivals (plus the number of initial customers), meaning that $\mathcal{R}_{a\mu}^{Learn}(T; \pi_{LTS}(\tau)) \leq \sum_{j=1}^J a_j (\mathbb{E}[E_j(\tau)] + X_j(0))$, and that $\mathbb{E}[E_j(\tau)]$ grows linearly in time.

Proposition 11 (Regret in the Learning Phase). Define $U_j := \sup_{x \geq 0} \frac{G_j^a(x) - \lambda_j \int_{u=0}^x \bar{G}_j^a(u) du}{G_j^a(x)} \in [0, \infty)$ for each $j \in [J]$. Then, given initial state $Y(0) = (\alpha(0), X(0), \nu(0), \eta(0)) \in \mathbb{Y}$, for all $T > 0$,

$$\mathcal{R}_{a\mu}^{Learn}(T; \pi_{LTS}(\tau)) \leq \sum_{j=1}^J a_j \left(\mathbb{E}[E_j(\tau)] + X_j(0) \right) \leq \left(\sum_{j=1}^J a_j \lambda_j \right) \tau + \sum_{j=1}^J a_j (X_j(0) + U_j).$$

The regret accumulated in the exploitation phase depends on whether or not the correct $a\mu$ -ranking is learned in the learning phase, as discussed below.

- (i) If $\hat{\pi}_{a\mu}(\tau) = \pi_{a\mu}$ (i.e., the LTS algorithm learns the correct $a\mu$ -ranking in the learning phase), then the regret accumulated in the exploitation phase from time τ to T captures the loss induced by different starting states at time τ .
- (ii) If $\hat{\pi}_{a\mu}(\tau) \neq \pi_{a\mu}$ (i.e., the LTS algorithm does not learn the correct $a\mu$ -ranking in the learning phase), then the regret accumulated in the exploitation phase captures the linear loss induced from using a suboptimal scheduling policy.

Combining the two cases above, it follows that

$$\begin{aligned} \mathcal{R}_{a\mu}^{Exploit}(T; \pi_{LTS}(\tau)) &= \mathbb{P}\{\hat{\pi}_{a\mu}(\tau) = \pi_{a\mu}\} \cdot \mathcal{R}_{a\mu}^{Exploit}(T; \pi_{LTS}(\tau) \mid \hat{\pi}_{a\mu}(\tau) = \pi_{a\mu}) \\ &\quad + \mathbb{P}\{\hat{\pi}_{a\mu}(\tau) \neq \pi_{a\mu}\} \cdot \mathcal{R}_{a\mu}^{Exploit}(T; \pi_{LTS}(\tau) \mid \hat{\pi}_{a\mu}(\tau) \neq \pi_{a\mu}). \end{aligned} \quad (2.10)$$

We next develop a bound for $\mathcal{R}_{a\mu}^{Exploit}(T; \pi_{LTS}(\tau) \mid \hat{\pi}_{a\mu}(\tau) \neq \pi_{a\mu})$ in (2.10). For this, we first observe that the cumulative number of abandonments in $[\tau, T]$ for each class is upper bounded by the cumulative number of arrivals in $[0, T]$ (plus the number of initial customers), and so $\mathcal{R}_{a\mu}^{Exploit}(T; \pi_{LTS}(\tau) \mid \hat{\pi}_{a\mu}(\tau) \neq \pi_{a\mu}) \leq \sum_{j=1}^J a_j (\mathbb{E}[E_j(T)] + X_j(0))$. Then, from the

second inequality in Proposition 11,

$$\mathcal{R}_{a\mu}^{Exploit}(T; \pi_{LTS}(\tau) \mid \hat{\pi}_{a\mu}(\tau) \neq \pi_{a\mu}) \leq \left(\sum_{j=1}^J a_j \lambda_j \right) T + \sum_{j=1}^J a_j (X_j(0) + U_j).$$

Substituting the above upper bound into (2.10), and noting that $\mathbb{P}\{\hat{\pi}_{a\mu}(\tau) = \pi_{a\mu}\} \leq 1$ shows that

$$\begin{aligned} \mathcal{R}_{a\mu}^{Exploit}(T; \pi_{LTS}(\tau)) &\leq \mathcal{R}_{a\mu}^{Exploit}(T; \pi_{LTS}(\tau) \mid \hat{\pi}_{a\mu}(\tau) = \pi_{a\mu}) \\ &\quad + \mathbb{P}\{\hat{\pi}_{a\mu}(\tau) \neq \pi_{a\mu}\} \cdot \left(\left(\sum_{j=1}^J a_j \lambda_j \right) T + \sum_{j=1}^J a_j (X_j(0) + U_j) \right). \end{aligned} \tag{2.11}$$

Equation (2.11) is the framework for bounding the regret accumulated in the exploitation phase. The first step is to bound the first term in (2.11). The issue when $\hat{\pi}_{a\mu}(\tau) = \pi_{a\mu}$ is that the expected system state at time τ differs depending on if the system was operated under the LTS policy or under the $a\mu$ -rule before time τ . Since we can regard the state at time τ as an initial state by the Markov property, the key is to show that the distance between the expected time-average cumulative number of abandonments becomes small quickly for two systems with distinct initial states that are both operated under the $a\mu$ -rule. We use Assumption 3 (ii) to do this. In particular, if the first condition in Assumption 3 (ii) holds, then an upper bound on $\mathcal{R}_{a\mu}^{Exploit}(T; \pi_{LTS}(\tau) \mid \hat{\pi}_{a\mu}(\tau) = \pi_{a\mu})$ can be immediately obtained. Otherwise, we can use conservation of mass to express the distance between the expected time-average cumulative number of abandonments in terms of the cumulative number of service completions and the system size. We can then bound each term using the second condition in Assumption 3 (ii), and an upper bound on the expected system size when no customers are served, in which case the system operates like J independent infinite server queues with service rates equal to the abandonment rates.

Lemma 9. *There exists some finite positive constant \check{X} such that $\mathbb{E}[X_j(t; \pi)] \leq \check{X}$ for all $t \geq 0$, $j \in [J]$ and $\pi \in \Pi$. Moreover,*

$$\mathcal{R}_{a\mu}^{Exploit}(T; \pi_{LTS}(\tau) \mid \hat{\pi}_{a\mu}(\tau) = \pi_{a\mu}) \leq (2\check{X} + \kappa) \left(\sum_{j=1}^J a_j \right),$$

where κ is given in Assumption 3.

The next term to study in the framework (2.11) is the probability of not correctly learning the true $a\mu$ -ranking, namely, $\mathbb{P}\{\hat{\pi}_{a\mu}(\tau) \neq \pi_{a\mu}\}$.

2.5.2 Bound on Estimation Error

Recall from the proposed LTS algorithm (Algorithm 1) that our sampling scheme involves observing the multiclass $GI/GI/N+GI$ queueing system over a continuous time period $[0, \tau]$, where $\tau \in [0, T]$ defines an appropriate length for learning. The data samples include $\{v_{j,i} : i = 1, \dots, D_j(\tau)\}_{j \in [J]}$, where $D_j(\tau)$ represents the total number of class $j \in [J]$ service completions incurred during the learning phase $[0, \tau]$.

To bound $\mathbb{P}\{\hat{\pi}_{a\mu}(\tau) \neq \pi_{a\mu}\}$, we must examine how well the empirical estimates $\hat{\mu}_j(\tau)$, $j \in [J]$, approximate the true service rates, $\mu_j, j \in [J]$. Specifically, the following result leverages the tail bound for sub-exponential random variables to bound the estimation error, under the condition that the number of samples grows at least linearly in τ .

Lemma 10. *For any $z \geq 0$ and $\epsilon > 0$, the following holds for all $\tau > 0$:*

$$\begin{aligned} & \mathbb{P}\{|a_j \hat{\mu}_j(\tau) - a_j \mu_j| \geq z \mid D_j(\tau) \geq \epsilon \tau\} \\ & \leq 2 \cdot \exp \left(-\frac{\epsilon \tau}{2} \min \left\{ \frac{1}{(\sigma_j^s)^2} \left(\frac{a_j \mu_j^2}{z} + \mu_j \right)^{-2}, \Upsilon^s \left(\frac{a_j \mu_j^2}{z} + \mu_j \right)^{-1} \right\} \right), \quad \forall j \in [J], \end{aligned}$$

where Υ^s is defined in Section 2.2.1.1.

The next step is to translate the conditional probability bound in Lemma 10 into a bound on the probability that the $\pi_{a\mu}$ rule is learned correctly (that is, that $\hat{\pi}_{a\mu}(\tau) = \pi_{a\mu}$). To do this, we use a large deviations bound for renewal processes (see, e.g., Appendix A in Bell and Williams (2001)) to establish a high-probability result that the sample size $D_j(\tau)$ grows linearly in τ for each $j \in [J]$, under the light tailed assumption made in Section 2.2.1.1. Then, we can uncondition, and rely on the fact that the sample sizes $D_j(\tau), j \in [J]$, are large enough to differentiate between the classes correctly (i.e., exactly as in the $a\mu$ -rule), given the index separation parameter Δ in Assumption 2.

Lemma 11. Define $\Lambda_j^{s,*}(x) := \sup_{l \in \mathbb{R}} (lx - \Lambda_j^s(l))$, $x \in \mathbb{R}$, for each $j \in [J]$. Under Assumption 2, there exists some finite positive constant $\psi \in (0, 1)$ such that the following holds for all large enough τ :

$$\mathbb{P}\{\hat{\pi}_{a\mu}(\tau) \neq \pi_{a\mu}\} \leq \left(3 + \max_{j \in [J]} \exp\left(\Lambda_j^s(\Upsilon^s)\right)\right) J e^{-\ell\tau},$$

where $\Lambda_j^s(\cdot)$ and Υ^s are defined in Section 2.2.1.1, and

$$\begin{aligned} \ell := & \frac{1}{J} \min \left\{ \frac{\psi}{6} \left(\min_{j \in [J]} \mu_j \right) \cdot \min_{j \in [J]} \frac{1}{(\sigma_j^s)^2} \left(\frac{2a_j \mu_j^2}{\Delta} + \mu_j \right)^{-2}, \frac{\psi}{6} \left(\min_{j \in [J]} \mu_j \right) \Upsilon^s \cdot \min_{j \in [J]} \left(\frac{2a_j \mu_j^2}{\Delta} + \mu_j \right)^{-1}, \right. \\ & \left. \frac{\Upsilon^s}{2}, \frac{\psi}{3} \left(\min_{j \in [J]} \mu_j \right) \cdot \min_{j \in [J]} \Lambda_j^{s,*} \left(\frac{3}{2} \left(\mu_j + \frac{1}{3} \min_{j \in [J]} \mu_j \right)^{-1} \right) \right\}. \end{aligned}$$

Remark 8. The constant ℓ , as defined in Lemma 11, is decreasing in $(\sigma_j^s)^2$, $j \in [J]$. This implies that the upper bound on $\mathbb{P}\{\hat{\pi}_{a\mu}(\tau) \neq \pi_{a\mu}\}$ becomes tighter as the variances in service times decrease.

2.5.3 Balancing the Regret Terms

Now, substituting the bounds in Lemma 9 and Lemma 11 into the framework (2.11) yields an upper bound on the regret accumulated in the exploitation phase.

Proposition 12 (Regret in the Exploitation Phase). *Given initial state $Y(0) = (\alpha(0), X(0), \nu(0), \eta(0)) \in \mathbb{Y}$, define $C_0 := (2\check{X} + \kappa) \left(\sum_{j=1}^J a_j \right)$, $C_1 := \left(3 + \max_{j \in [J]} \exp \left(\Lambda_j^s(\Upsilon^s) \right) \right) J \left(\sum_{j=1}^J a_j (X_j(0) + U_j) \right)$, and $C_2 := \left(3 + \max_{j \in [J]} \exp \left(\Lambda_j^s(\Upsilon^s) \right) \right) J \left(\sum_{j=1}^J a_j \lambda_j \right)$, where $\Lambda_j^s(\cdot)$ and Υ^s are defined in Section 2.2.1.1, \check{X} , κ and U_j are given in Lemma 9, Assumption 3 and Proposition 11, respectively. Then, under Assumptions 2 and 3, the following holds for all large enough T :*

$$\mathcal{R}_{a\mu}^{Exploit}(T; \pi_{LTS}(\tau)) \leq C_0 + C_1 e^{-\ell\tau} + C_2 T e^{-\ell\tau},$$

where ℓ is as defined in Lemma 11.

Finally, we combine the bounds on the regret accumulated in the learning and exploitation phases in Propositions 11 and 12 to establish the regret upper bound in Theorem 7. To achieve the stated $\mathcal{O}\left(\frac{J^2 \log T}{\Delta^2}\right)$ in Theorem 7, we balance the upper bound on the regret accumulated in the learning phase in Proposition 11 and the upper bound on the regret accumulated in the exploitation phase in Proposition 12. Specifically, we set τ so that the rate of regret accumulated in the learning phase, which is of order $J\tau$, is equal to the rate of regret accumulated in the exploitation phase, which is of order $J^2 T e^{-\ell\tau}$; that is, we find a finite positive constant ξ such that $J\tau = \xi J^2 T e^{-\ell\tau}$. After algebra, this can be equivalently written as $\tau = \ell^{-1} \log \xi J^2 + \ell^{-1} \log T - \ell^{-1} \log J\tau$, which implies $\tau = \mathcal{O}\left(\frac{\log T}{\ell}\right)$ for sufficiently large T . Since ℓ depends on $1/J$ and Δ^2 , $\tau = \mathcal{O}\left(\frac{J \log T}{\Delta^2}\right)$. Hence, when $\tau = \mathcal{O}\left(\frac{J \log T}{\Delta^2}\right)$, the upper bounds on the regrets accumulated in the learning and exploitation phases have the same order, so we have

$$\mathcal{R}_{a\mu}(T; \pi_{LTS}(\tau)) = \mathcal{R}_{a\mu}^{Learn}(T; \pi_{LTS}(\tau)) + \mathcal{R}_{a\mu}^{Exploit}(T; \pi_{LTS}(\tau)) = \mathcal{O}\left(\frac{J^2 \log T}{\Delta^2}\right),$$

which establishes Theorem 7.

2.6 Numerical Experiments

In this section, we use simulation to empirically evaluate the performance of our proposed LTS policy. Section 2.6.1 compares the cumulative abandonment cost under an optimal scheduling policy with respect to the long-run average cost (i.e., the optimal solution to (2.6)), the benchmark $a\mu$ -rule, and our proposed LTS policy. Section 2.6.2 provides evidence for the robustness of the algorithm's performance by varying the service and patience time distributions.

2.6.1 The LTS Policy Performance

We investigate the performance of the LTS policy in the 2-class $M/M/N+M$ queue (where the inter-arrival, service, and patience times are exponentially distributed). In this model, we can numerically compute an optimal policy. Then, when we calculate the cost incurred under the LTS policy, we can compare how much of that cost is due to learning the unknown service rates versus how much of that cost is due to the performance gap between the benchmark $a\mu$ -rule and an optimal policy.

The system state in the 2-class $M/M/N+M$ queue that we consider is much simpler than that for the more general model in Section 2.2.1. This is due to the memoryless property of the exponential distribution, and the assumption that the two classes have identical service time distributions ($\mu_1 = \mu_2$). Specifically, we need only track the number of customers waiting in the queue for each class, $Q_1(t)$ and $Q_2(t)$, as well as the number of customers in service, $B(t)$. Then, at any time $t \geq 0$, the system state $Y(t) = (Q_1(t), Q_2(t), B(t))$ satisfies $Y(t) \in \mathbb{Y}$, where $\mathbb{Y} = \mathbb{Z}_+^2 \times \{0, 1, \dots, N\}$.

We restrict attention to non-idling scheduling policies (because an optimal policy in the aforementioned queue is non-idling; see Section B.4.1 in the appendix to this chapter for the proof). Due to the non-idling condition, $Q_1(t) > 0$ or $Q_2(t) > 0$ if and only if $B(t) = N$.

The optimality equation of our long-run average cost MDP (see, e.g., Section 8.4.1 in Put-

erman (2014b)) is given below. For any state $(q_1, q_2, b) \in \mathbb{Y}$,

$$g + d(q_1, q_2, b) \cdot V(q_1, q_2, b) = \Xi V(q_1, q_2, b), \quad (2.12)$$

for some constant g , where $d(q_1, q_2, b) = \lambda_1 + \lambda_2 + q_1\theta_1 + q_2\theta_2 + b\mu$ is the rate at which transitions occur when the system is in state (q_1, q_2, b) , and

$$\begin{aligned} \Xi V(q_1, q_2, b) = & a_1 q_1 \theta_1 + a_2 q_2 \theta_2 + q_1 \theta_1 V(q_1 - 1, q_2, b) + q_2 \theta_2 V(q_1, q_2 - 1, b) \\ & + \begin{cases} \lambda_1 V(q_1 + 1, q_2, b), & \text{if } b = N; \\ \lambda_1 V(q_1, q_2, b + 1), & \text{if } b < N; \end{cases} + \begin{cases} \lambda_2 V(q_1, q_2 + 1, b), & \text{if } b = N; \\ \lambda_2 V(q_1, q_2, b + 1), & \text{if } b < N; \end{cases} \\ & + \begin{cases} 0, & \text{if } q_1 = 0, q_2 = 0, b = 0; \\ b\mu \cdot V(q_1, q_2, b - 1), & \text{if } q_1 = 0, q_2 = 0, b \geq 1; \\ b\mu \cdot V(q_1, q_2 - 1, b), & \text{if } q_1 = 0, q_2 \geq 1; \\ b\mu \cdot V(q_1 - 1, q_2, b), & \text{if } q_1 \geq 1, q_2 = 0; \\ b\mu \cdot \min\{V(q_1 - 1, q_2, b), V(q_1, q_2 - 1, b)\}, & \text{if } q_1 \geq 1, q_2 \geq 1. \end{cases} \end{aligned}$$

Note that $d(q_1, q_2, b)$ is unbounded, making the MDP not uniformizable, so we truncate the state space by imposing an arbitrarily large finite buffer size for both queues, denoted by L . In other words, when the queue length reaches L , newly arrived customers will be lost. As proved in Down et al. (2011) in a single server setting, an optimal MDP policy on the truncated space solves the original untruncated problem, as the truncation level L approaches infinity. In the experiment, we do not need to let L go to infinity, but set L large enough such that an optimal policy to the MDP becomes stable and identical to that for the original untruncated system. With such truncation, we can uniformize the continuous-time MDP to an equivalent discrete-time MDP using the uniform rate parameter $\gamma = \lambda_1 + \lambda_2 + L\theta_1 + L\theta_2 + N\mu$, and then solve it by the value iteration algorithm.

We adopt the same parameters as those used in Figure 2.3a in Section 2.2.3. We measure

the suboptimality gaps of the benchmark $a\mu$ -rule and our proposed LTS policy, respectively, with respect to an optimal MDP policy, by plotting the percentage increase of the cumulative abandonment cost versus simulation time from 100 to 2000. We calculate the cumulative abandonment costs (and the associated confidence intervals) using 1000 independent simulation runs with randomly sampled inter-arrival, service, and patience times. Figure 2.5 shows the suboptimality gaps of the benchmark $a\mu$ -rule (represented by red curves) and the proposed LTS policy (represented by blue curves) when $N = 2$ and $N = 8$. The independent replications allow us to reliably estimate the 95% confidence intervals, i.e., the 2.5th percentile to the 97.5th percentile of the 1000 simulation data, which are depicted by the shadows.

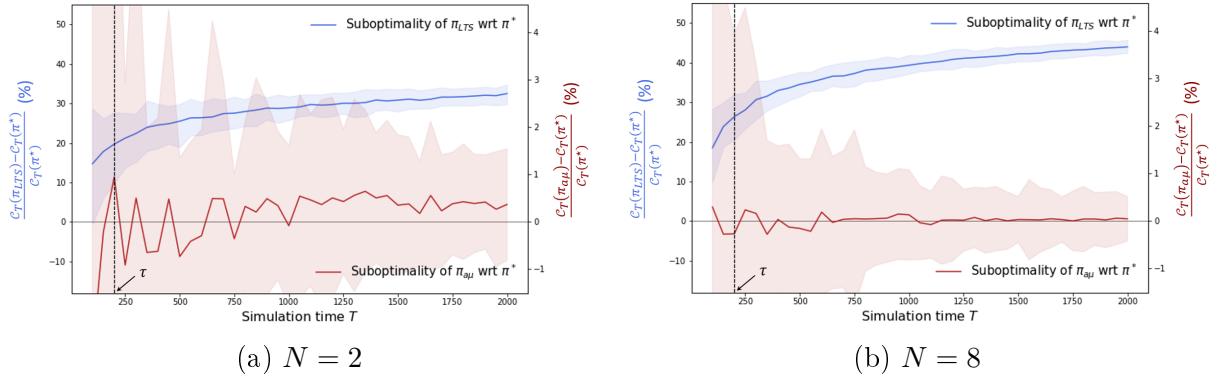


Figure 2.5: Comparison of the cumulative abandonment costs under the optimal MDP policy, the benchmark $a\mu$ -rule, and the proposed LTS policy, with 95% confidence intervals.

In this figure, $\lambda_1 = 2N$, $\lambda_2 = 4N$, $\mu_1 = \mu_2 = 4$, $\theta_1 = 1$, $\theta_2 = 4$, $a_1 = 1$, $\theta_2 = 0.4$.

We observe from Figure 2.5 that when the number of servers is small (Figure 2.5a), the percentage increase of the associated cost from optimal under our proposed LTS policy is a result of both the learning error arising out of the parameter estimation in the learning phase (represented by the gap between the two curves), and the suboptimality of the $a\mu$ -rule benchmark policy (represented by the red curve). This is because the $a\mu$ -rule (which performs well when the arrival rates and number of servers grow large) may not well approximate the optimal MDP policy when the number of servers N is small. On the other hand, when

$N = 8$ (Figure 2.5b), the percentage increase of the associated cost from optimal under our proposed LTS policy mostly comes from the learning error, with nearly zero suboptimality of the $a\mu$ -rule benchmark policy. This means that the $a\mu$ -rule approximates an optimal MDP policy very well when N is larger, which justifies using the $a\mu$ -rule as the benchmark when specifying the regret in Definition 6.

Consistent with Theorem 7, the regret grows logarithmically in time in Figure 2.5. Next, we confirm the robustness of this observation.

2.6.2 Performance Robustness

In this section, we plot regret over time in settings in which we cannot compute an optimal policy (in contrast to the Markovian setting in Section 2.6.1). Since Figure 2.5 in Section 2.6.1 suggests that $N = 8$ is large enough to ensure the benchmark policy $\pi_{a\mu}$ performs essentially identically to an optimal policy, we consider a many server queue with 8 servers; specifically, we consider a 2-class $M/GI/8+GI$ queue with Poisson arrival processes but non-exponential service and patience time distributions.

Figure 2.6 confirms the logarithmic growth of regret predicted by Theorems 6 and 7. The figure plots the regret of our proposed LTS policy, averaged over 1000 independent simulation runs, as a function of time from 100 to 2000, together with the 95% confidence intervals. We set the service and abandonment rates for each class identical to the specification used in Figure 2.5; that is, $\mu_1 = \mu_2 = 4$, $\theta_1 = 1$ and $\theta_2 = 4$. Figure 2.6a varies the service time distributions and Figure 2.6b varies the patience time distributions. Since all distributions have the same mean, we can focus on the effect of variability, which often hurts queueing system performance. However, we do not observe any monotonicity property of regret as we vary the variability of service or patience times.

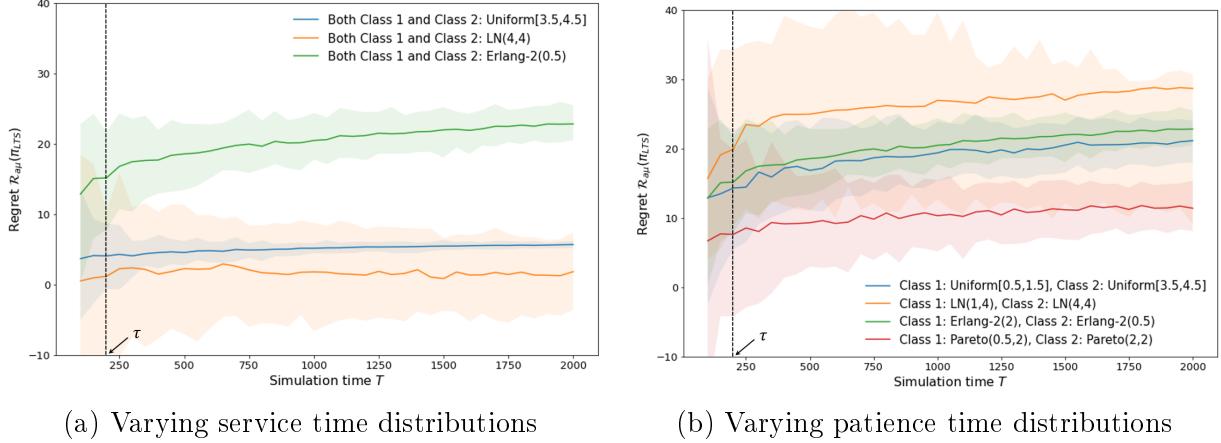


Figure 2.6: Regret of the LTS policy in the $M/GI/8+GI$ queue, for different service time and patience time distributions, with 95% confidence intervals. In this figure, $\mu_1 = \mu_2 = 4$, $\theta_1 = 1$, $\theta_2 = 4$, $a_1 = 1$, $\theta_2 = 0.4$. In panel (a), the patience time distributions for class 1 and class 2 are Erlang-2(2) and Erlang-2(0.5). In panel (b), the service time distributions for class 1 and class 2 are both Erlang-2(0.5). Note: LN denotes the log-normal distribution.

2.7 Asymptotic Optimality of the Benchmark $a\mu$ -Rule

We justify using $\pi_{a\mu}$ as our benchmark scheduling policy in the regret (see Definition 6) by first specifying a static scheduling optimization problem in Section 2.7.1, and second showing the asymptotic optimality of $\pi_{a\mu}$ under certain conditions in Section 2.7.2.

2.7.1 The Static Scheduling Problem

Suppose that we ignore the discrete and stochastic nature of arrivals, departures and abandonments, and instead assume that these flow at their long-run average rates. For each $j \in [J]$, the long-run average arrival rate for class j customers is λ_j , and the long-run average departure rate of class j customers depends on the long-run average fraction $b \in \mathbb{R}_+^J$ of server capacity given to that class, resulting in $b_j \mu_j \leq \lambda_j$, being the long-run average class j departure rate, and satisfying

$$\mathbb{B} := \left\{ b \in \mathbb{R}_+^J : b_j \leq \frac{\lambda_j}{\mu_j} \text{ for all } j \in [J] \text{ and } \sum_{j=1}^J b_j \leq 1 \right\}.$$

Then, to minimize the long-run average expected abandonment cost, we must decide on a scheduling policy that results in the optimal long-run average departure rate from each class. To do that, given $b \in \mathbb{B}$, we first observe that from conservation of mass, $\lambda_j - b_j\mu_j$ is the long-run average abandonment rate. Hence, the resulting static scheduling problem is

$$\inf_{b \in \mathbb{B}} \sum_{j=1}^J a_j(\lambda_j - b_j\mu_j) = \sum_{j=1}^J a_j\lambda_j - \sup_{b \in \mathbb{B}} \sum_{j=1}^J (a_j\mu_j)b_j. \quad (2.13)$$

We denote the solution to (2.13) by b^* , and then $b_j^*\mu_j$ represents the optimal long-run average departure rates from each class $j \in [J]$.

The solution to (2.13) motivates the $a\mu$ -rule, which is used as the benchmark scheduling policy $\pi_{a\mu}$ when defining the regret in Definition 6. Specifically, if we relabel the classes so that

$$a_1\mu_1 > a_2\mu_2 > \dots > a_J\mu_J, \quad (2.14)$$

then

$$b^* = \left(\frac{\lambda_1}{\mu_1}, \dots, \frac{\lambda_{j^*-1}}{\mu_{j^*-1}}, 1 - \sum_{k=1}^{j^*-1} \frac{\lambda_k}{\mu_k}, 0, \dots, 0 \right), \quad (2.15)$$

where

$$j^* := \min \left\{ j \in [J] : \sum_{k=1}^j \frac{\lambda_k}{\mu_k} > 1 \right\} \text{ if } \sum_{j=1}^J \frac{\lambda_j}{\mu_j} > 1, \text{ and } j^* := J + 1, \text{ otherwise.}$$

If we prioritize the classes in the stochastic system according to (2.14), then we expect the long-run average fraction of server capacity devoted to each class to be b^* . This scheduling policy assigns the maximum server effort required to ensure no long-run average abandonment cost, i.e., $b_j = \frac{\lambda_j}{\mu_j}$, to as many of the classes with a lower index $a_j\mu_j$ as possible.

Remark 9. *The objective function value in (2.13) at b^* is positive only when $\sum_{j=1}^J \frac{\lambda_j}{\mu_j} > 1$.*

Otherwise, when $j^* = J + 1$, the solution to (2.13) is $b_j^* = \lambda_j/\mu_j$, for all $j \in [J]$, and has objective function value 0 at b^* .

2.7.2 The Asymptotic Optimality Result

We consider a sequence of multiclass $GI/GI/N+GI$ queues, indexed by the number of servers N , with each N -server queue in the sequence following the model in Section 2.2.1, and having customer arrival rates $\lambda_j^N = N\lambda_j$ for each class $j \in [J]$. The N -server queue is operated under scheduling policy $\pi^N \in \Pi$ (see Definition 4), and incurs an expected cost $\mathcal{C}_T^N(\pi^N)$ over $[0, T]$ defined in (2.4).

To establish asymptotic optimality, it suffices to show that the static scheduling problem (2.13) arises as the limit of the stochastic scheduling problem (2.6), which requires the following assumption.

Assumption 5. For each $j \in [J]$, there exists $0 \leq L_j^s < H_j^s$ such that h_j^s is either bounded or lower-semicontinuous on (L_j^s, H_j^s) , and h_j^r is bounded.

The $a\mu$ -rule is known to be asymptotically optimal in the multiclass $GI/GI/N+M$ queue (see Proposition 5.1 and Theorem 5.1 in Atar et al. (2014)), and we show in Proposition 13 that the $a\mu$ -rule is also asymptotically optimal in the multiclass $GI/M/N+GI$ queue. A similar asymptotic optimality result was established for the multiclass $GI/M/N+GI$ queue in Theorem 4 in citelong2020dynamic, considering both abandonment and holding costs, under the condition of non-decreasing hazard functions for patience time distributions. Our next result observes that this additional assumption is unnecessary when considering only abandonment costs.

Proposition 13. Suppose Assumption 5 holds, G_j^r is strictly increasing for each $j \in [J]$, and either the service time or patience time distributions are all exponential.

(i) If $\{\pi^N\}_{N \in \mathbb{N}}$ is a sequence of admissible scheduling policies such that $\pi^N \in \Pi$ for each

$N \in \mathbb{N}$, and Equation (55) in Puha and Ward (2022) is satisfied for each $N \in \mathbb{N}$, then

$$\lim_{T \rightarrow \infty} \liminf_{N \rightarrow \infty} \frac{\mathcal{C}_T^N(\pi^N)}{NT} \geq \sum_{j=1}^J a_j(\lambda_j - b_j^* \mu_j). \quad (2.16)$$

(ii) If each N -server queue operates under scheduling policy $\pi_{a\mu}$, then

$$\lim_{T \rightarrow \infty} \lim_{N \rightarrow \infty} \frac{\mathcal{C}_T^N(\pi_{a\mu})}{NT} = \sum_{j=1}^J a_j(\lambda_j - b_j^* \mu_j).$$

Statement (i) in Proposition 13 establishes that the static scheduling problem (2.13) is an asymptotic lower bound for the stochastic scheduling problem objective (2.6), and statement (ii) in Proposition 13 establishes that $\pi_{a\mu}$ achieves that asymptotic lower bound. As a consequence, we conclude that $\pi_{a\mu}$ is asymptotically optimal for the multiclass $GI/M/N+GI$ and multiclass $GI/GI/N+M$ queues.

We conjecture Proposition 13 continues to hold for the multiclass $GI/GI/N+GI$ queue, and discuss the challenges in proving the conjecture following its statement.

Conjecture 1. *Proposition 13 holds for the multiclass $GI/GI/N+GI$ queue.*

Under the conditions in Proposition 13, Theorem 1 in Puha and Ward (2022) establishes that limit points of law-of-large-numbers scaled state descriptors solve the fluid model equations for the multiclass $GI/GI/N+GI$ queue given in Section 3 in Puha and Ward (2022) almost surely. Then, to prove Conjecture 1 (potentially under additional assumptions), the next step is to understand the long-time behavior of a fluid model solution. The most recent relevant results are Theorems 3.2 and 5.2 in Atar et al. (2021). Theorem 3.2 is for a single class setting, and assumes that the service time distribution is either bounded away from zero and infinity or has non-increasing associated hazard function. Theorem 5.2 additionally requires that service time distributions are class-independent, and assumes that patience time distributions are all exponential.

2.8 Concluding Remarks

In this chapter, we study a scheduling problem in a multiclass many server queueing system with abandonment, when the model primitives (including the inter-arrival, service, and patience time distributions) are unknown. The objective is to determine a scheduling policy that minimizes regret, which is the difference in expected abandonment cost between a proposed policy and a benchmark policy under full knowledge of model primitives. The benchmark policy we use is the $a\mu$ -rule, which is the solution to the static scheduling problem in an overloaded regime. We establish an order $\log T$ lower bound on the regret that any non-anticipating and non-preemptive scheduling policy can achieve over all multiclass $GI/GI/N+GI$ queues. Then, we prove an order $\log T$ upper bound on the regret accumulated under our proposed Learn-then-Schedule (LTS) policy, which matches the order of the regret lower bound. The LTS policy first learns empirical estimates of the service rates, and next schedules according to an empirical $a\mu$ -rule based on the estimated service rates.

Holding Costs

The cost (2.4) associated with a given scheduling policy uses only abandonment costs to penalize congestion. However, holding costs are also a natural way to penalize congestion, and are common in the literature. In our setting, if h_j is a class-dependent holding cost incurred per class $j \in [J]$ customer per unit time, then the static scheduling problem (2.13) becomes

$$\inf_{b \in \mathbb{B}} \sum_{j=1}^J a_j(\lambda_j - b_j \mu_j) + h_j q_j(b_j), \quad (2.17)$$

where, from Equation (54) in Puha and Ward (2022),

$$q_j(b) := \lambda_j \int_0^{(G_j^r)^{-1}(1-b_j \mu_j / \lambda_j)} (1 - G_j^r(x)) dx,$$

is the long-run average class $j \in [J]$ queue length under server capacity allocation $b \in \mathbb{B}$. The modified static scheduling problem (2.17) depends on the patience time distribution through $q_j(b)$. As a result, our proposed Learn-then-Schedule policy must be updated to learn information about both the patience time distributions and the service rates. Helpful information for such an update may be to recognize that when the patience time distributions have increasing hazard functions, then the objective function in (2.17) is concave, and when the patience time distributions have decreasing hazard functions, then the objective function in (2.17) is convex (see Lemma 1 in Puha and Ward (2019)).

In the special case that the patience time distributions are exponential, our proposed Learn-then-Schedule policy is easily updated to incorporate holding costs. This is because when the patience distributions are exponential, $q_j(b) = (\lambda_j - b_j\mu_j)/\theta_j$, with $b_j \in [0, \lambda_j/\mu_j]$ for each $j \in [J]$. Then, the solution to (2.17) is exactly as in Section 2.7.1, except with $\tilde{a}_j = a_j + h_j/\theta_j$ replacing a_j in (2.14). The Learn-then-Schedule policy in Definition 9 should be modified to also collect abandonment samples from the $J - 1$ classes not being served (to avoid censoring issues) during each of the J cycles in the learning phase. The proof of Theorem 7 can be updated in a straightforward manner to show that the regret under the modified Learn-then-Schedule policy also has an upper bound that is order $\log T$.

Regret Result Discussion

The LTS policy has regret that is optimal order; however, its regret upper bound depends on constants that may be further improved. Potentially, adaptive online learning algorithms such as the Upper Confidence Bound (UCB) algorithm (Auer et al. (2002a)) can do better, provided sufficient sampling from all classes in the learning phase.

Although the LTS policy achieves the optimal order of instance-dependent regret, we do not expect the LTS policy to achieve the optimal order of instance-independent regret, where an instance refers to a fixed multiclass $GI/GI/N+GI$ queue with given class-

dependent abandonment costs, as well as inter-arrival, service, and patience time distributions. This is because our LTS policy performs similarly to Explore-then-Commit algorithm in bandit settings. However, in that setting, Theorem 5.1 in Auer et al. (2002b) shows an instance-independent regret lower bound of order \sqrt{T} , whereas Theorem 1.1 in Slivkins (2019) shows that the Explore-then-Commit algorithm has the larger regret upper bound of order $T^{2/3}(\log T)^{1/3}$ (the smallest proved upper bound order). Ensuring a robust performance guarantee for worst-case regret is important, especially in safety-critical scenarios.

Other Open Questions

One open challenge is to quantify the performance gap between the $a\mu$ -rule, which we use to define regret in Definition 6, and an optimal policy, that achieves the smallest possible long-run average cost, as defined in (2.6). Our asymptotic optimality result guarantees that the performance gap is small as the arrival rates and number of servers become large, and Figures 2.3 and 2.5 show numerically that the performance gap can be small. However, more explicit bounds would help to understand when using the $a\mu$ -rule as the regret benchmark instead of an optimal policy is reasonable. This echoes the open question raised in Ward (2019) in a related context. We are hopeful that the recent paper by Braverman et al. (2020), which applies Taylor expansion to the value function to replace the Bellman equation with a differential equation, could be helpful.

Another open direction worth exploring is how fairness can be effectively integrated into the learn-to-schedule problem. Static priority policies perform well at the expense of lower priority classes. Fairness considerations in service system design are vital, although fairness could take on different definitions depending on the application. For example, in web servers and routers, equitable service for all job classes is essential, while in e-commerce applications, respecting the seniority of jobs in the queue becomes fundamental.

CHAPTER 3

ASYMPTOTICALLY OPTIMAL IDLING IN THE $GI/GI/N+GI$ QUEUE

3.1 Introduction

One common assumption when studying the $GI/GI/N+GI$ queue is that the service discipline is non-idling; that is, that servers do not idle when customers are present in the queue (Whitt (2006); Kang and Ramanan (2010); Kang et al. (2012); Zhang (2013); Kang and Pang (2019)). However, in the restricted $M/M/N+M$ setting, the paper Zhan and Ward (2019) (see Theorem 1, Proposition 1, and Example 1 therein) shows that in the presence of server utilization costs, a non-idling service discipline may not be asymptotically optimal. Our purpose in this chapter is to show that a similar phenomenon occurs in the $GI/GI/N+GI$ setting; that is, a non-idling service discipline might be suboptimal in the non-Markovian setting, when the system operates in a first-come, first-served (FCFS) manner.

The $GI/GI/N+GI$ queue is more difficult to analyze than the $M/M/N+M$ queue because the state descriptor is more complex. In particular, tracking the one-dimensional number-in-system process is sufficient when studying the $M/M/N+M$ queue, but more is needed when studying the $GI/GI/N+GI$ queue. This is because a Markovian state descriptor must also include knowledge regarding the time that has elapsed since the last arrival, the amount of time each job in service has been in service, and the amount of time each job in the queue has waited, resulting in a measure-valued state descriptor.

The control question is to determine when an available server should take the next customer into service, and when such a server should idle for some period of time. Too much idleness may lead to customer abandonment and excessive waiting, whereas too little rest may lead to server fatigue. To quantify these two competing interests, we consider an objective function that trades off the abandonment costs (and also, as an extension, holding

costs) with server utilization costs. Exact analysis of the $GI/GI/N+GI$ queue is intractable, and, therefore, we study the queue in an overloaded asymptotic regime in which the arrival rate and the number of servers become large. In that regime, we formulate a fluid control problem, and find that the solution to the fluid control problem sometimes motivates idling servers when customers are waiting (when operational costs are small compared to utilization costs). The policy we propose, and show is asymptotically optimal (see our main results in Theorems 8 and 9, and their extension to incorporate holding costs in Section 3.9), is one that “thins” the arrival process just enough to ensure the server utilization matches the solution to the fluid control problem.

Incorporating server utilization in the objective function is one way to ensure that the service discipline does not overwork servers. This can lead to increased employee retention, which can have performance benefits (discussed in Ward (2006)). Not overworking servers means ensuring sufficient idleness for all servers, an idea that arose earlier in papers that studied how to be fair to heterogeneous servers that can be grouped into statistically identical pools (see, e.g., Atar et al. (2011b), Ward and Armony (2013)), and how to exploit heterogeneous customers preferences so as to maximize revenue (see, e.g., Afeche and Pavlin (2016), Maglaras et al. (2018)).

Notation. We denote the set of integers endowed with the discrete topology by \mathbb{Z} , the set of non-negative integers by \mathbb{Z}_+ , the set of positive integers by \mathbb{N} , the set of real numbers endowed with the Euclidean topology by \mathbb{R} , and the set of non-negative real numbers by \mathbb{R}_+ . For F , a cumulative distribution function (abbreviated c.d.f. henceforth) on \mathbb{R}_+ with density f , we write $\bar{F} = 1 - F$ and recall that the right edge of the support is given by $x_r = \sup\{x \in \mathbb{R}_+ : \bar{F}(x) > 0\}$ and the hazard function is $x \mapsto f(x)/\bar{F}(x)$ for $x \in [0, x_r]$. For a measurable space (S, \mathcal{F}) and a measurable set $A \in \mathcal{F}$, 1_A is the indicator function of the set A , which is one when its argument is a member of the set A and is zero otherwise. In addition, when A is S , we use the shorthand notation 1 to mean 1_S . For $H \in (0, \infty]$, let

$\mathbf{M}[0, H)$ denote the set of finite, non-negative Borel measures on $[0, H)$ endowed with the topology of weak convergence. For a given $\eta \in \mathbf{M}[0, H)$ and a Borel measurable function $f : [0, H) \rightarrow \mathbb{R}_+$ that is integrable with respect to η , we write $\langle f, \eta \rangle = \int_{[0, H)} f(x)\eta(dx)$. The set $\mathbf{M}[0, H)$ endowed with the topology of weak convergence is a Polish space (Prokhorov (1956)). We let $\mathbf{0} \in \mathbf{M}[0, H)$ be the measure such that $\langle f, \mathbf{0} \rangle = 0$ for all Borel measurable functions $f : [0, H) \rightarrow \mathbb{R}_+$. Given $x \in [0, H)$, δ_x denotes the Dirac measure in $\mathbf{M}[0, H)$ such that for all Borel measurable functions $f : [0, H) \rightarrow \mathbb{R}_+$, $\langle f, \delta_x \rangle = f(x)$. Then let $\mathbf{M}_D[0, H)$ denote the subset of $\mathbf{M}[0, H)$ consisting of the measures $\eta \in \mathbf{M}[0, H)$ such that either $\eta = \mathbf{0}$ or η can be represented as a sum of finitely many Dirac measures, that is, $\eta = \sum_{i=1}^n a_i \delta_{x_i}$, for some finite $n \in \mathbb{N}$, $(a_1, \dots, a_n) \in (0, \infty)^n$ and $(x_1, \dots, x_n) \in [0, H]^n$. Given a Polish space \mathbb{S} , we use $\mathbf{D}(\mathbb{S})$ to denote the set of \mathbb{S} valued functions of \mathbb{R}_+ that are right continuous with finite lefts, endowed with the usual Skorokhod J_1 -topology. Finally, we use \Rightarrow to denote weak convergence and $\stackrel{d}{=}$ to denote equivalence in distribution.

3.2 The Model and Admissible Policy Class

In this chapter, we study a single-class many server queue with generally distributed inter-arrival, service, and patience times (i.e., a $GI/GI/N+GI$ queue) operating under a head-of-the-line (HL) control policy, that may or may not be non-idling. This is as specified in Puha and Ward (2021) specialized to a single customer class. In particular, we consider the model specified in Kang and Ramanan (2010), but with the non-idling condition (Kang and Ramanan, 2010, (2.30)) removed. Absent the non-idling condition, the system dynamics are not uniquely specified. Hence, one must specify a control policy to determine when each customer in system will commence service. Such control policies should satisfy natural conditions such as not using information about the future to make scheduling decisions. In what follows, we describe the model and admissible policy class in brief. We refer the interested reader to Puha and Ward (2021) for details.

3.2.1 The Model

Customers arrive according to a delayed renewal process E with rate $\lambda \in \mathbb{R}_+$, each with a service time sampled from c.d.f. G^s having finite mean $1/\mu \in (0, \infty)$, and a patience time (also known as reneging time) sampled from a c.d.f. G^r having finite mean $1/\theta \in (0, \infty)$. We denote the c.d.f. for the inter-arrival distribution associated with the renewal arrival as G . We assume G , G^s and G^r are absolutely continuous with density functions g , g^s and g^r respectively that have right edges of support H , H^s and H^r respectively and hazard function h , h^s and h^r respectively. We assume that there exists $0 \leq L^s < H^s$ such that h^s is either bounded or lower-semicontinuous on (L^s, H^s) and h^r is bounded and continuous. Boundedness of h^r implies that $H^r = \infty$. Finally, we assume G^r is strictly increasing with inverse function $(G^r)^{-1}$. The queue indexed by $N \in \mathbb{N}$ has N identical servers and is defined on a fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For the remainder of this chapter, we superscript all quantities that depend on N by N , e.g., G^N , g^N , H^N , λ^N and E^N depend on N , but G^s and G^r do not vary with N .

Following the notation in Section 2.2 in Puha and Ward (2021), the state descriptor for the N -server queue is denoted by $y^N = (\alpha^N, x^N, \nu^N, \eta^N) \in \mathbb{Y}_D$, where $\mathbb{Y}_D = \mathbb{R}_+ \times \mathbb{Z}_+ \times \mathbf{M}_D[0, H^s] \times \mathbf{M}_D[0, H^r]$. In particular, $\alpha^N \in [0, H^N]$ is the time that has elapsed since the last customer arrived to the system, $x^N \in \mathbb{Z}_+$ is the number of customers in system, $\nu^N \in \mathbf{M}_D[0, H^s]$ is a measure that has a unit mass at the age-in-service (amount of service received) of each customer currently in service, and $\eta^N \in \mathbf{M}_D[0, H^r]$ is a measure that has a unit mass at the potential waiting time of each customer “potentially” in system, (that is, each unit mass tracks the time passed since a customer’s arrival, until that customer’s patience time expires, at which point the unit atom is removed and tracking stops.) When $Y^N(0)$ denotes the initial state, the coordinate $\alpha^N(0)$ determines the distribution of the initial delay for E^N as the conditional distribution of G^N given $\alpha^N(0)$. That is, the initial delay distribution has density $g_0^N(x) = \frac{g^N(\alpha^N(0)+x)}{1-G^N(\alpha^N(0))}$ for $x \in [0, H^N - \alpha^N(0))$.

A state process for the N -server queue is a \mathbb{Y}_D valued, right continuous process Y^N with finite left limits that satisfies a set of dynamic equations for the N -server queue consistent with HL service. These are given as equations (5)-(26) in Puha and Ward (2021), which we omit here due to space constraints. With these, customers can only enter service at or after their arrival time and prior to their patience time expiring. An available server may idle or may take the customer in queue with the largest waiting time, the HL customer, into service. Once a server commences serving a customer, it works at rate one on the work associated with that customer until completely fulfilling that customer's service requirement, at which point the customer departs.

3.2.2 The Admissible Policy Class

The admissible policy class consists of all policies that only allow customers to enter service at moments of a customer departure or arrival, do not use information about the future, and are such that the state process Y^N is a Feller Markov process with respect to a natural filtration, and whose initial condition is policy compatible. The following leverages Puha and Ward (2021) to make this more precise.

As mentioned above, equations (5)-(26) in Puha and Ward (2021) do not uniquely specify the system dynamics. These are uniquely determined by the specification of an HL control policy $\pi^N = (\mathbb{S}^N, \{\mathbb{P}_y^N\}_{y \in \mathbb{S}^N})$. Here, as in Definition 1 in Puha and Ward (2021), \mathbb{S}^N is the Polish subspace of \mathbb{Y}_D that corresponds to the set of states that are achievable under the control policy. Also, for each initial state $y \in \mathbb{S}^N$, \mathbb{P}_y^N is a probability measure that uniquely determines the system dynamics when the system starts in state y . More formally, $\{\mathbb{P}_y^N\}_{y \in \mathbb{S}^N}$ is a collection of probability measures indexed by \mathbb{S}^N such that the mapping $y \mapsto \mathbb{P}_y^N(B)$ from \mathbb{S}^N to $[0, 1]$ is Borel measurable for each measurable $B \subset \mathbf{D}(\mathbb{S}^N)$ and, for

each $y \in \mathbb{S}^N$, \mathbb{P}_y^N almost surely,

$$Y^N(0) = y, \quad Y^N \in \mathbf{D}(\mathbb{S}^N) \text{ and satisfies (5) – (26) in Puha and Ward (2021).} \quad (3.1)$$

Given an HL control policy π^N , a state process Y^N satisfying (3.1) specifies an entry-into-service process K^N . Indeed, since a job has age-in-service equal to zero at the time of entering service, $\langle 1_{\{0\}}, \nu^N(t) \rangle$ is the number of jobs to enter service at time t , for each $t > 0$. Then K^N is a counting process such that $K^N(0) = 0$ and $K^N(t) - K^N(t-) = \langle 1_{\{0\}}, \nu^N(t) \rangle$ for each $t > 0$. In particular, $K^N(t)$ is the number of customers that enter service by time t for each $t \geq 0$. Then, for each $t \geq 0$, $D^N(t) = \langle 1, \nu^N(0) \rangle + K^N(t) - \langle 1, \nu^N(t) \rangle$ denotes the number of customers to depart the system due to service completion by time t . We restrict attention to HL policies that only allow customers to enter service at moments of a customer departure or arrival. We require that for each $y \in \mathbb{S}^N$, \mathbb{P}_y^N almost surely, for all $t \geq 0$,¹

$$K^N(t) - K^N(t-) \leq E^N(t) - E^N(t-) + D^N(t) - D^N(t-). \quad (3.2)$$

We allow for random initial states that are compatible with a given HL control policy $\pi^N = (\mathbb{S}^N, \{\mathbb{P}_y^N\}_{y \in \mathbb{S}^N})$. As in Definition 2 in Puha and Ward (2021), an initial distribution for π^N is a Borel probability measure ς^N on \mathbb{S}^N that determines the distribution of the initial state $Y^N(0)$. In particular, for each measurable $B \subset \mathbf{D}(\mathbb{S}^N)$, define $\mathbb{P}_\varsigma^N(B) = \int_{\mathbb{S}^N} \mathbb{P}_y^N(B) \varsigma^N(dy)$. Then \mathbb{P}_ς^N denotes the distribution of the state process Y^N under π^N for initial distribution ς^N . We say that an initial distribution ς^N for π^N is compatible if $\mathbb{E}_\varsigma^N [\langle 1, \eta^N(0) \rangle] < \infty$, where \mathbb{E}_ς^N denotes the expectation operator for \mathbb{P}_ς^N . Given an HL control policy π^N and a compatible initial distribution ς^N , we refer to the process Y^N with law \mathbb{P}_ς^N as the state process for (π^N, ς^N) .

In order to restrict attention to HL control policies that do not use information about

1. This condition is sufficient for a tightness result to hold as shown in Puha and Ward (2021).

the future, we require K^N to be non-anticipating. This amounts to requiring K^N to be adapted to a suitable filtration as in Definition 3 in Puha and Ward (2021). Because we consider long-run average cost, we make a further restriction in the definition of admissible HL control policies, which is used in Section 3.6 to establish the existence of a stationary distribution.

Definition 10 (Admissible Policies). *An admissible HL control policy for E^N is an HL control policy π^N such that for any compatible initial distribution ς^N , the pair (π^N, ς^N) (i) satisfies Definition 3 in Puha and Ward (2021) and (3.2) and (ii) is such that the state process Y^N for (π^N, ς^N) is a Feller Markov process with respect to the filtration used in Definition 3 in Puha and Ward (2021).*

Remark 10. *Our admissible policies focus on HL (equivalently, FCFS) control policies due to their common use in practice. However, non-HL control policies can be optimal in some settings; see Bassamboo and Randhawa (2016).*

Let Π^N denote the set of admissible HL control policies for E^N in Definition 10. For $\pi^N \in \Pi^N$, we will sometimes write $Y^N(\pi^N, \cdot)$, $X^N(\pi^N, \cdot)$, $\nu^N(\pi^N, \cdot)$, $\eta^N(\pi^N, \cdot)$, $K^N(\pi^N, \cdot)$ or $D^N(\pi^N, \cdot)$ to make the dependence on π^N explicit.

Proposition 14. *For any $\pi^N \in \Pi^N$, there exists a compatible initial distribution ξ^N such that the state process Y_∞^N for (π^N, ξ^N) is a stationary process.*

Proposition 14 follows as a special case of Lemma 12 stated in Section 3.7.

Given $\pi^N \in \Pi^N$ and a compatible initial distribution ξ^N such that the state process Y_∞^N for (π^N, ξ^N) is a stationary process, we refer to ξ^N as a compatible stationary distribution for π^N and we let $\mathcal{S}(\pi^N)$ denote the set of all compatible stationary distributions for π^N .

3.3 The Control Problem

Each customer abandonment incurs a cost $a \in (0, \infty)$ and the strictly increasing, continuous and convex function $g_U : [0, 1] \rightarrow [0, \infty)$ captures the cost of server utilization. The trade-off is between working the servers as much as possible, which incurs high utilization cost but low abandonment cost, and giving the servers more rest, which incurs lower utilization cost but higher abandonment cost. In particular, given $\pi^N \in \Pi^N$ and a compatible initial distribution ς^N , we define the long-run average cost of (π^N, ς^N) as

$$\mathcal{C}_\varsigma^N(\pi^N) := \limsup_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}_\varsigma^N \left[a \frac{R^N(\pi^N, T)}{N} + \int_0^T g_U \left(\frac{B^N(\pi^N, t)}{N} \right) dt \right],$$

where, for each $t > 0$, $R^N(\pi^N, t)$ is the cumulative number of abandonments by time t under π^N , and $B^N(\pi^N, t) \leq N$ is the number of busy servers at time t under π^N .

Proposition 15. *For any $\pi^N \in \Pi^N$ and compatible initial distribution ς^N , there exists $\xi^N \in \mathcal{S}(\pi^N)$ such that $\mathcal{C}_\varsigma^N(\pi^N) = \mathcal{C}_\xi^N(\pi^N)$.*

Proposition 15 follows as a special case of Lemma 13 stated in Section 3.7.

Given $\pi^N \in \Pi^N$, let $\mathcal{C}^N(\pi^N) := \sup_{\xi^N \in \mathcal{S}(\pi^N)} \mathcal{C}_\xi^N(\pi^N)$ denote the worst case cost. By Proposition 15, $\mathcal{C}^N(\pi^N)$ is the supremum of $\mathcal{C}_\varsigma^N(\pi^N)$ over all compatible initial distributions ς^N . Our objective is to find an admissible control policy π_{opt}^N such that

$$\mathcal{C}^N(\pi_{\text{opt}}^N) := \inf_{\pi^N \in \Pi^N} \mathcal{C}^N(\pi^N). \quad (3.3)$$

The objective is such that a non-idling control policy is not in general optimal. Based on the discrete-event queuing model, it is not possible to solve for π_{opt}^N exactly. Thus, we leverage an analytically tractable approximating fluid control problem to postulate an HL control policy that one might expect to perform well for the objective (3.3). Then, we show that this policy is asymptotically optimal (see Theorems 8 and 9 in Section 3.6).

3.4 The Fluid Control Problem

The fluid control problem is based on the fluid model and the fluid model solutions defined in Puha and Ward (2021). Fluid model solutions arise as functional law of large numbers limits of sequences of state descriptors for the stochastic system under fluid scaling. For each $N \in \mathbb{N}$, we define the fluid scaling for the N -server system as follows. Recall the constant λ^N and the processes $E^N, \alpha^N, X^N, \nu^N, \eta^N, K^N$ and D^N defined in Section 3.2, and the processes R^N and B^N defined in Section 3.3; also define the process $Q^N = X^N - B^N$ as the queue length, and the process $I^N = N - B^N$ as the number of idle servers. Then, let $\bar{\alpha}^N = \alpha^N$; also for $\triangle^N = \lambda^N, E^N, X^N, \nu^N, \eta^N, K^N, D^N, R^N, B^N, Q^N, I^N$, let $\bar{\triangle}^N = \triangle^N/N$. Then, the fluid-scaled state process for the N -server system is $\bar{Y}^N = (\bar{\alpha}^N, \bar{X}^N, \bar{\nu}^N, \bar{\eta}^N)$. Under suitable asymptotic conditions, limit points exist and are fluid model solutions almost surely (see Lemma 15 in Section 3.7.)

In particular, fluid model solutions are functions of time that take values in the set $\mathbb{X} = \mathbb{R}_+ \times \mathbf{M}[0, H^s] \times \mathbf{M}[0, H^r]$ endowed with the product topology. Then a state $(x, \nu, \eta) \in \mathbb{X}$ for the fluid model is a fluid analog of the state descriptor for the stochastic system with x , $\langle 1_{[0,z]}, \nu \rangle$ and $\langle 1_{[0,z]}, \eta \rangle$ corresponding to the total mass in system, the total mass in service with age-in-service less than or equal to z for each $z \in \mathbb{R}_+$, and the total mass potentially in system of age less than or equal to z for each $z \in \mathbb{R}_+$, respectively. They satisfy a set of conditions determined by a positive constant γ , which is the rate at which "fluid" or mass arrives to the system. These conditions are referred to as the fluid model for γ . We summarize the fluid model for γ and the definition of a fluid model solution for γ in C.1.

The invariant states for the fluid model for γ are fixed points of the fluid model for γ . From Proposition 1 in Puha and Ward (2021), an invariant state for γ is determined by the long-run average fraction of the collective server effort provided to the customers, denoted by b . It is clear that b must satisfy $b \in [0, \min\{1, \gamma/\mu\}]$, where we recall that μ is the reciprocal of the mean of G^s . Then, when the initial state for a fluid model solution for γ is an invariant

state for γ , it turns out that the departure rate of the fluid from the system is $b\mu$ and so, by conservation of mass, $\gamma - b\mu$ must be the rate at which fluid abandons. This implies that the abandonment rate is insensitive to the patience time distribution, which has a similar flavor to the insensitivity result for a single server queue in the large deviations regime in Atar et al. (2019).

Assumption 6. Let $\lambda \in (0, \infty)$. Suppose that $\lim_{N \rightarrow \infty} \bar{\lambda}^N = \lambda$.

Henceforth, λ satisfying the conditions in Assumption 6 is fixed. Our fluid control problem is based on the invariant states for λ . We expect to obtain the following fluid control problem for λ when letting $N \rightarrow \infty$ in problem (3.3).

Definition 11 (The Fluid Control Problem). *The fluid control problem for λ is given by*

$$\min_{b \in [0, \min\{1, \lambda/\mu\}]} a(\lambda - b\mu) + g_U(b). \quad (3.4)$$

We denote the solution to (3.4) by b_* (which exists and is unique because (3.4) optimizes a convex function over a compact set).

Example 2. Suppose $a = 1$ and $g_U(b) = b^2$. Then, the solution to (3.4) is $b_* = \min\{1, \frac{\mu}{2}, \frac{\lambda}{\mu}\}$.

The solution to (3.4) motivates a control policy that we expect to have good performance with respect to the original objective (3.3) when the arrival rate λ^N and the number of servers N are large. When $b_* = \min\{1, \lambda/\mu\}$, we expect a non-idling control policy to be optimal for (3.3). Otherwise, when $b_* < \min\{1, \lambda/\mu\}$, the solution to the fluid control motivates defining a policy that uses customer abandonments to trim congestion, in order to reduce server workload, and provide (additional) server idle time. In this case, for each $N \in \mathbb{N}$, consider the HL control policy $\tilde{\pi}^N$ such that each server idles after each service completion for the difference between the desired expected time between service completions, $(b_*\mu)^{-1}$, and the expected time between service completions when the server is always busy, μ^{-1} ; that

is, for $(b_*\mu)^{-1} \circ \mu^{-1} = (1 - b_*)(b_*\mu)^{-1}$ time units. Such a policy seems quite reasonable, and should be asymptotically optimal. However, establishing that for any sequence of compatible initial distributions $\{\varsigma^N\}_{N \in \mathbb{N}}$,

$$\begin{aligned}\lim_{N \rightarrow \infty} \lim_{t \rightarrow \infty} \frac{1}{t} \mathbb{E}_\varsigma^N \left[\bar{R}^N(\tilde{\pi}^N, t) \right] &= \lambda - b_*\mu \quad \text{and} \\ \lim_{N \rightarrow \infty} \lim_{t \rightarrow \infty} \mathbb{E}_\varsigma^N \left[g_U \left(\bar{B}^N(\tilde{\pi}^N, t) \right) \right] &= g_U(b_*)\end{aligned}\tag{3.5}$$

is difficult. This difficulty is related to a lack of results providing sufficient conditions for fluid model solutions to converge to invariant states in the time infinity limit (see Section 7.1 in Kang et al. (2012)). Instead, we propose to expand the admissible policy class to include thinned arrival processes and then rely on results in the literature for non-idling many server queues to show that (3.5) holds. If we can show a policy is asymptotically optimal for an enlarged policy class, then we know that no policy in the original smaller policy class can perform better.

3.5 The Proposed Policy π_*^N

The solution $0 \leq b_* \leq \min\{1, \lambda/\mu\}$ to (3.4) represents the optimal long-run average fraction of busy servers, which suggests that a control policy that thins the arrival process to rate $b_*\mu$ and forces the servers to work in a non-idling fashion, but builds in idleness due to admission control, should perform well for the original objective (3.3). This motivates us to enlarge the admissible policy class in Definition 10 to allow for admission control. Specifically, at the time of each arrival, let $p \in (0, 1]$ be the probability the arrival is admitted for service and $1 - p$ the probability the arrival is rejected, which incurs a cost a . Given $p \in (0, 1]$, we denote the admitted arrival process by E_p^N , and we refer to the N -server queue with arrival process E_p^N as the p -admitted queue. It is clear that the thinned arrival process E_p^N is a suitably delayed renewal process with arrival rate $p\lambda^N$, because the admitted arrivals remain i.i.d..

Definition 12 (Enlarged Admissible Policies). *For any $p \in (0, 1]$, an admissible HL control policy for E_p^N satisfies Definition 10 with E^N replaced by E_p^N .*

For $p \in (0, 1]$, let Π_p^N denote the set of admissible HL control policies for E_p^N . Note that $\Pi_1^N = \Pi^N$. For $p \in (0, 1]$, $\pi_p^N \in \Pi_p^N$ and $\Delta^N = Y^N, X^N, \nu^N, \eta^N, K^N, D^N, R^N, B^N, Q^N$ or I^N , $\Delta^N(\pi_p^N, \cdot)$ refers to the process for the p -admitted queue under π_p^N .

Given $p \in (0, 1]$, $\pi_p^N \in \Pi_p^N$ and a compatible initial distribution ς^N , the long-run average cost of (π_p^N, ς^N) is

$$\begin{aligned} \mathcal{C}_\varsigma^N(\pi_p^N) := & \limsup_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}_\varsigma^N \left[a \left(\bar{E}^N(T) - \bar{E}_p^N(T) + \bar{R}^N(\pi_p^N, T) \right) \right. \\ & \left. + \int_0^T g_U \left(\bar{B}^N(\pi_p^N, t) \right) dt \right]. \end{aligned} \quad (3.6)$$

When the initial state for the fluid model for $p\lambda$ is an invariant state for $p\lambda$ associated with $b \in [0, p\lambda/\mu]$, $p\lambda - b\mu$ is the rate at which fluid abandons and $(1-p)\lambda$ is the rate at which fluid is rejected. Since $p \in (0, 1]$ is a parameter that can be optimized over, the resulting fluid control problem is given by

$$\begin{aligned} & \min_{p \in (0, 1], b \in [0, \min\{1, p\lambda/\mu\}]} a(1-p)\lambda + a(p\lambda - b\mu) + g_U(b) \\ &= \min_{b \in [0, \min\{1, \lambda/\mu\}]} a(\lambda - b\mu) + g_U(b). \end{aligned} \quad (3.7)$$

The solution to (3.7) does not depend on the admission control parameter $p \in (0, 1]$ and is identical to the solution to (3.4). This observation crucially relies on the abandonment cost being linear with the per unit cost equal to the per unit cost of rejection.

This gives us flexibility to propose a policy in Π_p^N for various choices of $p \in (0, 1]$. We first observe that an optimal admission control parameter must lie in $[b_*\mu/\lambda, 1]$, because otherwise the admitted arrivals would not be sufficient for servers to work at busyness level

b_* . Let

$$p_* := b_*\mu/\lambda. \quad (3.8)$$

We next observe that if the p_* -admitted queue satisfies the non-idling condition (that is, the servers never idle when customers are waiting), the long-run average fraction of busy servers achieves b_* . The non-idling condition, together with (5)-(26) in Puha and Ward (2021) uniquely specifies \mathbb{P}_y^N for each $y \in \mathbb{S}^N = \left\{ y^N \in \mathbb{Y}^D : N - \langle 1, \nu^N \rangle = (N - x^N)^+ \text{ and } x^N \leq \langle 1, \eta^N \rangle \right\}$ and satisfies (3.2). Moreover, for any compatible initial distribution, the state process that satisfies the non-idling condition is a Feller, strong Markov process (see Proposition 4.2 in Kang et al. (2012)). Thus, for any $p \in (0, 1]$, the non-idling policy (the control policy that obeys the non-idling condition) is an admissible HL control policy for E_p^N , and thus is in Π_p^N .

Definition 13 (The Proposed Policy). *For each $N \in \mathbb{N}$, let π_*^N be the non-idling policy in $\Pi_{p_*}^N$, where p_* is given by (3.8).*

3.6 Asymptotic Optimality of π_*^N

In this section, we state our main results concerning asymptotic optimality of $\{\pi_*^N\}_{N \in \mathbb{N}}$ under fluid scaling.

Theorem 8 (Convergence under the Proposed Policy). *Suppose that Assumption 6 holds and that h^s is non-increasing when $b_* = 1$. Then the sequence $\{\pi_*^N\}_{N \in \mathbb{N}}$ satisfies*

$$\lim_{N \rightarrow \infty} \mathcal{C}^N(\pi_*^N) = a(\lambda - b_*\mu) + g_U(b_*).$$

Let $\hat{\Pi}^N := \cup_{p \in (0, 1]} \Pi_p^N$ denote the enlarged policy class, and given $\hat{\pi}^N \in \hat{\Pi}^N$, let $\hat{p}^N \in (0, 1]$ denote the associated admission control parameter.

Theorem 9 (Asymptotic Lower Bound). *Suppose that Assumption 6 holds, $\hat{\pi}^N \in \hat{\Pi}^N$ for each $N \in \mathbb{N}$ and the sequence $\{\hat{p}^N\}_{N \in \mathbb{N}}$ satisfies $\lim_{N \rightarrow \infty} \hat{p}^N = p$ for some $p \in (0, 1]$. Then,*

$$\liminf_{N \rightarrow \infty} \mathcal{C}^N(\hat{\pi}^N) \geq a(\lambda - b_*\mu) + g_U(b_*).$$

Remark 11. *The condition that $\lim_{N \rightarrow \infty} \hat{p}^N = p$ for some $p \in (0, 1]$ implies that $\{\hat{p}^N \bar{\lambda}^N\}_{N \in \mathbb{N}}$ satisfies $\lim_{N \rightarrow \infty} \hat{p}^N \bar{\lambda}^N = p\lambda$.*

Theorem 8 establishes that the solution to the fluid control problem (3.4) is achieved in the limiting system, when, for each N , the N -server system operates under π_*^N in Definition 13, and in case $b_* = 1$, h^s is non-increasing. Theorem 9 establishes that the fluid control problem (3.4) is an asymptotic lower bound for the objective (3.6). As a consequence, we conclude that the proposed sequence of policies $\{\pi_*^N\}_{N \in \mathbb{N}}$ is asymptotically optimal.

The proof of Theorem 8 given in Section 3.8 is facilitated by the fact that, for each $N \in \mathbb{N}$, under π_*^N the p_* -admitted N -server queue is non-idling, and thus, we can appeal to results in Kang et al. (2012); Atar et al. (2021) to establish the weak convergence of the sequence of fluid-scaled stationary distributions. The additional condition that h^s is non-increasing when $b_* = 1$, is needed for this in order to apply part (3) of Theorem 3.2 in Atar et al. (2021) in that case. This implies that the limit is the unique invariant state with zero queue mass.

The proof of Theorem 9 in Section 3.8 requires first adapting one of the arguments in Kang et al. (2012) (wherein the non-idling condition is assumed throughout) to show that a sequence of fluid-scaled stationary distributions is tight, and second arguing that the fluid control problem (3.7) provides an asymptotic lower bound on the cost along any convergent subsequence.

In the next section, we establish some preliminary results for stationary distributions (for both the stochastic N -server queue model and the fluid model) that help to prove Theorems 8 and 9, which may also be of independent interest. The proofs of Theorems 8 and 9 will be provided in Section 3.8.

3.7 Preliminary Results

In order to prove our main results (Theorems 8 and 9), we begin by establishing two foundational results concerning stationary distributions for the N -server queue. Then, we provide a fluid limit theorem, which shows that the distributional limit points of stationary distributions are fluid model solutions almost surely under suitable asymptotic conditions. Finally, we show some properties of stationary fluid model solutions for γ . The proofs are delayed to Section C.2 in the appendix to this chapter.

3.7.1 Stationary Distributions of the N -Server Queue

The following lemmas confirm the existence of a stationary distribution under any admissible HL control policy for E_p^N and $p \in (0, 1]$, and derive an expression for the long-run average cost. We denote by Δ_∞^N a stationary process associated with the process Δ^N , for $\Delta^N = E^N, Y^N, X^N, \nu^N, \eta^N, K^N, D^N, R^N, B^N, Q^N, I^N$.

Lemma 12. *Let $p \in (0, 1]$. For any $\pi_p^N \in \Pi_p^N$, there exists a compatible initial distribution ξ^N such that the state process Y_∞^N for (π_p^N, ξ^N) is stationary. Moreover, $\mathbb{E}_\xi^N [\langle 1, \eta_\infty^N(t) \rangle] = p\lambda^N\theta^{-1} < \infty$, for all $t \geq 0$.*

Remark 12. *Proposition 14 in Section 3.2 follows by setting $p = 1$.*

Given $p \in (0, 1]$, $\pi_p^N \in \Pi_p^N$ and a compatible initial distribution ς^N , let

$$\chi^N(t) := \inf\{x \geq 0 : \langle 1_{[0,x]}, \eta^N(t) \rangle \geq Q^N(t)\} \quad (3.9)$$

represent the waiting time of the HL customer at time t for each $t \geq 0$. Then, for $t \geq 0$,

$$Q^N(t) = \langle 1_{[0,\chi^N(t)]}, \eta^N(t) \rangle. \quad (3.10)$$

The associated stationary process is denoted by χ_∞^N .

Lemma 13. Let $p \in (0, 1]$. For any $\pi_p^N \in \Pi_p^N$ and compatible initial distribution ς^N , there exists $\xi^N \in \mathcal{S}(\pi_p^N)$ such that

$$\limsup_{T \rightarrow \infty} \mathbb{E}_\varsigma^N \left[\frac{\bar{R}^N(\pi_p^N, T)}{T} \right] = \mathbb{E}_\xi^N \left[\left\langle 1_{[0, \chi_\infty^N(0)]} h^r, \bar{\eta}_\infty^N(0) \right\rangle \right], \quad (3.11)$$

and

$$\limsup_{T \rightarrow \infty} \mathbb{E}_\varsigma^N \left[\frac{1}{T} \int_0^T g_U \left(\bar{B}^N(\pi_p^N, t) \right) dt \right] = \mathbb{E}_\xi^N \left[g_U \left(\bar{B}_\infty^N(0) \right) \right]. \quad (3.12)$$

If $\varsigma^N \in \mathcal{S}(\pi_p^N)$, then $\xi^N = \varsigma^N$.

In light of (3.10), one can interpret the right-hand side of (3.11) as an expected stationary reneging rate for the N -server queue.

Remark 13. For any $p \in (0, 1]$, $\pi_p^N \in \Pi_p^N$ and compatible initial distribution ς^N , there exists $\xi^N \in \mathcal{S}(\pi_p^N)$ such that

$$\mathcal{C}_\varsigma^N(\pi_p^N) = \mathbb{E}_\xi^N \left[a(1-p)\bar{\lambda}^N + a \left\langle 1_{[0, \chi_\infty^N(0)]} h^r, \bar{\eta}_\infty^N(0) \right\rangle + g_U \left(\bar{B}_\infty^N(0) \right) \right].$$

Proposition 15 in Section 3.2 follows by setting $p = 1$.

3.7.2 A Fluid Limit Theorem

Here we provide asymptotic assumptions under which it is shown in Puha and Ward (2021) that fluid limit points are almost surely fluid model solutions. Such a result is crucial for the proof of Theorem 9, which will appear in Section 3.8.

Assumption 7. Suppose for each $N \in \mathbb{N}$, $p^N \in (0, 1]$, $\pi_{p^N}^N \in \Pi_{p^N}^N$ for $E_{p^N}^N$ and ς^N is a compatible initial distribution for $\pi_{p^N}^N$. Assume that $\lim_{N \rightarrow \infty} p^N = p$ and $(\bar{X}^N(0), \bar{\nu}^N(0), \bar{\eta}^N(0)) \Rightarrow (X^0, \nu^0, \eta^0)$, as $N \rightarrow \infty$, for some random variable (X^0, ν^0, η^0) taking values in \mathbb{X} such that $\sup_{N \in \mathbb{N}} \mathbb{E}_\varsigma^N \left[\left\langle 1, \bar{\eta}^N(0) \right\rangle \right] < \infty$.

Remark 14. Under Assumptions 6 and 7 and the conditions on $E_{p^N}^N$, K^N , G^s , g^s , h^s , G^r , g^r , and h^r specified in Sections 3.2 and 3.5, one can without loss of generality assume that the convergence of the initial condition in Assumption 7 is almost sure and then check that Assumptions 1, 2, 3(1), 3(3), 3(4), 4, 5(1) and 5(3) in Puha and Ward (2021) hold, i.e., Assumptions 3(2), 3(5) and 5(2) may not hold.

Lemma 14. Suppose Assumptions 6 and 7 hold. Then, $\bar{\eta}^N \Rightarrow \eta$, as $N \rightarrow \infty$, where $\eta(0) \stackrel{d}{=} \eta^0$ and η satisfies (C.15) almost surely for $E(t) = p\lambda t$, $t \geq 0$.

In fact, Assumptions 3(2) and 3(5) in Puha and Ward (2021) can be replaced by the condition $\sup_{N \in \mathbb{N}} \mathbb{E}_\xi^N [\langle 1, \bar{\eta}^N(0) \rangle] < \infty$ and Assumption 5(2) (η^0 has no atoms) is used to establish convergence of the scaled reneging processes to the expression in (C.8). Thus, the result in Theorem 1 in Puha and Ward (2021) continues to hold. We obtain the following slightly restated version of Theorem 1 in Puha and Ward (2021).

Lemma 15 (Theorem 1 in Puha and Ward (2021)). *Suppose that $\{(\pi^N, \varsigma^N)\}_{N \in \mathbb{N}}$ is such that Assumptions 6 and 7 hold, η^0 has no atoms, and (X, ν, η) is a distributional limit point of $\{(\bar{X}^N, \bar{\nu}^N, \bar{\eta}^N)\}_{N \in \mathbb{N}}$. Then $(X(0), \nu(0), \eta(0)) \stackrel{d}{=} (X^0, \nu^0, \eta^0)$ and (X, ν, η) is almost surely a fluid model solution for $p\lambda$.*

3.7.3 Properties of Stationary Fluid Model Solutions

Fix $\gamma > 0$. Here we consider the fluid model for γ with random initial states such that the resulting fluid model solution is a stationary process. Lemmas 16 and 17 below, provide properties of such solutions. The proof of Theorem 9 relies on Lemmas 16 and 17.

In what follows, we fix a fluid model solution $Z_\infty = (X_\infty, \nu_\infty, \eta_\infty)$ for γ such that Z_∞ is a stationary process. We denote the law of $Z_\infty(0)$ by ξ and the expectation operator by \mathbb{E}_ξ . In addition, we define a Borel probability measure η_e satisfying $d\eta_e(x) = \theta \bar{G}^r(x)dx$ for all $x \in \mathbb{R}_+$, where the subscript e is mnemonic for excess life distribution.

Lemma 16. For all $t \geq 0$, $\eta_\infty(t) = \gamma\theta^{-1}\eta_e$. In particular, for all $t \geq 0$, $\eta_\infty(t)$ has no atoms, $x \mapsto \langle 1_{[0,x]}, \eta_\infty(t) \rangle$ is a continuous strictly increasing function on \mathbb{R}_+ , and $\langle 1, \eta_\infty(t) \rangle = \gamma\theta^{-1}$.

Lemma 17. There exists $b \in [0, \min\{1, \gamma/\mu\}]$ such that for all $t \geq 0$, $\mathbb{E}_\xi[B_\infty(t)] = b$ and $\mathbb{E}_\xi[\langle 1_{[0,\chi_\infty(t)]} h^r, \eta_\infty(t) \rangle] = \gamma - b\mu$.

3.8 Proofs of Main Results (Theorems 8 and 9)

Proof of Theorem 8. For each $N \in \mathbb{N}$, let $\xi^N \in \mathcal{S}(\pi_*^N)$ which exists by Lemma 12, and recall that $Y_\infty^N(0)$ has distribution ξ^N . Consider the sequence $\{(\bar{X}_\infty^N(0), \bar{\nu}_\infty^N(0), \bar{\eta}_\infty^N(0))\}_{N \in \mathbb{N}}$. We wish to show that $\lim_{N \rightarrow \infty} \mathcal{C}(\pi_*^N) = a(\lambda - b_*\mu) + g_U(b_*)$. By Lemma 13, it suffices to show

$$\begin{aligned} & \lim_{N \rightarrow \infty} \mathbb{E}_\xi^N \left[a(1 - p_*)\bar{\lambda}^N + a \left\langle 1_{[0,\chi_\infty^N(0)]} h^r, \bar{\eta}_\infty^N(0) \right\rangle + g_U(\bar{B}_\infty^N(0)) \right] \\ &= a(\lambda - b_*\mu) + g_U(b_*). \end{aligned} \quad (3.13)$$

Note that $p_*\bar{\lambda}^N \rightarrow p_*\lambda$, as $N \rightarrow \infty$ (from Assumption 6). This, together with the assumptions on E^N (which $E_{p_*}^N$ inherits), G^s , g^s , h^s , G^r , g^r , and h^r given in Section 3.2, implies that Assumptions 3.1-3.5 in Kang et al. (2012) hold for $\{(E_{p_*}^N, \pi_*^N, \xi^N)\}_{N \in \mathbb{N}}$. In addition, since it is assumed that h^s is non-increasing when $b_* = 1$, the result in Theorem 3.3 in Kang et al. (2012) holds², which establishes

$$(\bar{X}_\infty^N(0), \bar{\nu}_\infty^N(0), \bar{\eta}_\infty^N(0)) \Rightarrow (b_*, b_*\nu_e, p_*\lambda\theta^{-1}\eta_e), \quad (3.14)$$

2. There is a gap in the original proof of Theorem 3.3 in Kang et al. (2012), where a stationary distribution for the fluid model is assumed to coincide with the invariant state, which is unique since G^r is strictly increasing. Under the conditions of Theorem 3.3 in Kang et al. (2012), Theorem 3.2(1) in Atar et al. (2021) implies that this is true when $b_* < 1$. With the added condition that h^s is non-increasing, Theorem 3.2(3) in Atar et al. (2021) implies that this is true when $b_* = 1$. Hence, the result in Theorem 3.3 in Kang et al. (2012) holds in the present setting. See the discussion in Atar et al. (2021) that follows the statement of Theorem 3.2 for a detailed explanation.

as $N \rightarrow \infty$, where $d\nu_e(x) = \mu\bar{G}^s(x)dx$ and $d\eta_e(x) = \theta\bar{G}^r(x)dx$ for each $x \in \mathbb{R}_+$. This, together with (C.5), (C.6), and $p_* = b_*\mu/\lambda$, gives that as $N \rightarrow \infty$,

$$\bar{B}_\infty^N(0) = \left\langle 1, \bar{\nu}_\infty^N(0) \right\rangle \Rightarrow \langle 1, b_*\nu_e \rangle = b_*, \quad (3.15)$$

$$\bar{Q}_\infty^N(0) = \bar{X}_\infty^N(0) - \bar{B}_\infty^N(0) \Rightarrow b_* - b_* = 0. \quad (3.16)$$

The function g_U is continuous. Hence, by (3.15) and the continuous mapping theorem,

$$g_U(\bar{B}_\infty^N(0)) \Rightarrow g_U(b_*), \text{ as } N \rightarrow \infty. \quad (3.17)$$

Then, since g_U is bounded, (3.17) and the bounded convergence theorem yield that

$$\lim_{N \rightarrow \infty} \mathbb{E}_\xi^N \left[g_U \left(\bar{B}_\infty^N(0) \right) \right] = g_U(b_*). \quad (3.18)$$

From (3.10) and (3.16),

$$\left\langle 1_{[0, \chi_\infty^N(0)]}, \bar{\eta}_\infty^N(0) \right\rangle = \bar{Q}_\infty^N(0) \Rightarrow 0, \text{ as } N \rightarrow \infty. \quad (3.19)$$

Note that for each $N \in \mathbb{N}$,

$$0 \leq a \left\langle 1_{[0, \chi_\infty^N(0)]} h^r, \bar{\eta}_\infty^N(0) \right\rangle \leq a \|h^r\|_\infty \bar{Q}_\infty^N(0),$$

which, together with (3.19) and boundedness of h^r , implies

$$a \left\langle 1_{[0, \chi_\infty^N(0)]} h^r, \bar{\eta}_\infty^N(0) \right\rangle \Rightarrow 0, \text{ as } N \rightarrow \infty. \quad (3.20)$$

By Lemma 12, $\lim_{N \rightarrow \infty} p_* \bar{\lambda}^N = p_* \lambda$, and (3.14),

$$\lim_{N \rightarrow \infty} \mathbb{E}_\xi^N \left[\left\langle 1, \bar{\eta}_\infty^N(0) \right\rangle \right] = \lim_{N \rightarrow \infty} p_* \bar{\lambda}^N \theta^{-1} = p_* \lambda \theta^{-1} = \left\langle 1, p_* \lambda \theta^{-1} \eta_e \right\rangle.$$

This together with (3.14) implies that $\left\{ \left\langle 1, \bar{\eta}_\infty^N(0) \right\rangle \right\}_{N \in \mathbb{N}}$ is uniformly integrable. Note that $\left\langle 1_{[0, \chi_\infty^N(0)]} h^r, \bar{\eta}_\infty^N(0) \right\rangle \leq \|h^r\|_\infty \left\langle 1, \bar{\eta}_\infty^N(0) \right\rangle$ for each $N \in \mathbb{N}$ and h^r is bounded. Thus, $\left\{ \left\langle 1_{[0, \chi_\infty^N(0)]} h^r, \bar{\eta}_\infty^N(0) \right\rangle \right\}_{N \in \mathbb{N}}$ is uniformly integrable. This together with (3.20) implies that

$$\lim_{N \rightarrow \infty} \mathbb{E}_\xi^N \left[a \left\langle 1_{[0, \chi_\infty^N(0)]} h^r, \bar{\eta}_\infty^N \right\rangle \right] = 0. \quad (3.21)$$

Finally, by Assumption 6, it follows that

$$\lim_{N \rightarrow \infty} a(1 - p_*) \bar{\lambda}^N = a(1 - p_*) \lambda = a(\lambda - b_* \mu). \quad (3.22)$$

Combining (3.18), (3.21) and (3.22) establishes (3.13), as desired. \square

Proof of Theorem 9. Fix a sequence $\{\hat{\pi}^N\}_{N \in \mathbb{N}}$ satisfying the conditions of Theorem 9. For each $N \in \mathbb{N}$, let $\xi^N \in \mathcal{S}(\hat{\pi}^N)$ be such that $\mathcal{C}^N(\hat{\pi}^N) = \mathcal{C}_\xi^N(\hat{\pi}^N)$ which exists by Lemma 12 and the definition of $\mathcal{C}^N(\hat{\pi}^N)$. For each $N \in \mathbb{N}$, let Y_∞^N be the state process for $(\hat{\pi}^N, \xi^N)$. It suffices to show that $\lim_{i \rightarrow \infty} \mathcal{C}_\xi^{N_i}(\hat{\pi}^{N_i}) \geq a(\lambda - b_* \mu) + g_U(b_*)$, for any convergent subsequence of cost functions $\left\{ \mathcal{C}_\xi^{N_i}(\hat{\pi}^{N_i}) \right\}_{i=1}^\infty$. Fix such a subsequence $\{N_i\}_{i=1}^\infty$. We consider the fluid scaled sequence $\{(\bar{X}_\infty^{N_i}, \bar{\nu}_\infty^{N_i}, \bar{\eta}_\infty^{N_i})\}_{i=1}^\infty$. By Lemma 13, it suffices to show

$$\begin{aligned} & \lim_{i \rightarrow \infty} \mathbb{E}_\xi^{N_i} \left[a \left(1 - \hat{p}^{N_i} \right) \bar{\lambda}^{N_i} + a \left\langle 1_{[0, \chi_\infty^{N_i}(0)]} h^r, \bar{\eta}_\infty^{N_i}(0) \right\rangle + g_U(\bar{B}_\infty^{N_i}(0)) \right] \\ & \geq a(\lambda - b_* \mu) + g_U(b_*). \end{aligned} \quad (3.23)$$

We begin by noting that the sequence $\{(\bar{X}_\infty^{N_i}(0), \bar{\nu}_\infty^{N_i}(0), \bar{\eta}_\infty^{N_i}(0))\}_{i=1}^\infty$ is tight. This follows

by Theorem 6.2 in Kang et al. (2012) and its proof since in the present setting, the result in Lemma 6.1 in Kang et al. (2012) holds, $\bar{K}_\infty^{N_i}(t) \leq \bar{E}_\infty^{N_i}(t) + \langle 1, \bar{\eta}_\infty^{N_i}(0) \rangle$ for all $i \in \mathbb{N}$ and $t \geq 0$, and $\bar{X}_\infty^{N_i}(0) \leq 1 + \langle 1, \bar{\eta}_\infty^{N_i}(0) \rangle$ for all $i \in \mathbb{N}$. Since $\{(\bar{X}_\infty^{N_i}(0), \bar{\nu}_\infty^{N_i}(0), \bar{\eta}_\infty^{N_i}(0))\}_{i=1}^\infty$ is tight, there exists a further subsequence $\{N_{i_k}\}_{k=1}^\infty$ such that

$$\left(\bar{X}_\infty^{N_{i_k}}(0), \bar{\nu}_\infty^{N_{i_k}}(0), \bar{\eta}_\infty^{N_{i_k}}(0) \right) \Rightarrow \left(X_\infty^0, \nu_\infty^0, \eta_\infty^0 \right), \quad (3.24)$$

as $k \rightarrow \infty$. Without loss of generality, we can replace $\{N_{i_k}\}_{k=1}^\infty$ with $\{N_i\}_{i=1}^\infty$ by eliminating some members if necessary. In what follows, we verify that (3.23) holds along this subsequence. For this, we will first show that

$$\begin{aligned} & \lim_{i \rightarrow \infty} \mathbb{E}_\xi^{N_i} \left[a \left(1 - \hat{p}^{N_i} \right) \bar{\lambda}^{N_i} + a \left\langle 1_{[0, \chi_\infty^{N_i}(0)]} h^r, \bar{\eta}_\infty^{N_i}(0) \right\rangle + g_U \left(\bar{B}_\infty^{N_i}(0) \right) \right] \\ &= a(1-p)\lambda + a\mathbb{E}_\xi \left[\left\langle 1_{[0, \chi_\infty^0]} h^r, \eta_\infty^0 \right\rangle \right] + \mathbb{E}_\xi \left[g_U(B_\infty^0) \right], \end{aligned} \quad (3.25)$$

where ξ denotes the distribution of $(X_\infty^0, \nu_\infty^0, \eta_\infty^0)$ and \mathbb{E}_ξ is the expectation operator for ξ . Then we will establish process level convergence to a stationary fluid model solution for $p\lambda$ in order to apply Lemma 17 to the right-hand side of (3.25).

We begin by showing that η_∞^0 has no atoms and that $\left\{ \left\langle 1_{[0, \chi_\infty^{N_i}(0)]} h^r, \bar{\eta}_\infty^{N_i}(0) \right\rangle \right\}_{i=1}^\infty$ is uniformly integrable. By Lemma 12, Assumption 6 and $\lim_{i \rightarrow \infty} \hat{p}^{N_i} = p$,

$$\lim_{i \rightarrow \infty} \mathbb{E}_\xi^{N_i} \left[\left\langle 1, \bar{\eta}_\infty^{N_i}(0) \right\rangle \right] = p\lambda\theta^{-1}, \quad (3.26)$$

so $\sup_{i \in \mathbb{N}} \mathbb{E}_\xi^{N_i} \left[\left\langle 1, \bar{\eta}_\infty^{N_i}(0) \right\rangle \right] < \infty$. This together with (3.24) implies that Assumption 7 holds with $\{E_{\hat{p}^{N_i}}\}_{i=1}^\infty$, $\{\hat{\pi}^{N_i}\}_{i=1}^\infty$ and $\{\xi^{N_i}\}_{i=1}^\infty$ replacing $\{E_{p^N}\}_{N \in \mathbb{N}}$, $\{\pi_{p^N}^N\}_{N \in \mathbb{N}}$ and $\{\varsigma^N\}_{N \in \mathbb{N}}$ respectively. Thus, by Lemma 14, $\bar{\eta}_\infty^{N_i} \Rightarrow \eta_\infty$, as $i \rightarrow \infty$, where $\eta_\infty(0) \stackrel{d}{=} \eta_\infty^0$ and η_∞ satisfies (C.15) almost surely for $E_\infty(t) = p\lambda t$, $t \geq 0$. Moreover, since $\bar{\eta}_\infty^{N_i}$ is a stationary process for each $i \in \mathbb{N}$, η_∞ is a stationary process such that $\eta_\infty(t) \stackrel{d}{=} \eta_\infty^0$ for all

$t \geq 0$. Hence, by Lemma 16, $\eta_\infty^0 = p\lambda\theta^{-1}\eta_e$, so that η_∞^0 has no atoms, $\langle 1, \eta_\infty^0 \rangle = p\lambda\theta^{-1}$ and $x \mapsto \langle 1_{[0,x]}, \eta_\infty^0 \rangle$ is a continuous, strictly increasing function on \mathbb{R}_+ . Then recalling (3.26), $\lim_{i \rightarrow \infty} \mathbb{E}_\xi^{N_i} [\langle 1, \bar{\eta}_\infty^{N_i}(0) \rangle] = \langle 1, \eta_\infty^0 \rangle$. This together with (3.24) implies that $\{\langle 1, \bar{\eta}_\infty^{N_i}(0) \rangle\}_{i=1}^\infty$ is uniformly integrable. Since h^r is bounded and $\left\langle 1_{[0, \chi_\infty^{N_i}(0)]} h^r, \bar{\eta}_\infty^{N_i}(0) \right\rangle \leq \|h^r\|_\infty \langle 1, \bar{\eta}_\infty^{N_i}(0) \rangle$ for each $i \in \mathbb{N}$, uniform integrability of $\left\{ \left\langle 1_{[0, \chi_\infty^{N_i}(0)]} h^r, \bar{\eta}_\infty^{N_i}(0) \right\rangle \right\}_{i=1}^\infty$ follows.

Next we show (3.25). For this without loss of generality, we assume that the convergence in (3.24) is almost sure which we abbreviate as a.s. By (3.24), we have

$$\begin{aligned} \lim_{i \rightarrow \infty} \bar{B}_\infty^{N_i}(0) &= \lim_{i \rightarrow \infty} \langle 1, \bar{\nu}_\infty^{N_i}(0) \rangle = \langle 1, \nu_\infty^0 \rangle = B_\infty^0, \quad \text{a.s., and} \\ \lim_{i \rightarrow \infty} \bar{Q}_\infty^{N_i}(0) &= \lim_{i \rightarrow \infty} \bar{X}_\infty^{N_i}(0) - \lim_{i \rightarrow \infty} \bar{B}_\infty^{N_i}(0) = X_\infty^0 - B_\infty^0 = Q_\infty^0, \quad \text{a.s.} \end{aligned} \quad (3.27)$$

This implies that

$$\lim_{i \rightarrow \infty} \left\langle 1_{[0, \chi_\infty^{N_i}(0)]}, \bar{\eta}_\infty^{N_i}(0) \right\rangle = \lim_{i \rightarrow \infty} \bar{Q}_\infty^{N_i}(0) = Q_\infty^0 = \left\langle 1_{[0, \chi_\infty^0]}, \eta_\infty^0 \right\rangle, \quad \text{a.s.}$$

Thus, since $x \mapsto \langle 1_{[0,x]}, \eta_\infty^0 \rangle$ is a continuous strictly increasing function on \mathbb{R}_+ , $\lim_{i \rightarrow \infty} \chi_\infty^{N_i}(0) = \chi_\infty^0$ a.s. This together with the above display and that h^r is continuous and bounded implies

$$\lim_{i \rightarrow \infty} \left\langle 1_{[0, \chi_\infty^{N_i}(0)]} h^r, \bar{\eta}_\infty^{N_i}(0) \right\rangle = \left\langle 1_{[0, \chi_\infty^0]} h^r, \eta_\infty^0 \right\rangle, \quad \text{a.s.} \quad (3.28)$$

Now, as in the proof of Theorem 8, (3.25) follows from $\lim_{i \rightarrow \infty} \hat{p}^{N_i} \bar{\lambda}^{N_i} = p\lambda$, (3.27) and g_U is bounded and continuous, and (3.28) and the uniform integrability of $\left\{ \left\langle 1_{[0, \chi_\infty^{N_i}(0)]} h^r, \bar{\eta}_\infty^{N_i}(0) \right\rangle \right\}_{i=1}^\infty$.

Finally, we argue process level convergence to a stationary fluid model solution for $p\lambda$. Since Assumption 7 holds for $\{N_i\}_{i=1}^\infty$ (as noted above) and η_∞^0 has no atoms (also noted above), Lemma 15 implies that $(\bar{X}_\infty^{N_i}, \bar{\nu}_\infty^{N_i}, \bar{\eta}_\infty^{N_i}) \Rightarrow (X_\infty, \nu_\infty, \eta_\infty)$, as $i \rightarrow \infty$, where

$(X_\infty, \nu_\infty, \eta_\infty)$ is almost surely a fluid model solution for $p\lambda$ such that $(X_\infty(0), \nu_\infty(0), \eta_\infty(0)) \stackrel{d}{=} (X_\infty^0, \nu_\infty^0, \eta_\infty^0)$. Moreover, $(X_\infty, \nu_\infty, \eta_\infty)$ is a stationary fluid model solution for $p\lambda$ by the stationarity of $(\bar{X}_\infty^{N_i}, \bar{\nu}_\infty^{N_i}, \bar{\eta}_\infty^{N_i})$ for each $i \in \mathbb{N}$. Then, from Lemma 17, there exists $b \in [0, \min\{1, p\lambda/\mu\}]$ such that $\mathbb{E}_\xi[B_\infty^0] = \mathbb{E}_\xi[B_\infty(0)] = b$. Since g_U is convex, Jensen's inequality further implies that

$$\mathbb{E}_\xi \left[g_U(B_\infty^0) \right] \geq g_U \left(\mathbb{E}_\xi \left[B_\infty^0 \right] \right) = g_U(b).$$

This together with (3.25) and the second part of Lemma 17 gives

$$\begin{aligned} & \lim_{i \rightarrow \infty} \mathbb{E}_\xi^{N_i} \left[a(1 - \hat{p}^{N_i}) \bar{\lambda}^{N_i} + a \left\langle 1_{[0, \chi_\infty^{N_i}(0)]} h^r, \bar{\eta}_\infty^{N_i}(0) \right\rangle + g_U(\bar{B}_\infty^{N_i}(0)) \right] \\ & \geq a(1 - p)\lambda + a(p\lambda - b\mu) + g_U(b) \\ & = a(\lambda - b\mu) + g_U(b) \geq a(\lambda - b_*\mu) + g_U(b_*) , \end{aligned} \tag{3.29}$$

which completes the proof that (3.23) holds, as desired. \square

3.9 Extension: Holding Costs

In this section, we include holding costs in the objective function to penalize congestion, and we show similar results as in the case with abandonment cost only, using the enlarged admissible policy class with admission control.

Let c be the holding cost incurred per customer per unit time. Then, given $\pi^N \in \Pi^N$ (Definition 10) and a compatible initial distribution ς^N , the long-run average cost of (π^N, ς^N)

is modified as

$$\begin{aligned}\mathcal{H}_\varsigma^N(\pi^N) := \limsup_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}_\varsigma^N & \left[c \int_0^T \bar{Q}^N(\pi^N, t) dt + a \bar{R}^N(\pi^N, T) \right. \\ & \left. + \int_0^T g_U(\bar{B}^N(\pi^N, t)) dt \right],\end{aligned}$$

and the worst case cost under π^N is

$$\mathcal{H}^N(\pi^N) := \sup_{\xi^N \in \mathcal{S}(\pi^N)} \mathcal{H}_\xi^N(\pi^N).$$

Then, the objective is to determine π_{opt}^N and $\mathcal{H}^N(\pi_{\text{opt}}^N)$ such that

$$\mathcal{H}^N(\pi_{\text{opt}}^N) := \inf_{\pi^N \in \Pi^N} \mathcal{H}^N(\pi^N). \quad (3.30)$$

We begin by noting that the fluid model equations are not changed, and hence the fluid control problem can be obtained based on the unchanged fluid model invariant states. Also, the weak convergence result (Lemma 15) continues to hold because the proof of Lemma 15 does not rely on the objective function.

Assumption 8. *The function h^r is non-increasing.*

Assumption 8 is crucial to prove the main asymptotic optimality result, Theorem 9, when holding costs are considered. To explain this point, for $p \in (0, 1]$ and $b \in [0, \min\{1, p\lambda/\mu\}]$, define

$$q(b, p) := p\lambda \int_0^{(G^r)^{-1}(1-b\mu/p\lambda)} (1 - G^r(x)) dx. \quad (3.31)$$

From Equation (54) in Puha and Ward (2021), $q(b, p)$ represents the invariant fluid queue length for $p \in (0, 1]$ and $b \in [0, \min\{1, p\lambda/\mu\}]$ when the arrival rate is thinned to $p\lambda$. As

a consequence of Assumption 8, $q(\cdot, p)$ is a convex function on $[0, \min\{1, p\lambda/\mu\}]$ for each $p \in (0, 1]$ (see Remark 10 in Puha and Ward (2019)). This convexity plays an integral role in our analysis. The modified fluid control problem is

$$\min_{b \in [0, \min\{1, \lambda/\mu\}]} cq(b, 1) + a(\lambda - b\mu) + g_U(b). \quad (3.32)$$

Different from (3.4) when the holding costs were not included, the optimization problem (3.32) becomes sensitive to the patience distribution, because the fluid queue length depends on the patience distribution; see the right-hand side of (3.31). We denote the solution to (3.32) by $b_{*,hc}$ (which is unique and guaranteed to exist because $q(b, 1)$ is a convex function). As in Section 3.4, if $b_{*,hc} < \min\{1, \lambda/\mu\}$, then we expect an idling control policy to be optimal for (3.30). In particular, servers can be allowed to take a rest for $(1 - b_{*,hc})(b_{*,hc}\mu)^{-1}$ time units after each service completion.

As in Sections 3.5 and 3.6, in order to show the asymptotic optimality property, we work with the enlarged admissible policy class $\hat{\Pi}^N$, which incorporates the potential of admission control. The unit abandonment and holding cost for the admitted arrivals remain a and c . For every rejected arrival, in addition to a cost of a (as in (3.7)), we need to further account for the holding cost that would have been incurred if under a control policy in Π^N without admission control that may idle. Suppose that $\tilde{\mathcal{H}}(b, p)$ is the overall holding costs for the rejected arrivals for $p \in (0, 1]$ and $b \in [0, \min\{1, p\lambda/\mu\}]$, and we call it the fluid-scaled holding cost compensator.

Definition 14 (The Fluid-Scaled Holding Cost Compensator). *Given λ is fixed, $p \in (0, 1]$ and $b \in [0, \min\{1, p\lambda/\mu\}]$, $\tilde{\mathcal{H}}(b, p)$ is given by*

$$\tilde{\mathcal{H}}(b, p) = c(q(b, 1) - q(b, p)).$$

Then, the modified fluid control problem under $\hat{\Pi}^N$ is

$$\begin{aligned} & \min_{p \in (0,1], b \in [0, \min\{1, p\lambda/\mu\}]} cq(b, p) + \tilde{\mathcal{H}}(b, p) + a(p\lambda - b\mu) \\ & \quad + a(1-p)\lambda + g_U(b) \\ &= \min_{b \in [0, \min\{1, p\lambda/\mu\}]} cq(b, 1) + a(\lambda - b\mu) + g_U(b). \end{aligned} \quad (3.33)$$

Under Definition 14, it is clear that the solution to (3.33) is identical to the solution to (3.32). Then, from (3.33), given $p \in (0, 1]$, $\pi_p^N \in \Pi_p^N$ (Definition 12) and a compatible initial distribution ς^N , the modified objective function of (π_p^N, ς^N) is given by

$$\begin{aligned} \mathcal{H}_\varsigma^N(\pi_p^N) &= \limsup_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}_\varsigma^N \left[c \int_0^T \bar{Q}^N(\pi_p^N, t) dt \right. \\ &\quad + \int_0^T \tilde{\mathcal{H}}(\bar{B}^N(\pi_p^N, t), p) dt + a \left(\bar{E}^N(T) - \bar{E}_p^N(T) + \bar{R}^N(\pi_p^N, T) \right) \\ &\quad \left. + \int_0^T g_U(\bar{B}^N(\pi_p^N, t)) dt \right]. \end{aligned} \quad (3.34)$$

Noting that the fluid solution to (3.32), $b_{*,hc}$, is independent of p , we can consider the same proposed policy as defined in Definition 13 with b_* replaced by $b_{*,hc}$; denote it by $\pi_{*,hc}^N$. In the remainder of the section, we outline the proof of asymptotic optimality for $\{\pi_{*,hc}^N\}_{N \in \mathbb{N}}$.

Lemma 12 continues to hold, because its proof does not rely on the objective function. For the objective (3.32) with linear holding cost penalties, Lemma 13 can be modified such that for any admissible HL control policy $\pi_p^N \in \Pi_p^N$, the following holds for a stationary distribution $\xi^N \in \mathcal{S}(\pi_p^N)$:

$$\begin{aligned} \mathcal{H}_\xi^N(\pi_p^N) &= \mathbb{E}_\xi^N \left[c \bar{Q}_\infty^N(0) + \tilde{\mathcal{H}}(\bar{B}_\infty^N(0), p) \right. \\ &\quad \left. + a(1-p)\bar{\lambda}^N + a \left\langle 1_{[0, \chi_\infty^N(0)]} h^r, \bar{\eta}_\infty^N(0) \right\rangle + g_U(\bar{B}_\infty^N(0)) \right]. \end{aligned}$$

This is because for each $N \in \mathbb{N}$, if $\{\tau(n)\}_{n \in \mathbb{N}} \subset \mathbb{R}_+$ is a strictly increasing subsequence along which $L_{\tau(n)}^N$ converges to ξ^N , then $Q_{\tau(n)}^N(\pi_p^N, 0) \Rightarrow Q_\infty^N(0)$ as $n \rightarrow \infty$, and

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{\tau(n)} \mathbb{E}_\xi^N \left[\int_0^{\tau(n)} Q^N(\pi_p^N, t) dt \right] &= \lim_{n \rightarrow \infty} \mathbb{E}_{L_{\tau(n)}}^N [Q_{\tau(n)}^N(0)] \\ &= \mathbb{E}_\xi^N [Q_\infty^N(0)], \end{aligned}$$

because $Q^N(t) \leq \langle 1, \eta^N(t) \rangle$ for all $t \geq 0$, and $\{\langle 1, \eta_{\tau(n)}^N(0) \rangle\}_{n \in \mathbb{N}}$ is uniformly integrable as in the proof of Lemma 13 in Section C.2 in the appendix to this chapter (see the paragraph immediately above Claim 20).

Thus, following a similar proof to that of Theorem 8, the sequence $\{\pi_{*,hc}^N\}_{N \in \mathbb{N}}$ satisfies

$$\begin{aligned} &\lim_{N \rightarrow \infty} \mathcal{H}^N(\pi_{*,hc}^N) \\ &= cq(b_{*,hc}, p) + \tilde{\mathcal{H}}(b_{*,hc}, p) + a(\lambda - b_{*,hc}\mu) + g_U(b_{*,hc}) \\ &= cq(b_{*,hc}, 1) + a(\lambda - b_{*,hc}\mu) + g_U(b_{*,hc}). \end{aligned}$$

We wish to show that for $\{\hat{\pi}^N\}_{N \in \mathbb{N}}$ that satisfy the conditions in the statement of Theorem 9,

$$\liminf_{N \rightarrow \infty} \mathcal{H}^N(\hat{\pi}^N) \geq cq(b_{*,hc}, 1) + a(\lambda - b_{*,hc}\mu) + g_U(b_{*,hc}).$$

Let $\{(\bar{X}_\infty^{Ni}, \bar{\nu}_\infty^{Ni}, \bar{\eta}_\infty^{Ni})\}_{N \in \mathbb{N}}$ and $(X_\infty, \nu_\infty, \eta_\infty)$ be as in the proof of Theorem 9; that is,

$$(\bar{X}_\infty^{Ni}, \bar{\nu}_\infty^{Ni}, \bar{\eta}_\infty^{Ni}) \Rightarrow (X_\infty, \nu_\infty, \eta_\infty), \text{ as } i \rightarrow \infty,$$

with $(X_\infty, \nu_\infty, \eta_\infty)$ being almost surely a stationary fluid model solution for $p\lambda$. Note that,

by definition,

$$\begin{aligned}
& cQ_\infty(0) + \tilde{\mathcal{H}}(p, B_\infty(0)) \\
&= cq(B_\infty(0), p) + c(q(B_\infty(0), 1) - q(B_\infty(0), p)) \\
&= cq(B_\infty(0), 1).
\end{aligned}$$

Since $q(\cdot, 1)$ is convex on $[0, \min\{1, \lambda/\mu\}]$, Jensen's inequality together with Lemma 17 implies that

$$\begin{aligned}
& \mathbb{E}_\xi[cQ_\infty(0) + \tilde{\mathcal{H}}(p, B_\infty(0))] = \mathbb{E}_\xi[cq(B_\infty(0), 1)] \\
& \geq cq(\mathbb{E}_\xi[B_\infty(0)], 1) = cq(\mathbb{E}_\xi[B_\infty^0], 1) = cq(b, 1). \tag{3.35}
\end{aligned}$$

Then, as in the proof of Theorem 9,

$$\begin{aligned}
& \lim_{i \rightarrow \infty} \mathbb{E}_\xi^{N_i} \left[c\bar{Q}_\infty^{N_i}(0) + \tilde{\mathcal{H}}(\bar{B}_\infty^{N_i}(0), p) \right. \\
& \quad \left. + a(1 - \hat{p}^{N_i})\bar{\lambda}^{N_i} + a \left\langle 1_{[0, \chi_\infty^{N_i}(0)]} h^r, \bar{\eta}_\infty^{N_i}(0) \right\rangle + g_U(\bar{B}_\infty^{N_i}(0)) \right] \\
&= a(1 - p)\lambda + \mathbb{E}_\xi \left[cQ_\infty(0) + \tilde{\mathcal{H}}(B_\infty(0), p) \right. \\
& \quad \left. + a \left\langle 1_{[0, \chi_\infty(0)]} h^r, \eta_\infty(0) \right\rangle + g_U(B_\infty(0)) \right] \\
&\geq cq(b, 1) + a(\lambda - b\mu) + g_U(b) \\
&\geq cq(b_{*,hc}, 1) + a(\lambda - b_{*,hc}\mu) + g_U(b_{*,hc}),
\end{aligned}$$

where the first inequality follows from the first inequality in (3.29) in the proof of Theorem 9, and (3.35).

Hence, the proposed policy $\pi_{*,hc}^N$ with $p_{*,hc} = b_{*,hc}\mu/\lambda$ is asymptotically optimal, under the generalized objective function involving holding costs.

Remark 15. All the results in this section continue to stand if we consider a non-decreasing,

continuous, convex holding cost that maintains uniform integrability of the sequence of holding costs rather than a linear holding cost. In that case, the corrected fluid-scaled holding cost compensator defined in Definition 14 is the difference of the holding cost function evaluated at $q(b, 1)$ and at $q(b, p)$.

3.10 Conclusion

In this chapter, we examine the intricate trade-off between long-term average operational costs (specifically abandonment costs and, as an extension, holding costs) and server utilization costs in the $GI/GI/N+GI$ queue. Due to the difficulty of exact analysis, we resort to large-system fluid approximation in a many-server asymptotic regime, where the arrival rate and number of servers become large. The solution to the fluid control problem suggests a compelling insight: non-idling service disciplines are not in general optimal, when server utilization costs are take into consideration. To ensure sufficient idle time for human servers, our proposed solution comes in the form of an admission control policy, which is shown to be asymptotically optimal. This policy balances the imperative of enhancing operational efficiency with the preservation of the well-being of human servers, which carries profound implications for the sustainable long-term growth of service operations.

An intriguing avenue for future exploration centers on the practical feasibility and ethical considerations of the proposed policy. While our analysis provides a strong theoretical foundation, the real-world implementation of admission control can raise questions about fairness, necessitating a deeper examination of advanced mechanisms to screen heterogeneous customers, such as a coupon system. This research direction not only aligns with the evolving landscape of service industry practices but also underscores our commitment to promoting fair and responsible service management strategies.

CHAPTER 4

CONCLUSION

In this thesis, we embark on a journey to explore the intricate world of modern service systems, where the interplay of human server behavior, uncertain system characteristics, and managerial decisions presents a multifaceted challenge. Our primary goal is to bridge the gap between classical queueing theory and the complexities inherent in today's service environments. We integrate human server behavior and statistical learning into classical many-server queueing models, ultimately providing a framework for the analysis and optimization of behavior-aware and prediction-driven modern service systems.

Chapter 1 delves into the realm of strategic human server behavior. We develop a game-theoretic model to unveil how managerial decisions impact human server work speed. Contrary to conventional wisdom, we find that reducing workload does not always decrease customer waiting times. Chapter 2 navigates the nontrivial interactions between statistical learning and queueing dynamics in complex service systems. We develop a policy that can efficiently learn the unknown system characteristics and obtain optimal regret performance. This chapter underscores the significance of utilizing large-system asymptotic approximation to simplify challenging learning problems with complicated state space. Chapter 3 further exploits the power of fluid approximation, revealing that non-idling service disciplines are not universally optimal. This finding highlights the need to ensure servers have sufficient idle time to sustain both well-being and organizational growth.

Contributions and Implications

The research in this thesis has contributed significant insights to the field of modern service system design and management. These novel managerial insights have far-reaching implications for both the theoretical understanding and practical applications of queueing theory in modern service environments.

The integration of human server behavior into classical queueing models sheds light on the consequences of ignoring strategic behavior. Our findings based on behavior-aware queueing models challenge commonly accepted rules of thumb, providing a more realistic foundation for service system design. Moreover, by incorporating human server behavior into the economic cost structure, we contribute to a growing understanding of the relationship between operational efficiency and employee well-being.

The integration of statistical learning techniques into queueing models addresses the inherent uncertainties in system characteristics. This innovative approach equips service systems with the capability to adapt and optimize their operations in response to evolving conditions, opening new avenues for improving service quality and efficiency.

Future Directions

As we conclude this thesis, we recognize that our journey is far from over. The evolving landscape of modern service systems presents a rich tapestry of complexities waiting to be unraveled. Future research can further extend our understanding of how human behavior, statistical learning, and system design interact in various contexts. Rather than reiterating specific model-related open problems from each chapter, our focus here is on higher-level directions that merit future research.

Interactions between Strategic Customers and Strategic Servers

One promising avenue for future research lies in exploring the interactions between strategic customers and strategic servers within service systems. This dynamic interplay, where customers selfishly decide whether to join the queue and servers selfishly determine their work speed, poses intriguing challenges and opportunities. Understanding how self-centric customer behavior influences server responses and vice versa is a complex yet vital task, providing invaluable insights for optimal system design. Furthermore, it is worth investigating how strategic interactions between customers and servers can impact the distribution

of social welfare, shedding light on whether one side benefits at the expense of the other. Moreover, a deeper exploration of behavioral economics and incentive design within service systems can illuminate the design of effective mechanisms that encourage desirable customer and server behaviors. By aligning incentives with desired outcomes, system managers can design systems that promote better performance.

This line of research carries profound implications for the practices on online marketplaces, where understanding the dynamics between demand and supply sides is crucial for effective operation. Additionally, it can offer valuable insights for experimental design, particularly in studies where the interaction between these two sides can significantly affect data collection and analysis.

Bridging Theory and Practice

The long-term grand vision for advancing the field of service operations management centers on bridging the gap between theoretical and empirical research, ultimately fostering a research ecosystem where theory and practice synergistically reinforce each other. In this thesis, we have pursued this vision by developing analytical queueing models that can predict human behavior documented in empirical or psychological studies. Concurrently, it is also important to empirically validate the managerial insights derived from these analytical models using real data or through behavioral experiments. Such integrative approach ensures that theoretical frameworks are firmly rooted in practice, and that empirical insights contribute to the refinement and relevance of theoretical models.

Dynamic and Decentralized Learning in Queueing

Exploring dynamic learning policies capable of adapting to changing system characteristics in real-time holds profound significance. This research direction involves the development of algorithms and frameworks that continuously monitor and update learned models based on evolving data. For instance, when significant shifts in customer or server behavior occur,

these policies should swiftly adapt to maintain high-quality decision-making. Investigating the trade-offs between model stability and adaptability, and developing strategies that strike a balance between the two, is a promising avenue.

Furthermore, the study of multi-agent learning scenarios, where multiple agents simultaneously engage in learning, presents an intriguing opportunity for exploration. Queueing systems frequently consist of interconnected, geographically and logically diverse components. Centralized coordination is often unfeasible, and simulating the entire system may not be practical. Thus, optimizing individual components becomes essential, contributing to the overall optimality of the system as a whole. This research direction involves examining the dynamics of competition and cooperation among agents. Investigating how agents' strategic decisions, based on their learned models, impact not only their individual performance but also the overall system performance, is a complex challenge. Exploring scenarios where agents can collaborate to optimize the overall system performance, even when their individual interests may conflict, can yield valuable insights into the emergent behaviors and equilibrium states in multi-agent learning settings in queues.

In conclusion, this thesis serves as a foundational step in the quest for a deeper understanding of modern service systems, offering a bridge between traditional queueing theory and the complexities of modern service environments. It is our hope that the methodologies and insights developed herein will inspire future research in this realm.

APPENDIX A

APPENDIX FOR CHAPTER 1

In this chapter, we provide proofs for the results stated in Chapter 1. The proofs of these results are in the order in which they appear in Chapter 1.

A.1 Preliminaries

Our analysis requires knowledge of the Erlang B and C formulae, which are arguably the most fundamental formulae for studying queueing systems (Cooper (1981)). The Erlang B formula represents the steady-state blocking probability in the $M/M/N/N$ queue, given by

$$ErlB(N, \rho) = \frac{\rho^N / N!}{\sum_{i=0}^N \rho^i / i!}, \quad \rho > 0. \quad (\text{A.1})$$

For a constant $\rho > 0$, $ErlB(N, \rho)$ satisfies the recursion

$$ErlB(N, \rho) = \frac{\rho ErlB(N-1, \rho)}{N + \rho ErlB(N-1, \rho)}, \quad (\text{A.2})$$

where $ErlB(0, \rho) \equiv 1$; see, e.g., pp.82 of Cooper (1981). The Erlang C formula represents the steady-state probability of delay in the $M/M/N$ queue, given by

$$ErlC(N, \rho) = \frac{\rho^N \frac{1}{N!} \frac{N}{N-\rho}}{\sum_{i=0}^{N-1} \rho^i \frac{1}{i!} + \rho^N \frac{1}{N!} \frac{N}{N-\rho}}, \quad \rho > 0. \quad (\text{A.3})$$

For a constant $\rho > 0$, $ErlC(N, \rho)$ connects to $ErlB(N-1, \rho)$ by (Cooper, 1981, p.92)

$$ErlC(N, \rho) = \frac{\rho ErlB(N-1, \rho)}{\rho ErlB(N-1, \rho) + N - \rho}. \quad (\text{A.4})$$

Lemma 18 (Monotonicity of Erlang B and C). *The following hold:*

(a) $ErlB(N, \rho)$ is strictly decreasing in N and strictly increasing in ρ ;

(b) $ErlC(N, \rho)$ is strictly decreasing in N and strictly increasing in ρ .

Lemma 19 (More Properties of Erlang B and C). *The following hold:*

$$(a) \quad ErlC(N, \rho) = N \left(\frac{N-\rho}{ErlB(N, \rho)} + \rho \right)^{-1}.$$

$$(b) \quad ErlC(N, \rho) \begin{cases} < 1, & \rho < N \\ = 1, & \rho = N \\ > 1, & \rho > N \end{cases}$$

Moreover, $\lim_{\rho \downarrow 0} ErlC(N, \rho) = 0$ and $\lim_{\rho \rightarrow \infty} ErlC(N, \rho)/\rho = 1$.

$$(c) \quad \frac{1-ErlC(N, \rho)}{N-\rho} \in (0, 1), \forall \rho > 0.$$

Lemma 20 (Derivative of Erlang C). $\frac{\partial ErlC(N, \rho)}{\partial \rho} = ErlC(N, \rho) \left(\frac{1-ErlC(N, \rho)}{N-\rho} + \frac{N-\rho}{\rho} \right).$

A.1.1 Proof of Lemma 18

(a): The proof can be found in Problems 4 and 6 in Section 1 of Whitt (2002).

(b): The proof can be found in Problem 2 in Section 2 of Whitt (2002).

■

A.1.2 Proof of Lemma 19

(a): From (A.2),

$$ErlB(N, \rho) = \frac{\rho ErlB(N-1, \rho)}{N + \rho B(N-1, \rho)},$$

which implies

$$\frac{1}{ErlB(N-1, \rho)} = \left(\frac{1}{ErlB(N, \rho)} - 1 \right) \frac{\rho}{N}. \quad (\text{A.5})$$

From (A.4),

$$ErlC(N, \rho) = \frac{\frac{\rho}{N} ErlB(N-1, \rho)}{\frac{\rho}{N} ErlB(N-1, \rho) + 1 - \frac{\rho}{N}} = \frac{\frac{\rho}{N}}{\frac{1 - \frac{\rho}{N}}{ErlB(N-1, \rho)} + \frac{\rho}{N}},$$

which, by (A.5), evaluates to

$$\frac{\frac{\rho}{N}}{(1 - \frac{\rho}{N}) \left(\frac{1}{ErlB(N, \rho)} - 1 \right) \frac{\rho}{N} + \frac{\rho}{N}} = \frac{1}{(1 - \frac{\rho}{N}) \left(\frac{1}{ErlB(N, \rho)} - 1 \right) + 1} = \frac{N}{\frac{N-\rho}{ErlB(N, \rho)} + \rho}.$$

Hence,

$$ErlC(N, \rho) = N \left(\frac{N - \rho}{ErlB(N, \rho)} + \rho \right)^{-1}.$$

(b): Letting $\rho = N$ in the Erlang-C formula (A.3):

$$ErlC(N, N) = \frac{\frac{N^N}{N!} \frac{N}{N-N}}{\sum_{i=0}^{N-1} \frac{N^i}{i!} + \frac{N^N}{N!} \frac{N}{N-N}} = 1.$$

From Lemma 18 (b), $ErlC(N, \rho)$ is strictly increasing in ρ . Thus, $ErlC(N, \rho) < 1$ when $\rho < N$, and $ErlC(N, \rho) > 1$ when $\rho > N$.

Moreover, note that

$$\lim_{\rho \downarrow 0} ErlB(N, \rho) = \lim_{\rho \downarrow 0} \left(\sum_{i=0}^N \frac{\rho^i}{i!} \frac{N!}{\rho^N} \right)^{-1} = \lim_{\rho \downarrow 0} \left(\sum_{i=0}^N \frac{(i+1)(i+2)\dots N}{\rho^{N-i}} \right)^{-1} = 0.$$

Then, it follows from part (a) that

$$\lim_{\rho \downarrow 0} ErlC(N, \rho) = \lim_{\rho \downarrow 0} \frac{N}{\frac{N-\rho}{ErlB(N, \rho)} + \rho} = 0.$$

Additionally, from part (a),

$$\begin{aligned}
& \lim_{\rho \rightarrow \infty} \frac{ErlC(N, \rho)}{\rho} = \lim_{\rho \rightarrow \infty} \frac{N/\rho}{\frac{N-\rho}{ErlB(N, \rho)} + \rho} = \lim_{\rho \rightarrow \infty} \frac{N/\rho}{(N-\rho) \left(\sum_{i=0}^N \frac{\rho^i}{i!} \frac{N!}{\rho^N} \right) + \rho} \\
&= \lim_{\rho \rightarrow \infty} \frac{N/\rho}{(N-\rho) \left(1 + \frac{N}{\rho} + \frac{N(N-1)}{\rho^2} + \dots + \frac{N!}{\rho^{N-1}} + \frac{N!}{\rho^N} \right) + \rho} \\
&= \lim_{\rho \rightarrow \infty} \frac{N}{(N-\rho) \left(\rho + N + \frac{N(N-1)}{\rho} + \dots + \frac{N!}{\rho^N} + \frac{N!}{\rho^{N-1}} \right) + \rho^2} \\
&= \lim_{\rho \rightarrow \infty} N \left[N\rho + N^2 + \frac{N^2(N-1)}{\rho} + \dots + \frac{N!N}{\rho^N} + \frac{N!N}{\rho^{N-1}} - \rho^2 - \rho N - N(N-1) \right. \\
&\quad \left. - \frac{N(N-1)(N-2)}{\rho} - \dots - \frac{N!}{\rho^{N-1}} - \frac{N!}{\rho^{N-2}} + \rho^2 \right]^{-1} \\
&= \lim_{\rho \rightarrow \infty} \frac{N}{\frac{N^2(N-1)}{\rho} + \dots + \frac{N!N}{\rho^N} + \frac{N!N}{\rho^{N-1}} + N - \frac{N(N-1)(N-2)}{\rho} - \dots - \frac{N!}{\rho^{N-1}} - \frac{N!}{\rho^{N-2}}} = 1.
\end{aligned}$$

(c): Expanding $ErlC(N, \rho)$ as finite summations:

$$\frac{1 - ErlC(N, \rho)}{N - \rho} = \frac{\frac{1}{N} \sum_{i=0}^{N-1} \frac{\rho^i}{i!}}{(1 - \frac{\rho}{N}) \sum_{i=0}^{N-1} \frac{\rho^i}{i!} + \frac{\rho^N}{N!}} = \frac{\frac{1}{N} \sum_{i=0}^{N-1} \frac{\rho^i}{i!}}{\sum_{i=0}^N \frac{\rho^i}{i!} - \frac{\rho}{N} \sum_{i=0}^{N-1} \frac{\rho^i}{i!}} = \frac{\sum_{i=0}^{N-1} \frac{1}{N} \frac{\rho^i}{i!}}{\sum_{i=0}^{N-1} \left(1 - \frac{i}{N} \right) \frac{\rho^i}{i!}}.$$

Since $0 < \frac{1}{N} < 1 - \frac{i}{N}$, $\forall i < N - 1$ and $\frac{1}{N} = 1 - \frac{i}{N}$ for $i = N - 1$, it follows that $\frac{1 - ErlC(N, \rho)}{N - \rho} \in (0, 1)$, $\forall \rho > 0$ and $\forall N \geq 2$.

■

A.1.3 Proof of Lemma 20

Differentiating $ErlC(N, \rho)$ with respect to ρ yields

$$\begin{aligned}
\frac{\partial ErlC(N, \rho)}{\partial \rho} &= \left(\sum_{i=0}^{N-1} \frac{\rho^i}{i!} + \frac{\rho^N}{N!} \frac{N}{N-\rho} \right)^{-2} \left[\left(\frac{N\rho^{N-1}}{N!} \frac{N}{N-\rho} + \frac{\rho^N}{N!} \frac{N}{(N-\rho)^2} \right) \left(\sum_{i=0}^{N-1} \frac{\rho^i}{i!} + \frac{\rho^N}{N!} \frac{N}{N-\rho} \right) - \right. \\
&\quad \left. \left(\frac{\rho^N}{N!} \frac{N}{N-\rho} \right) \left(\sum_{i=0}^{N-1} \frac{i\rho^{i-1}}{i!} + \frac{N\rho^{N-1}}{N!} \frac{N}{N-\rho} + \frac{\rho^N}{N!} \frac{N}{(N-\rho)^2} \right) \right]
\end{aligned}$$

$$\begin{aligned}
&= \left(\sum_{i=0}^{N-1} \frac{\rho^i}{i!} + \frac{\rho^N}{N!} \frac{N}{N-\rho} \right)^{-2} \left[\left(\frac{N\rho^{N-1}}{N!} \frac{N}{N-\rho} + \frac{\rho^N}{N!} \frac{N}{(N-\rho)^2} \right) \left(\sum_{i=0}^{N-1} \frac{\rho^i}{i!} \right) \right. \\
&\quad \left. - \left(\frac{\rho^N}{N!} \frac{N}{N-\rho} \right) \sum_{i=0}^{N-1} \frac{i\rho^{i-1}}{i!} \right] \\
&= \frac{\frac{\rho^N}{N!} \frac{N}{N-\rho}}{\sum_{i=0}^{N-1} \frac{\rho^i}{i!} + \frac{\rho^N}{N!} \frac{N}{N-\rho}} \frac{\left(\frac{N}{\rho} + \frac{1}{N-\rho} \right) \left(\sum_{i=0}^{N-1} \frac{\rho^i}{i!} \right) - \sum_{i=0}^{N-1} \frac{i\rho^{i-1}}{i!}}{\sum_{i=0}^{N-1} \frac{\rho^i}{i!} + \frac{\rho^N}{N!} \frac{N}{N-\rho}} \\
&= ErlC(N, \rho) \frac{\left(\frac{1}{N-\rho} \sum_{i=0}^{N-1} \frac{\rho^i}{i!} \right) + \left(\left(\frac{N}{\rho} - 1 \right) \sum_{i=0}^{N-1} \frac{\rho^i}{i!} + \sum_{i=0}^{N-1} \frac{\rho^i}{i!} - \sum_{i=0}^{N-1} \frac{i\rho^{i-1}}{i!} \right)}{\sum_{i=0}^{N-1} \frac{\rho^i}{i!} + \frac{\rho^N}{N!} \frac{N}{N-\rho}} \\
&= ErlC(N, \rho) \frac{\left(\frac{1}{N-\rho} \sum_{i=0}^{N-1} \frac{\rho^i}{i!} \right) + \left(\left(\frac{N}{\rho} - 1 \right) \sum_{i=0}^{N-1} \frac{\rho^i}{i!} + \sum_{i=0}^{N-1} \frac{\rho^i}{i!} - \sum_{i=0}^{N-1} \frac{i\rho^{i-1}}{i!} \right)}{\sum_{i=0}^{N-1} \frac{\rho^i}{i!} + \frac{\rho^N}{N!} \frac{N}{N-\rho}} \\
&= ErlC(N, \rho) \left[\frac{1}{N-\rho} \frac{\sum_{i=0}^{N-1} \frac{\rho^i}{i!}}{\sum_{i=0}^{N-1} \frac{\rho^i}{i!} + \frac{\rho^N}{N!} \frac{N}{N-\rho}} + \frac{\left(\frac{N}{\rho} - 1 \right) \sum_{i=0}^{N-1} \frac{\rho^i}{i!} + \frac{\rho^{N-1}}{(N-1)!}}{\sum_{i=0}^{N-1} \frac{\rho^i}{i!} + \frac{\rho^N}{N!} \frac{N}{N-\rho}} \right] \\
&= ErlC(N, \rho) \left(\frac{1 - ErlC(N, \rho)}{N - \rho} + \frac{N - \rho}{\rho} \right).
\end{aligned}$$

Thus,

$$\frac{\partial ErlC(N, \rho)}{\partial \rho} = ErlC(N, \rho) \left(\frac{1 - ErlC(N, \rho)}{N - \rho} + \frac{N - \rho}{\rho} \right).$$

■

A.2 Proofs from Section 1.1.3

A.2.1 Preliminaries

Lemma 21 (Properties of Idle Time). $I_i(\mu; \lambda, k, N)$ satisfies the following monotonicity properties.

(a) $I_i(\mu; \lambda, k, N)$ is a strictly increasing function of μ_j , $1 \leq j \leq N$.

(b) $I_i(\mu; \lambda, k, N)$ is a strictly decreasing function of k .

(c) $I_i(\mu; \lambda, k, N)$ is a strictly decreasing function of λ .

A.2.1.1 Proof of Lemma 21

We begin by considering an $M/M/1/k$ queueing system with arrival rate λ and service rate μ . The birth-death process is shown in Figure A.1. Let P_ℓ denote the steady-state probability of ℓ jobs in the system. Then, the balance equations are given by

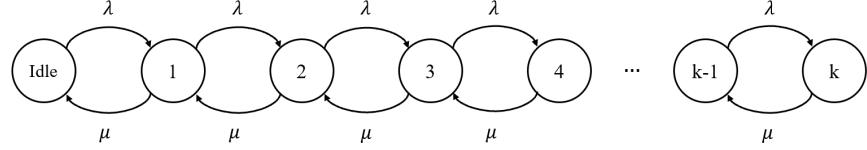


Figure A.1: Birth-death process for the $M/M/1/k$ system

$$\begin{aligned} -\lambda P_0 + \mu P_1 &= 0, \\ -(\lambda + \mu)P_i + \lambda P_{i-1} + \mu P_{i+1} &= 0, \quad 1 \leq i \leq k-1 \\ -\mu P_k + \lambda P_{k-1} &= 0. \end{aligned}$$

Solving the above system of equations yields the steady-state probabilities:

$$P_i = \left(\frac{\lambda}{\mu}\right)^i P_0, \quad 1 \leq i \leq k, \quad \text{where } P_0 = \frac{1}{\sum_{i=1}^k \left(\frac{\lambda}{\mu}\right)^i}. \quad (\text{A.6})$$

Here, P_0 is the idle time, i.e., the probability that the server is idle. It is straightforward that P_0 is strictly increasing in μ , strictly decreasing in λ , and strictly decreasing in k .

Now, we extend the analysis to an $M/M/N/k$ system with arrival rate λ and heterogeneous service rates μ_i ($i = 1, 2, \dots, N$). The birth-death process is demonstrated in Figure A.2. Note that all the states framed in the red dashed square represent some servers

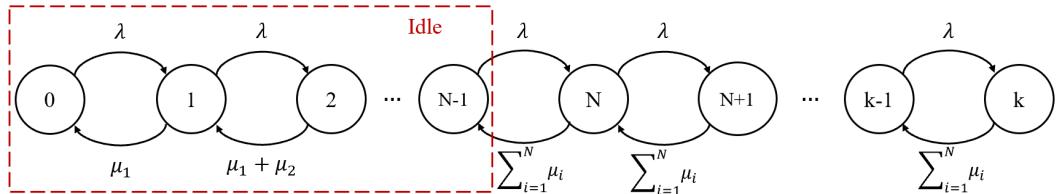


Figure A.2: Birth-death process for the $M/M/N/k$ system

are idle. Thus, it is plausible to regard states $\{0, 1, 2, \dots, N-1\}$ as one “super” idle state and obtain an equivalent birth-death process, as shown in Figure A.3. The equivalent birth-death

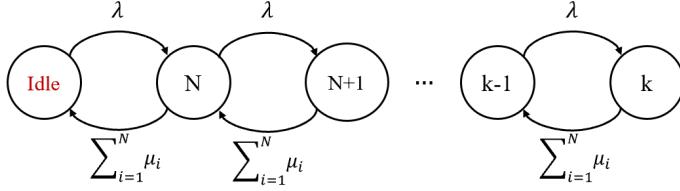


Figure A.3: Equivalent birth-death process for the $M/M/N/k$ system

process for the $M/M/N/k$ system can be viewed as an $M/M/1/k$ system with arrival rate λ , service rate $\sum_{i=1}^N \mu_i$ and a state truncation at N . Thus, following the stationary result (A.6) in the $M/M/1/k$ system, the steady-state probabilities of this equivalent birth-death process can be written as

$$P'_i = \left(\frac{\lambda}{\sum_{i=1}^N \mu_i} \right)^{i-N+1} P'_0, \quad N \leq i \leq k, \quad \text{where} \quad P'_0 = \frac{1}{\sum_{i=1}^{k-N+1} \left(\frac{\lambda}{\sum_{i=1}^N \mu_i} \right)^i}, \quad (\text{A.7})$$

and P'_0 is the probability of the “super” idle state. Denote the probably that all servers are busy with no jobs waiting in queue by P'_N , then it follows that

$$P'_N = P'_0 \frac{\lambda}{\sum_{i=1}^N \mu_i} = \frac{1}{\sum_{i=0}^{k-N} \left(\frac{\lambda}{\sum_{i=1}^N \mu_i} \right)^i}. \quad (\text{A.8})$$

Next, we examine the idle probability of each individual server. Let $\mathcal{I} \subseteq \{1, 2, \dots, N\}$ be the set of idle servers ($\mathcal{I} = \emptyset$ when all servers are busy), and state $\mathbf{a} = (a_1, a_2, \dots, a_{|\mathcal{I}|})$ be the ordered vector of idle servers where server j became idle before server k whenever $1 \leq j < k \leq |\mathcal{I}|$. Denote by $P(\mathbf{a})$ the steady-state probability of state \mathbf{a} . From Gopalakrishnan et al. (2016a), all idle-time-order-based routing policies have the same steady-state probabilities

as random routing, and satisfy

$$P(\mathbf{a}) = P'_N \prod_{\ell=1}^{|\mathcal{I}(\mathbf{a})|} \frac{\mu_\ell}{\lambda}, \quad \text{for all } \mathbf{a} = (a_1, a_2, \dots, a_{|\mathcal{I}(\mathbf{a})|}) \text{ with } |\mathcal{I}(\mathbf{a})| > 0,$$

where $\mathcal{I}(\mathbf{a})$ denote the set of idle servers in state \mathbf{a} . Then, the steady-state probability that server i is idle (i.e., idleness fraction), denoted by I_i , can be written as

$$I_i = \sum_{\mathbf{a}: \{i\} \in \mathcal{I}(\mathbf{a})} P(\mathbf{a}) = \sum_{\mathbf{a}: \{i\} \in \mathcal{I}(\mathbf{a})} P'_N \prod_{\ell=1}^{|\mathcal{I}(\mathbf{a})|} \frac{\mu_\ell}{\lambda}, \quad \text{for all } i \in \{1, 2, \dots, N\}. \quad (\text{A.9})$$

From (A.8), it is straightforward to observe that

- P'_N is strictly increasing in μ_j for all $j \in \{1, 2, \dots, N\}$;
- P'_N is strictly decreasing in k .
- P'_N is strictly decreasing in λ ;

(a): As μ_j ($j \in \{1, 2, \dots, N\}$) increases, P'_N increases, and from (A.9), I_i is also increasing in all μ_j . Thus, I_i is increasing in μ_j ($j \in \{1, 2, \dots, N\}$).

(b): As k increases, P'_N decreases, and from (A.9), k influences I_i only through P'_N . Thus, it is clear that I_i is decreasing in k .

(c): As λ increases, P'_N decreases, and from (A.9), I_i is also decreasing in λ . Thus, I_i is decreasing in λ .

■

A.2.2 Proof of Lemma 1

We first observe that the steady-state probabilities when at least one server is idle have an identical form to those for the infinite-buffer system given in (21) in Gopalakrishnan et al. (2016a). Then, it is sufficient to verify the detailed balance equations of the corresponding

Markov chain. Identical to the proof of Theorem 9 in Gopalakrishnan et al. (2016a), one can show that, for any state $\mathbf{a} = (a_1, a_2, \dots, a_{|\mathcal{I}|})$, where \mathcal{I} is the set of idle servers,

- (i) Rate into state \mathbf{a} due to an arrival = Rate out of state \mathbf{a} due to a departure; and
- (ii) Rate into state \mathbf{a} due to a departure = Rate out of state \mathbf{a} due to an arrival.

■

A.2.3 Proof of Lemma 2

The proof is closely related to the proof of Theorem 1 in Gopalakrishnan et al. (2016a), which relies on Gumbel (1960), and utilizes the same techniques. The difference is that their results are for the infinite buffer system ($M/M/N$). For the finite-buffer system ($M/M/N/k$), the Markov chain is the same as that for the infinite-buffer system except that it is truncated at state k . As a result, the balance equations for all the states except that for state k remain the same, and the normalization constant is different.

We let $(a_1, a_2, \dots, a_\ell)$ denote the state of the $M/M/N/k$ system when there are ℓ jobs in the system ($0 < \ell < N$) and the busy servers are $\{a_1, a_2, \dots, a_\ell\}$, where $1 \leq a_1 < a_2 < \dots < a_\ell \leq N$. Let $P(a_1, a_2, \dots, a_\ell)$ denote the steady-state probability of the $M/M/N/k$ system being in state $(a_1, a_2, \dots, a_\ell)$. Also let P_ℓ denote the steady-state probability of ℓ jobs in the $M/M/N/k$ system. We first note that Equations (EC.1) and (EC.6) in Gopalakrishnan et al. (2016a) continue to hold for the finite-buffer $M/M/N/k$ system, and so

- When there are $\ell \in \{1, 2, \dots, N-1\}$ jobs in the system, and the tagged server (server 1) is idle,

$$P(a_1, a_2, \dots, a_\ell) = \frac{(N-\ell)! P_0 \rho^\ell}{N!}, \quad 2 \leq a_1 \leq \dots \leq a_\ell \leq N. \quad (\text{A.10})$$

- When there are $\ell \in \{1, 2, \dots, N\}$ jobs in the system, and the tagged server (server 1)

is busy,

$$P(1, a_2, \dots, a_\ell) = \frac{(N - \ell)! P_0 \rho_1 \rho^{\ell-1}}{N!}, \quad 2 \leq a_2 \leq \dots \leq a_\ell \leq N. \quad (\text{A.11})$$

Combining (A.10) and (A.11), we can write down the steady-state probability of $\ell \in \{1, 2, \dots, N-1\}$ jobs in the system as

$$\begin{aligned} P_\ell &= \sum_{2 \leq a_1 \leq \dots \leq a_\ell \leq N} P(a_1, a_2, \dots, a_\ell) + \sum_{2 \leq a_2 \leq \dots \leq a_\ell \leq N} P(1, a_2, \dots, a_\ell) \\ &= P_0 \left(\sum_{2 \leq a_1 \leq \dots \leq a_\ell \leq N} \frac{(N - \ell)! \rho^\ell}{N!} + \sum_{2 \leq a_2 \leq \dots \leq a_\ell \leq N} \frac{(N - \ell)! \rho_1 \rho^{\ell-1}}{N!} \right). \end{aligned} \quad (\text{A.12})$$

Letting $\ell = N$ in (A.11) implies

$$P_N = P(1, 2, \dots, N) = \frac{P_0 \rho_1 \rho^{N-1}}{N!} = P_0 \frac{\rho_1 \rho^N}{\rho N!}.$$

When there are $\ell \geq N$ jobs in the system, the system behaves as a single-server queue with service rate $(N-1)\mu + \mu_1$, and so, from the balance equations,

$$P_\ell = P_N \left(\frac{\lambda}{(N-1)\mu + \mu_1} \right)^{\ell-N} = P_0 \left(\frac{\rho_1 \rho^N}{\rho N!} \left(\frac{\rho}{N - (1 - \frac{\rho}{\rho_1})} \right)^{\ell-N} \right), \quad N \leq \ell \leq k. \quad (\text{A.13})$$

Thus, we can obtain the expression for P_0 using the normalization constraint, which yields

$$P_0 = \left(1 + \sum_{\ell=1}^k \frac{P_\ell}{P_0} \right)^{-1}.$$

Note that the ratio $\frac{P_\ell}{P_0}$ is independent of k , in both (A.12) and (A.13). Thus, letting $k \rightarrow \infty$ in the above display implies that the steady-state probability of an empty infinite-

buffer $M/M/N$ system, denoted by $P_0^{M/M/N}$, satisfies

$$P_0^{M/M/N} = \left(1 + \sum_{\ell=1}^{\infty} \frac{P_\ell}{P_0} \right)^{-1} = \left(1 + \sum_{\ell=1}^k \frac{P_\ell}{P_0} + \sum_{\ell=k+1}^{\infty} \frac{P_\ell}{P_0} \right)^{-1} = \left((P_0)^{-1} + \sum_{\ell=k+1}^{\infty} \frac{P_\ell}{P_0} \right)^{-1}.$$

Hence, we can express P_0 in terms of $P_0^{M/M/N}$ as

$$P_0 = \left((P_0^{M/M/N})^{-1} - \sum_{\ell=k+1}^{\infty} \frac{P_\ell}{P_0} \right)^{-1}.$$

Using the expression for $P_0^{M/M/N}$, as shown in the display following (EC.8) in Gopalakrishnan et al. (2016a), and substituting that equivalence and (A.13) into the above expression:

$$\begin{aligned} P_0 &= \left(\left(1 - \frac{\rho}{N} \left(1 - \frac{\rho_1}{\rho} \right) \right) \sum_{\ell=0}^{N-1} \frac{\rho^\ell}{\ell!} + \frac{\rho^N}{N!} \left(1 + \frac{\rho_1}{(N-\rho) - \left(1 - \frac{\rho}{\rho_1} \right)} \right) - \frac{\rho_1}{\rho} \frac{\rho^N}{N!} \sum_{\ell=k+1}^{\infty} \left(\frac{\rho}{N - \left(1 - \frac{\rho}{\rho_1} \right)} \right)^{\ell-N} \right)^{-1} \\ &= \left(\left(1 - \frac{\rho}{N} \left(1 - \frac{\rho_1}{\rho} \right) \right) \sum_{\ell=0}^{N-1} \frac{\rho^\ell}{\ell!} + \frac{\rho^N}{N!} \left(1 + \frac{\rho_1}{(N-\rho) - \left(1 - \frac{\rho}{\rho_1} \right)} \left(1 - \left(\frac{\rho}{N - \left(1 - \frac{\rho}{\rho_1} \right)} \right)^{k-N} \right) \right) \right)^{-1}. \end{aligned}$$

Note that the term in red appearing in P_0 vanishes when $k \rightarrow \infty$, recovering the expression for $P_0^{M/M/N}$ in the infinite-buffer system, as given in (4) in Gopalakrishnan et al. (2016a).

(This is a sanity check.)

Next, we use the same manoeuvre used to obtain (EC.9) in Gopalakrishnan et al. (2016a) to express P_0 in terms of $ErlC(N, \rho)$; i.e., we add and subtract the term $\frac{N}{N-\rho} \frac{\rho^N}{N!}$, and follow similar algebraic simplifications to obtain

$$\begin{aligned} P_0 &= \left(\sum_{\ell=0}^{N-1} \frac{\rho^\ell}{\ell!} + \frac{N}{N-\rho} \frac{\rho^N}{N!} \right)^{-1} \\ &\quad \cdot \left(1 - \frac{\rho}{N} \left(1 - \frac{\rho_1}{\rho} + \left(1 - \frac{\rho_1}{\rho} + \frac{N\rho_1}{\rho} \left(1 - \frac{\rho}{N} \right) \left(\frac{\rho}{N - \left(1 - \frac{\rho}{\rho_1} \right)} \right)^{k-N} \right) \frac{ErlC(N, \rho)}{(N-\rho) - \left(1 - \frac{\rho}{\rho_1} \right)} \right) \right)^{-1}. \end{aligned}$$

Identical to Equation (EC.5) in Gopalakrishnan et al. (2016a), the formula for the tagged

server's idle time is

$$I(\mu_1, \mu; \lambda, k, N) = P_0 + \sum_{\ell=1}^{N-1} \sum_{2 \leq a_1 \leq \dots \leq a_\ell \leq N} P(a_1, a_2, \dots, a_\ell).$$

Following similar final steps as those in Gopalakrishnan et al. (2016a), we find:

$$\begin{aligned} & I(\mu_1, \mu; \lambda, k, N) \\ &= \left(1 - \frac{\rho}{N}\right) \left(\sum_{\ell=0}^{N-1} \frac{\rho^\ell}{\ell!} + \frac{N}{N-\rho} \frac{\rho^N}{N!} \right) P_0 \\ &= \left(1 - \frac{\rho}{N}\right) \left(1 - \frac{\rho}{N} \left(1 - \frac{\mu}{\mu_1} + \left(1 - \frac{\mu}{\mu_1} + \frac{N\mu}{\mu_1} \left(1 - \frac{\rho}{N} \right) \left(\frac{\rho}{N - (1 - \frac{\mu_1}{\mu})} \right)^{k-N} \right) \frac{ErlC(N, \rho)}{(N - \rho) - (1 - \frac{\mu_1}{\mu})} \right) \right)^{-1} \\ &= \left(\frac{N}{N-\rho} - \rho \left(\left(1 - \frac{\mu}{\mu_1} \right) \left(1 + \frac{ErlC(N, \rho)}{(N - \rho) - (1 - \frac{\mu_1}{\mu})} \right) \frac{1}{N - \rho} + \frac{\mu}{\mu_1} \left(\frac{\rho}{N - (1 - \frac{\mu_1}{\mu})} \right)^{k-N} \frac{ErlC(N, \rho)}{(N - \rho) - (1 - \frac{\mu_1}{\mu})} \right) \right)^{-1} \\ &= \left(1 + \rho \frac{\mu}{\mu_1} \left(\frac{1 - ErlC(N, \rho)}{N - \rho} + \left(1 - \left(\frac{\rho}{N - (1 - \frac{\mu_1}{\mu})} \right)^{k-N} \right) \frac{ErlC(N, \rho)}{(N - \rho) - (1 - \frac{\mu_1}{\mu})} \right) \right)^{-1}, \end{aligned}$$

which establishes (1.1).

Finally, let $x = \frac{\rho}{N - (1 - \frac{\mu_1}{\mu})}$. Note that $(N - \rho) - (1 - \frac{\mu_1}{\mu}) = [N - (1 - \frac{\mu_1}{\mu})] - \rho = [N - (1 - \frac{\mu_1}{\mu})] - x [N - (1 - \frac{\mu_1}{\mu})] = (1 - x) [N - (1 - \frac{\mu_1}{\mu})]$. Then,

$$\begin{aligned} & \left(1 - \left(\frac{\rho}{N - (1 - \frac{\mu_1}{\mu})} \right)^{k-N} \right) \frac{ErlC(N, \rho)}{(N - \rho) - (1 - \frac{\mu_1}{\mu})} = (1 - x^{k-N}) \frac{ErlC(N, \rho)}{(1 - x) [N - (1 - \frac{\mu_1}{\mu})]} \\ &= (1 - x) \left(\sum_{i=0}^{k-N-1} x^i \right) \frac{ErlC(N, \rho)}{(1 - x) [N - (1 - \frac{\mu_1}{\mu})]} = \left(\sum_{i=0}^{k-N-1} x^i \right) \frac{ErlC(N, \rho)}{N - (1 - \frac{\mu_1}{\mu})}. \end{aligned}$$

When $\lambda \geq (N - 1)\mu + \mu_1$, we have $\rho \geq N - (1 - \frac{\mu_1}{\mu})$, i.e., $x \geq 1$. Thus, $\sum_{i=0}^{k-N-1} x^i \geq k - N \rightarrow \infty$, as $k \rightarrow \infty$. Hence, from (1.1) and the above display, $I(\mu_1, \mu; \lambda, \infty, N) := \lim_{k \rightarrow \infty} I(\mu_1, \mu; \lambda, k, N) = 0$.

■

A.3 Proofs from Section 1.2

A.3.1 Proof of Proposition 1

From (1.1), $\lim_{\mu_1 \downarrow 0} I(\mu_1, \mu) = 0$ for all $\mu > 0$. Together with $c(0) = 0$, this implies that $\lim_{\mu_1 \downarrow 0} U(\mu_1, \mu) = 0$ for all $\mu > 0$. Thus, any equilibrium $\mu^* > 0$ satisfies $U(\mu^*, \mu^*) \geq 0$.

■

A.3.2 Proof of Lemma 3

We first note that the cost function c is strictly increasing with $c(0) = 0$, and is therefore invertible in $[0, \infty)$. From Proposition 1, $U(\mu^*, \mu^*) = p\mu^* + (v - p\mu^*)I(\mu^*, \mu^*) - c(\mu^*) \geq 0$. Using this inequality and the trivial bound $I(\mu^*, \mu^*) < 1$, we show the next claim, whose proof will appear at the end.

Claim 1. *If $c(\mu^*) > p\mu^*$, then $\mu^* < \frac{v}{p}$ and $\mu^* < (c')^{-1}(v)$.*

Case (I): If $c'(0) \geq p$, because c is strictly convex, we have $c(\mu^*) > c'(0)(\mu^* - 0) \geq p\mu^*$.

Then, it follows from Claim 1 that $\mu^* \leq \min \left\{ \frac{v}{p}, (c')^{-1}(v) \right\}$.

Case (II): If $c'(0) < p$, because c is strictly increasing and strictly convex with $c(0) = 0$, and $p\mu$ is linear with zero intercept, there exists a unique $\mu_0 > 0$ such that

$$c(\mu_0) = p\mu_0 \quad \text{and} \quad c(\mu) < p\mu \Leftrightarrow \mu < \mu_0. \quad (\text{A.14})$$

Next, we discuss two cases.

Case (II-1): If $c(\frac{v}{p}) < v$, which can be equivalently written as $c(\frac{v}{p}) < p \cdot \frac{v}{p}$, then it follows from (A.14) that $\frac{v}{p} < \mu_0$. We show that $\mu^* \leq \mu_0$ by contradiction. Suppose $\mu^* > \mu_0$, then $c(\mu^*) > p\mu^*$ using (A.14), which implies $\mu^* < \frac{v}{p}$ from Claim 1. This contradicts $\mu^* > \mu_0 > \frac{v}{p}$.

Case (II-2): If $c(\frac{v}{p}) \geq v$, which can be equivalently written as

$$\frac{v}{p} \geq c^{-1}(v) \Leftrightarrow v \geq p \cdot c^{-1}(v) \Leftrightarrow c(c^{-1}(v)) \geq p \cdot c^{-1}(v).$$

Then, (A.14) implies that $c^{-1}(v) \geq \mu_0$. Thus,

$$\mu_0 \leq c^{-1}(v) \leq \frac{v}{p}. \quad (\text{A.15})$$

- If $\mu^* \leq \mu_0$, then it follows from (A.15) that $\mu^* \leq \mu_0 \leq c^{-1}(v)$.
- If $\mu^* > \mu_0$, then (A.14) implies that $c(\mu^*) > p\mu^*$. Then, it follows from Claim 1 that $\mu^* \leq c^{-1}(v)$, recalling from (A.15) that $c^{-1}(v) \leq \frac{v}{p}$.

Combining all the cases above,

$$\mu^* \leq \mu_{\max}^*(p, v) = \begin{cases} \mu_0, & c'(0) < p \text{ and } c(\frac{v}{p}) < v, \\ c^{-1}(v), & c'(0) < p \text{ and } c(\frac{v}{p}) \geq v, \\ \min \left\{ \frac{v}{p}, c^{-1}(v) \right\}, & c'(0) \geq p. \end{cases}$$

■

A.3.2.1 Proof of Claim 1

Since $c(\mu^*) > p\mu^*$, $U(\mu^*, \mu^*) = p\mu^* + (v - p\mu^*)I(\mu^*, \mu^*) - c(\mu^*) \geq 0$ holds only if $(v - p\mu^*)I(\mu^*, \mu^*) > 0$, implying that $v - p\mu^* > 0$ (since $I(\mu^*, \mu^*) > 0$), i.e., $\mu^* < \frac{v}{p}$.

In addition, since $I(\mu^*, \mu^*) < 1$, $U(\mu^*, \mu^*) = p\mu^* + (v - p\mu^*)I(\mu^*, \mu^*) - c(\mu^*) < v - c(\mu^*)$ (recalling that $v - p\mu^* > 0$). Thus, $U(\mu^*, \mu^*) = p\mu^* + (v - p\mu^*)I(\mu^*, \mu^*) - c(\mu^*) \geq 0$ holds only if its strict upper bound is strictly positive; that is, $v - c(\mu^*) > 0$, i.e., $\mu^* < c^{-1}(v)$.

■

A.4 Proofs from Section 1.3

A.4.1 Preliminaries

We start by providing closed-form expressions for the first two partial derivatives of the idle time with respect to μ_1 .

Lemma 22 (Expressions for Derivatives of Idle Time). *In an $M/M/N/k$ system where server 1 operates at rate $\mu_1 > 0$ and the other $N - 1$ servers operate at rate $\mu > 0$, the first two partial derivatives of $I(\mu_1, \mu; \lambda, k, N)$, from (1.1), with respect to μ_1 are given by*

$$\frac{\partial I(\mu_1, \mu)}{\partial \mu_1} = \frac{1}{\mu_1} I(1 - I) + \frac{I^2}{\mu_1} \left(\frac{C}{d_2} \left[\left(1 - \left(\frac{\rho}{d_1} \right)^{k-N} \right) \frac{\rho}{d_2} - (k - N) \left(\frac{\rho}{d_1} \right)^{k-N} \right] + \frac{k - N}{d_1} \left(\frac{\rho}{d_1} \right)^{k-N} C \right), \quad (\text{A.16})$$

$$\begin{aligned} \frac{\partial^2 I(\mu_1, \mu)}{\partial \mu_1^2} = & \frac{I^2}{\mu_1^2} \left\{ -2(1 - I) + \frac{2\rho C}{d_2^2} \left[1 - \left(\frac{\rho}{d_1} \right)^{k-N} \right] \left[(1 - 2I) - \frac{\frac{\mu_1}{\mu} + \rho \left(\frac{1-C}{N-\rho} \right)}{d_2} I \right] \right. \\ & - \frac{\rho C}{d_2} \frac{k - N}{d_1} \left(\frac{\rho}{d_1} \right)^{k-N} \left[2(1 - 2I) + \left(\frac{2}{d_2} - \frac{1}{d_1} \right) \frac{\mu_1}{\mu} - 4 \frac{\frac{\mu_1}{\mu} + \rho \left(\frac{1-C}{N-\rho} \right)}{d_2} I \right] \\ & \left. - \rho C \left(\frac{k - N}{d_1} \right)^2 \left(\frac{\rho}{d_1} \right)^{k-N} \left[\frac{1}{d_2} \frac{\mu_1}{\mu} - 2 \frac{\frac{\mu_1}{\mu} + \rho \left(\frac{1-C}{N-\rho} \right)}{d_2} I - \frac{2\rho C}{d_2^2} I \right] \right\}, \end{aligned} \quad (\text{A.17})$$

where

$$C := ErlC(N, \rho), \quad d_1 := N - \left(1 - \frac{\mu_1}{\mu} \right) \quad \text{and} \quad d_2 := d_1 - \rho.$$

Corollary 2 ($\mu_1 = \mu$ in Lemmas 2 and 22).

$$I(\mu, \mu) = \left(1 + \rho \left(\frac{1 - ErlC(N, \rho)}{N - \rho} \right) + ErlC(N, \rho) \sum_{i=1}^{k-N} \left(\frac{\rho}{N} \right)^i \right)^{-1} \quad (\text{A.18})$$

$$= \left(\left(1 + \sum_{i=0}^{N-1} \frac{N!}{i!} \left(\frac{\mu}{\lambda} \right)^{N-i} + \sum_{i=1}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i \right) ErlC \left(N, \frac{\lambda}{\mu} \right) \right)^{-1} \quad (\text{A.19})$$

$$= \left(1 - \frac{\rho}{N} \right) \left(1 - ErlC(N, \rho) \left(\frac{\rho}{N} \right)^{k-N+1} \right)^{-1}, \quad (\text{A.20})$$

$$\frac{\partial I(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} = \frac{1}{\mu} I(\mu, \mu)(1 - I(\mu, \mu)) + I(\mu, \mu)^2 \frac{ErlC(N, \rho)}{N\mu} \sum_{i=1}^{k-N} i \left(\frac{\rho}{N} \right)^i \quad (\text{A.21})$$

$$= \frac{\rho}{N\mu} \frac{1 - \frac{\rho}{N} + ErlC(N, \rho) \left(\frac{1}{N} - \left(\frac{\rho}{N} \right)^{k-N} \left(\frac{1}{N} + \frac{k}{N} (1 - \frac{\rho}{N}) \right) \right)}{\left(1 - \left(\frac{\rho}{N} \right)^{k-N+1} ErlC(N, \rho) \right)^2}. \quad (\text{A.22})$$

Next, we present upper bounds on the derivative of idle time, which can help simplify proof.

Lemma 23. *The following hold for all $\lambda > 0$ and $k \geq N \geq 2$:*

(a)

$$I(\mu, \mu)^2 ErlC \left(N, \frac{\lambda}{\mu} \right) \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^i < 2\sqrt{N}, \quad \forall \mu > 0.$$

(b)

$$I(\mu, \mu)^2 ErlC \left(N, \frac{\lambda}{\mu} \right) \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^{i+1} < 2\sqrt{N}, \quad \forall \mu > 0.$$

Corollary 3. *The following hold for all $\lambda > 0$ and $k \geq N \geq 2$:*

(a)

$$\mu \frac{\partial I(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} < I(\mu, \mu)(1 - I(\mu, \mu)) + \frac{2}{\sqrt{N}}, \quad \forall \mu > 0.$$

(b)

$$\frac{\partial I(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} < \frac{I(\mu, \mu)}{\mu} + \frac{2\sqrt{N}}{\lambda}, \quad \forall \mu > 0.$$

The next result expresses the derivative of idle time in terms of the partial derivative of idle time, which is useful for some proof.

Lemma 24.

$$\frac{dI(\mu, \mu)}{d\mu} = N \cdot \frac{\partial I(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} - \frac{I(\mu, \mu)}{\mu} \frac{\lambda}{N\mu} \left(N - \frac{1 - ErlC(N, \frac{\lambda}{\mu})}{1 - \frac{\lambda}{N\mu}} \right).$$

Finally, we provide a useful monotonicity property of the idle time.

Lemma 25. *The difference between the idle time in a finite-buffer M/M/N/k system and that in an infinite-buffer M/M/N system, $I(\mu, \mu; \lambda, k, N) - \left(1 - \frac{\lambda}{N\mu}\right)$, is strictly decreasing in μ for $\mu \in (0, \infty)$, for all $\lambda > 0$ and $k \geq N \geq 2$, and satisfies $I(\mu, \mu; \lambda, k, N) - \left(1 - \frac{\lambda}{N\mu}\right) \geq 0$, with equality holding only when $k = \infty$.*

A.4.1.1 Proof of Lemma 22

First-order derivative:

This is useful to first observe that taking the reciprocal of (1.1) shows

$$\left(\frac{1}{I} - 1\right) \frac{\mu_1}{\mu} \frac{1}{\rho} = \frac{1 - C}{N - \rho} + \frac{C}{d_2} - \frac{C}{d_2} \left(\frac{\rho}{d_1}\right)^{k-N} \quad (\text{A.23})$$

$$\Leftrightarrow \frac{C(N - \rho)}{d_2} \left(\frac{\rho}{d_1}\right)^{k-N} = 1 + \frac{N - \rho}{\rho} \left(1 - \frac{1}{I}\right) \frac{\mu_1}{\mu} + \left(1 - \frac{\mu_1}{\mu}\right) \frac{C}{d_2}. \quad (\text{A.24})$$

Differentiating $I(\mu_1, \mu)$ using (1.1) with respect to μ_1 yields

$$\begin{aligned} \frac{\partial I}{\partial \mu_1} &= -I^2 \cdot \rho \mu \left\{ -\frac{1}{\mu_1^2} \left[\frac{1 - C}{N - \rho} + \left(1 - \left(\frac{\rho}{d_1}\right)^{k-N}\right) \frac{C}{d_2} \right] \right. \\ &\quad \left. + \frac{1}{\mu_1} \left[\left((k - N) \left(\frac{\rho}{d_1}\right)^{k-N-1} \frac{\rho}{d_1^2 \mu}\right) \frac{C}{d_2} - \left(1 - \left(\frac{\rho}{d_1}\right)^{k-N}\right) \frac{C}{d_2^2 \mu} \right] \right\} \\ &= \frac{I^2}{\mu_1^2} \frac{\lambda}{N - \rho} \left\{ 1 + C \frac{(N - \rho) - \left(1 - \frac{\mu_1}{\mu}\right)^2}{d_2^2} - \frac{C(N - \rho)}{d_2} \left(\frac{\rho}{d_1}\right)^{k-N} \left(1 + \frac{\mu_1}{\mu} \frac{1}{d_2} + \frac{\mu_1}{\mu} \frac{k - N}{d_1}\right) \right\}. \end{aligned}$$

Substitution for $\frac{C(N-\rho)}{d_2} \left(\frac{\rho}{d_1}\right)^{k-N}$ using (A.24) and additional algebra shows

$$\frac{\partial I}{\partial \mu_1} = \frac{I^2}{\mu_1} \left\{ \left(1 + \frac{\mu_1}{\mu} \frac{1}{d_2}\right) \left(\frac{1}{I} - 1\right) - \frac{\rho \left(\frac{1-C}{N-\rho}\right)}{d_2} + \frac{k-N}{d_1} \left(\frac{\mu_1}{\mu} \left(\frac{1}{I} - 1\right) - \left(\rho \left(\frac{1-C}{N-\rho}\right) + \frac{\rho C}{d_2}\right)\right) \right\}.$$

Thus,

$$\begin{aligned} \frac{\mu_1}{I^2} \frac{\partial I}{\partial \mu_1} &= \left(1 + \frac{\mu_1}{\mu} \frac{1}{d_2}\right) \left(\frac{1}{I} - 1\right) - \frac{\rho \left(\frac{1-C}{N-\rho}\right)}{d_2} + \frac{k-N}{d_1} \left(\frac{\mu_1}{\mu} \left(\frac{1}{I} - 1\right) - \left(\rho \left(\frac{1-C}{N-\rho}\right) + \frac{\rho C}{d_2}\right)\right) \\ &= \left(\frac{1}{I} - 1\right) + \left(\frac{1}{d_2} + \frac{k-N}{d_1}\right) \frac{\mu_1}{\mu} \left(\frac{1}{I} - 1\right) - \frac{\rho \left(\frac{1-C}{N-\rho}\right)}{d_2} - \frac{k-N}{d_1} \left(\rho \left(\frac{1-C}{N-\rho}\right) + \frac{\rho C}{d_2}\right). \end{aligned} \quad (\text{A.25})$$

Substitution for the second $\left(\frac{1}{I} - 1\right)$ using (A.23) yields

$$\begin{aligned} \frac{\mu_1}{I^2} \frac{\partial I}{\partial \mu_1} &= \left(\frac{1}{I} - 1\right) + \left(\frac{1}{d_2} + \frac{k-N}{d_1}\right) \frac{\mu_1}{\mu} \cdot \rho \frac{\mu}{\mu_1} \left(\frac{1-C}{N-\rho} + \left(1 - \left(\frac{\rho}{d_1}\right)^{k-N}\right) \frac{C}{d_2}\right) - \frac{\rho \left(\frac{1-C}{N-\rho}\right)}{d_2} \\ &\quad - \frac{k-N}{d_1} \left(\rho \left(\frac{1-C}{N-\rho}\right) + \frac{\rho C}{d_2}\right) \\ &= \left(\frac{1}{I} - 1\right) + \frac{C}{d_2} \left[\left(1 - \left(\frac{\rho}{d_1}\right)^{k-N}\right) \frac{\rho}{d_2} - (k-N) \left(\frac{\rho}{d_1}\right)^{k-N} \left(\frac{\rho}{d_1} - 1 + 1\right) \right] \\ &= \left(\frac{1}{I} - 1\right) + \frac{C}{d_2} \left[\left(1 - \left(\frac{\rho}{d_1}\right)^{k-N}\right) \frac{\rho}{d_2} - (k-N) \left(\frac{\rho}{d_1}\right)^{k-N} \right] - \frac{C}{d_2} \frac{k-N}{d_1} d_1 \left(\frac{\rho}{d_1}\right)^{k-N} \left(\frac{\rho}{d_1} - 1\right). \end{aligned}$$

Recall that $d_1 - \rho = d_2$, which implies that $\frac{d_1}{d_2} \left(\frac{\rho}{d_1} - 1\right) = -1$, and so

$$\frac{\mu_1}{I^2} \frac{\partial I}{\partial \mu_1} = \left(\frac{1}{I} - 1\right) + \frac{C}{d_2} \left[\left(1 - \left(\frac{\rho}{d_1}\right)^{k-N}\right) \frac{\rho}{d_2} - (k-N) \left(\frac{\rho}{d_1}\right)^{k-N} \right] + \frac{k-N}{d_1} \left(\frac{\rho}{d_1}\right)^{k-N} C,$$

which establishes (A.16).

Second-order derivative:

From (A.25),

$$\frac{\mu_1}{I} \frac{\partial I}{\partial \mu_1} = \left(1 + \frac{\mu_1}{\mu} \frac{1}{d_2}\right) (1 - I) - \frac{\rho \left(\frac{1-C}{N-\rho}\right)}{d_2} I + \frac{k-N}{d_1} \left(\frac{\mu_1}{\mu} (1-I) - \left(\rho \left(\frac{1-C}{N-\rho}\right) + \frac{\rho C}{d_2}\right) I\right).$$

Set $LHS(\mu_1, \mu)$ and $RHS(\mu_1, \mu)$ equal to the left-hand and right-hand sides of the above equation, respectively. Then,

$$\frac{\partial LHS(\mu_1, \mu)}{\partial \mu_1} = \frac{\mu_1}{I} \frac{\partial^2 I}{\partial \mu_1^2} - \frac{1}{\mu_1} \left[\left(\frac{\mu_1}{I} \frac{\partial I}{\partial \mu_1} \right)^2 - \left(\frac{\mu_1}{\mu} \frac{\partial I}{\partial \mu_1} \right) \right],$$

and

$$\begin{aligned} \frac{\partial RHS(\mu_1, \mu)}{\partial \mu_1} &= \frac{N-\rho-1}{d_2^2} \frac{1}{\mu} - \frac{(N-1)-\rho \left(1 + \frac{1-C}{N-\rho}\right)}{d_2^2} \frac{I}{\mu} - \left(2 - \frac{(N-1)-\rho \left(1 + \frac{1-C}{N-\rho}\right)}{d_2}\right) \frac{\partial I}{\partial \mu_1} \\ &\quad + (k-N) \left[\frac{N-1}{d_1^2} \frac{1}{\mu} - \left(\frac{N \left(1 - \frac{1-C}{N-\rho}\right)}{d_1^2} - \frac{C}{d_2^2} \right) \frac{I}{\mu} - \left(1 - \frac{N \left(1 - \frac{1-C}{N-\rho}\right)}{d_1} + \frac{C}{d_2}\right) \frac{\partial I}{\partial \mu_1} \right] \\ &= \frac{1}{\mu_1} \left[\frac{N-\rho-1}{d_2^2} \frac{\mu_1}{\mu} - \frac{(N-1)-\rho \left(1 + \frac{1-C}{N-\rho}\right)}{d_2^2} \frac{\mu_1}{\mu} I - \left(2 - \frac{(N-1)-\rho \left(1 + \frac{1-C}{N-\rho}\right)}{d_2}\right) I \frac{\mu_1}{I} \frac{\partial I}{\partial \mu_1} \right. \\ &\quad \left. + (k-N) \left(\frac{N-1}{d_1^2} \frac{\mu_1}{\mu} - \left(\frac{N \left(1 - \frac{1-C}{N-\rho}\right)}{d_1^2} - \frac{C}{d_2^2} \right) \frac{\mu_1}{\mu} I - \left(1 - \frac{N \left(1 - \frac{1-C}{N-\rho}\right)}{d_1} + \frac{C}{d_2}\right) I \frac{\mu_1}{I} \frac{\partial I}{\partial \mu_1} \right) \right] \\ &= \frac{1}{\mu_1} \left[\left\{ \frac{N-\rho-1}{d_2} - \frac{(N-1)-\rho \left(1 + \frac{1-C}{N-\rho}\right)}{d_2} I + (k-N) \left(\frac{N-1}{d_1} - \left(\frac{N \left(1 - \frac{1-C}{N-\rho}\right)}{d_1} - \frac{C}{d_2} \right) I \right) \right. \right. \\ &\quad \left. \left. - (k-N) \left(\frac{\rho(N-1)}{d_1^2} - \frac{\rho N \left(1 - \frac{1-C}{N-\rho}\right)}{d_1^2} I \right) \right\} \cdot \left(\frac{1}{d_2} \frac{\mu_1}{\mu} \right) - \left\{ \left(2 - \frac{(N-1)-\rho \left(1 + \frac{1-C}{N-\rho}\right)}{d_2}\right) I \right. \right. \\ &\quad \left. \left. + (k-N) \left(1 - \frac{N \left(1 - \frac{1-C}{N-\rho}\right)}{d_1} + \frac{C}{d_2} \right) I \right\} \cdot \left(\frac{\mu_1}{I} \frac{\partial I}{\partial \mu_1} \right) \right]. \end{aligned} \tag{A.26}$$

Additionally, (A.25) can be alternatively written as

$$\begin{aligned} \frac{\mu_1}{I} \frac{\partial I}{\partial \mu_1} = & 2 - \frac{N - \rho - 1}{d_2} - \left(2 - \frac{(N - 1) - \rho \left(1 + \frac{1-C}{N-\rho} \right)}{d_2} \right) I \\ & + (k - N) \left(1 - \frac{N - 1}{d_1} - \left(1 - \frac{N \left(1 - \frac{1-C}{N-\rho} \right)}{d_1} + \frac{C}{d_2} \right) I \right), \end{aligned}$$

which implies

$$\begin{aligned} \frac{N - \rho - 1}{d_2} - \frac{(N - 1) - \rho \left(1 + \frac{1-C}{N-\rho} \right)}{d_2} I + (k - N) \left(\frac{N - 1}{d_1} - \left(\frac{N \left(1 - \frac{1-C}{N-\rho} \right)}{d_1} - \frac{C}{d_2} \right) I \right) \\ = (k - N + 2)(1 - I) - \left(\frac{\mu_1}{I} \frac{\partial I}{\partial \mu_1} \right). \end{aligned}$$

Substitution into (A.26) yields

$$\begin{aligned} \frac{\partial RHS(\mu_1, \mu)}{\partial \mu_1} = & \frac{1}{\mu_1} \left[\left\{ (k - N + 2)(1 - I) - \left(\frac{\mu_1}{I} \frac{\partial I}{\partial \mu_1} \right) - (k - N) \left(\frac{\rho(N - 1)}{d_1^2} - \frac{\rho N \left(1 - \frac{1-C}{N-\rho} \right)}{d_1^2} I \right) \right\} \left(\frac{1}{d_2} \frac{\mu_1}{\mu} \right) \right. \\ & \left. - \left\{ 2 - \frac{N - \rho - 1}{d_2} + (k - N) \left(1 - \frac{N - 1}{d_1} \right) - \frac{\mu_1}{I} \frac{\partial I}{\partial \mu_1} \right\} \cdot \left(\frac{\mu_1}{I} \frac{\partial I}{\partial \mu_1} \right) \right] \\ = & \frac{1}{\mu_1} \left[\frac{2(1 - I)}{d_2} \frac{\mu_1}{\mu} + (k - N) \left(1 - I - \frac{\rho}{d_1^2} \left(N - 1 - N \left(1 - \frac{1-C}{N-\rho} \right) I \right) \right) \frac{1}{d_2} \frac{\mu_1}{\mu} \right. \\ & \left. - \left(\frac{\mu_1}{I} \frac{\partial I}{\partial \mu_1} \right) \left(\left(\frac{2}{d_2} + \frac{k - N}{d_1} \right) \frac{\mu_1}{\mu} + 2 \right) + \left(\frac{\mu_1}{I} \frac{\partial I}{\partial \mu_1} \right) + \left(\frac{\mu_1}{I} \frac{\partial I}{\partial \mu_1} \right)^2 \right] \\ = & \frac{1}{\mu_1} \left[\frac{2(1 - I)}{d_2} \frac{\mu_1}{\mu} + \frac{k - N}{d_1} \frac{\mu_1}{\mu} \left((1 - I) + \left(\frac{1}{d_2} - \frac{1}{d_1} \right) \left(\frac{\mu_1}{\mu} (1 - I) - \left(\rho \left(\frac{1 - C}{N - \rho} \right) - C \right) I \right) \right) \right. \\ & \left. + \left(\frac{\mu_1}{I} \frac{\partial I}{\partial \mu_1} \right) \left(2 \left(\frac{\mu_1}{I} \frac{\partial I}{\partial \mu_1} \right) - 2 \left(1 + \frac{1}{d_2} \frac{\mu_1}{\mu} \right) - \frac{k - N}{d_1} \frac{\mu_1}{\mu} \right) + \left(\frac{\mu_1}{I} \frac{\partial I}{\partial \mu_1} \right) - \left(\frac{\mu_1}{I} \frac{\partial I}{\partial \mu_1} \right)^2 \right]. \end{aligned}$$

Hence,

$$\begin{aligned} \frac{\partial^2 I}{\partial \mu_1^2} = & \frac{I}{\mu_1^2} \left[\frac{2(1 - I)}{d_2} \frac{\mu_1}{\mu} + \frac{k - N}{d_1} \frac{\mu_1}{\mu} \left((1 - I) + \left(\frac{1}{d_2} - \frac{1}{d_1} \right) \left(\frac{\mu_1}{\mu} (1 - I) - \left(\rho \left(\frac{1 - C}{N - \rho} \right) - C \right) I \right) \right) \right. \\ & \left. + \left(\frac{\mu_1}{I} \frac{\partial I}{\partial \mu_1} \right) \left(2 \left(\frac{\mu_1}{I} \frac{\partial I}{\partial \mu_1} \right) - 2 \left(1 + \frac{1}{d_2} \frac{\mu_1}{\mu} \right) - \frac{k - N}{d_1} \frac{\mu_1}{\mu} \right) \right]. \end{aligned}$$

Substituting for $\frac{\mu_1}{I} \frac{\partial I}{\partial \mu_1}$ from (A.25), and after additional algebra, we obtain

$$\begin{aligned} \frac{\partial^2 I}{\partial \mu_1^2} = & \frac{I}{\mu_1^2} \left\{ -2I(1-I) + \frac{2}{d_2} \left[(1-2I) - \frac{\frac{\mu_1}{\mu} + \rho \left(\frac{1-C}{N-\rho} \right)}{d_2} I \right] \left[\frac{\mu_1}{\mu}(1-I) - \rho \left(\frac{1-C}{N-\rho} \right) I \right] \right. \\ & + \frac{k-N}{d_1} \left[2(1-2I) + \left(\frac{2}{d_2} - \frac{1}{d_1} \right) \frac{\mu_1}{\mu} - 4 \frac{\frac{\mu_1}{\mu} + \rho \left(\frac{1-C}{N-\rho} \right)}{d_2} I \right] \left[\frac{\mu_1}{\mu}(1-I) - \left(\rho \left(\frac{1-C}{N-\rho} \right) + \frac{\rho C}{d_2} \right) I \right] \\ & \left. + \left(\frac{k-N}{d_1} \right)^2 \left[\frac{\mu_1}{\mu} - 2 \left(\frac{\mu_1}{\mu} + \rho \left(\frac{1-C}{N-\rho} \right) + \frac{\rho C}{d_2} \right) I \right] \left[\frac{\mu_1}{\mu}(1-I) - \left(\rho \left(\frac{1-C}{N-\rho} \right) + \frac{\rho C}{d_2} \right) I \right] \right\}. \end{aligned} \quad (\text{A.27})$$

We further simplify $\left[\frac{\mu_1}{\mu}(1-I) - \rho \left(\frac{1-C}{N-\rho} \right) I \right]$ and $\left[\frac{\mu_1}{\mu}(1-I) - \left(\rho \left(\frac{1-C}{N-\rho} \right) + \frac{\rho C}{d_2} \right) I \right]$ as follows. Recall from (A.24):

$$\frac{C(N-\rho)}{d_2} \left(\frac{\rho}{d_1} \right)^{k-N} = 1 + \frac{N-\rho}{\rho} \left(1 - \frac{1}{I} \right) \frac{\mu_1}{\mu} + \left(1 - \frac{\mu_1}{\mu} \right) \frac{C}{d_2},$$

which implies

$$\frac{\mu_1}{\mu}(1-I) - \left(\rho \left(\frac{1-C}{N-\rho} \right) + \frac{\rho C}{d_2} \right) I = -\frac{\rho C}{d_2} \left(\frac{\rho}{d_1} \right)^{k-N} I, \quad (\text{A.28})$$

and furthermore,

$$\frac{\mu_1}{\mu}(1-I) - \rho \left(\frac{1-C}{N-\rho} \right) I = \frac{\rho C}{d_2} \left(1 - \left(\frac{\rho}{d_1} \right)^{k-N} \right) I. \quad (\text{A.29})$$

Substituting for $\left[\frac{\mu_1}{\mu}(1-I) - \rho \left(\frac{1-C}{N-\rho} \right) I \right]$ and $\left[\frac{\mu_1}{\mu}(1-I) - \left(\rho \left(\frac{1-C}{N-\rho} \right) + \frac{\rho C}{d_2} \right) I \right]$ in (A.27) using (A.28) and (A.29) respectively:

$$\begin{aligned} \frac{\partial^2 I}{\partial \mu_1^2} = & \frac{I^2}{\mu_1^2} \left\{ -2(1-I) + \frac{2\rho C}{d_2^2} \left[1 - \left(\frac{\rho}{d_1} \right)^{k-N} \right] \left[(1-2I) - \frac{\frac{\mu_1}{\mu} + \rho \left(\frac{1-C}{N-\rho} \right)}{d_2} I \right] \right. \\ & \left. - \frac{\rho C}{d_2} \frac{k-N}{d_1} \left(\frac{\rho}{d_1} \right)^{k-N} \left[2(1-2I) + \left(\frac{2}{d_2} - \frac{1}{d_1} \right) \frac{\mu_1}{\mu} - 4 \frac{\frac{\mu_1}{\mu} + \rho \left(\frac{1-C}{N-\rho} \right)}{d_2} I \right] \right\} \end{aligned}$$

$$-\rho C \left(\frac{k-N}{d_1} \right)^2 \left(\frac{\rho}{d_1} \right)^{k-N} \left[\frac{1}{d_2} \frac{\mu_1}{\mu} - 2 \frac{\frac{\mu_1}{\mu} + \rho \left(\frac{1-C}{N-\rho} \right)}{d_2} I - \frac{2\rho C}{d_2^2} I \right] \},$$

which establishes (A.17). ■

A.4.1.2 Proof of Corollary 2

Idle time (A.18): From Lemma 2, substituting $\mu_1 = \mu$ into (1.1) yields

$$\begin{aligned} I(\mu, \mu) &= \left(1 + \rho \left(\frac{1 - ErlC(N, \rho)}{N - \rho} + \left(1 - \left(\frac{\rho}{N} \right)^{k-N} \right) \frac{ErlC(N, \rho)}{N - \rho} \right) \right)^{-1} \\ &= \left(1 + \rho \left(\frac{1 - ErlC(N, \rho)}{N - \rho} \right) + ErlC(N, \rho) \frac{\rho \left(1 - \left(\frac{\rho}{N} \right)^{k-N} \right)}{N - \rho} \right)^{-1} \\ &= \left(1 + \rho \left(\frac{1 - ErlC(N, \rho)}{N - \rho} \right) + ErlC(N, \rho) \sum_{i=1}^{k-N} \left(\frac{\rho}{N} \right)^i \right)^{-1}. \end{aligned} \quad (\text{A.30})$$

Idle time (A.19): Using the relationship between $ErlB$ and $ErlC$ (from Lemma 19 (a)), note that

$$\begin{aligned} 1 + \rho \left(\frac{1 - ErlC(N, \rho)}{N - \rho} \right) &= 1 + \rho \frac{1 - \left(\frac{N}{\frac{N-\rho}{ErlB(N,\rho)} + \rho} \right)}{N - \rho} = 1 + \rho \frac{\frac{1}{ErlB(N,\rho)} - 1}{\frac{N-\rho}{ErlB(N,\rho)} + \rho} \\ &= 1 + \rho \frac{1 - ErlB(N, \rho)}{N - (1 - ErlB(N, \rho))\rho} = \frac{N}{N - (1 - ErlB(N, \rho))\rho} = \frac{N}{\frac{N-\rho}{ErlB(N,\rho)} + \rho} \frac{1}{ErlB(N, \rho)} = \frac{ErlC(N, \rho)}{ErlB(N, \rho)}. \end{aligned}$$

Substitution into (A.30) using the above display and by definition of $ErlB(N, \rho)$ from (A.1):

$$\begin{aligned} I(\mu, \mu) &= \left(\left(\frac{1}{ErlB(N, \rho)} + \sum_{i=1}^{k-N} \left(\frac{\rho}{N} \right)^i \right) ErlC(N, \rho) \right)^{-1} \\ &= \left(\left(1 + \sum_{i=0}^{N-1} \frac{N!}{i!} \left(\frac{\mu}{\lambda} \right)^{N-i} + \sum_{i=1}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i \right) ErlC \left(N, \frac{\lambda}{\mu} \right) \right)^{-1}. \end{aligned}$$

Idle time (A.20): From Lemma 2, substituting $\mu_1 = \mu$ into (1.1) yields

$$\begin{aligned} I(\mu, \mu) &= \left(1 + \rho \left(\frac{1 - ErlC(N, \rho)}{N - \rho} + \left(1 - \left(\frac{\rho}{N} \right)^{k-N} \right) \frac{ErlC(N, \rho)}{N - \rho} \right) \right)^{-1} \\ &= \left(1 + \rho \left(\frac{1}{N - \rho} - \frac{ErlC(N, \rho) \left(\frac{\rho}{N} \right)^{k-N}}{N - \rho} \right) \right)^{-1} \\ &= \left(1 - \frac{\rho}{N} \right) \left(1 - ErlC(N, \rho) \left(\frac{\rho}{N} \right)^{k-N+1} \right)^{-1}. \end{aligned}$$

First-order derivative (A.21): From Lemma 22, substituting $\mu_1 = \mu$ into (A.16) yields

$$\begin{aligned} \frac{\partial I(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} &= \frac{I(\mu, \mu)^2}{\mu} \left(\frac{1}{I(\mu, \mu)} - 1 + \frac{ErlC(N, \rho)}{N - \rho} \left(\left(1 - \left(\frac{\rho}{N} \right)^{k-N} \right) \frac{\rho}{N - \rho} - (k - N) \left(\frac{\rho}{N} \right)^{k-N} \right) \right. \\ &\quad \left. + \frac{k - N}{N} \left(\frac{\rho}{N} \right)^{k-N} ErlC(N, \rho) \right) \\ &= \frac{I(\mu, \mu)}{\mu} \left(1 - I(\mu, \mu) + I(\mu, \mu) \cdot ErlC(N, \rho) \left(\frac{\rho}{(N - \rho)^2} - \frac{\rho}{(N - \rho)^2} \left(\frac{\rho}{N} \right)^{k-N} - (k - N) \left(\frac{\rho}{N} \right)^{k-N} \frac{\rho}{N(N - \rho)} \right) \right). \end{aligned}$$

Note that

$$\begin{aligned} \sum_{i=1}^{k-N} i \left(\frac{\rho}{N} \right)^i &= \frac{\rho}{N} \frac{(k - N) \left(\frac{\rho}{N} \right)^{k-N+1} - (k - N + 1) \left(\frac{\rho}{N} \right)^{k-N} + 1}{\left(\frac{\rho}{N} - 1 \right)^2} \\ &= N \left(\frac{\rho}{(N - \rho)^2} - \frac{\rho}{(N - \rho)^2} \left(\frac{\rho}{N} \right)^{k-N} - \frac{k - N}{N - \rho} \left(\frac{\rho}{N} \right)^{k-N+1} \right), \end{aligned}$$

and thus $\frac{\partial I(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu}$ can be equivalently written as

$$\begin{aligned} \frac{\partial I(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} &= \frac{I(\mu, \mu)}{\mu} \left(1 - I(\mu, \mu) + I(\mu, \mu) \cdot ErlC(N, \rho) \cdot \frac{1}{N} \sum_{i=1}^{k-N} i \left(\frac{\rho}{N} \right)^i \right) \\ &= \frac{1}{\mu} I(\mu, \mu) (1 - I(\mu, \mu)) + I(\mu, \mu)^2 \frac{ErlC(N, \rho)}{N \mu} \sum_{i=1}^{k-N} i \left(\frac{\rho}{N} \right)^i. \end{aligned}$$

First-order derivative (A.22): From (A.16) in Lemma 22, letting $\mu_1 = \mu$ and substi-

tuting for $1 - I(\mu, \mu)$ using (A.18) in Corollary 2:

$$\begin{aligned}
& \frac{\partial I(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} \\
&= \frac{I^2(\mu, \mu)}{\mu} \left[\rho \frac{1 - ErlC(N, \rho)}{N - \rho} + \left(1 - \left(\frac{\rho}{N}\right)^{k-N}\right) \frac{\rho ErlC(N, \rho)}{N - \rho} \right] \\
&\quad + \frac{I^2(\mu, \mu)}{\mu} \left[\frac{ErlC(N, \rho)}{N - \rho} \left(1 - \left(\frac{\rho}{N}\right)^{k-N}\right) \frac{\rho}{N - \rho} - \frac{ErlC(N, \rho)}{N - \rho} (k - N) \left(\frac{\rho}{N}\right)^{k-N} \right. \\
&\quad \left. + \frac{k - N}{N} \left(\frac{\rho}{N}\right)^{k-N} ErlC(N, \rho) \right] \\
&= \frac{I^2(\mu, \mu)}{\mu} \left[\frac{\rho (1 - ErlC(N, \rho))}{N - \rho} + \frac{\rho ErlC(N, \rho)}{N - \rho} - \frac{\rho ErlC(N, \rho)}{N - \rho} \left(\frac{\rho}{N}\right)^{k-N} + \frac{\rho ErlC(N, \rho)}{(N - \rho)^2} \right. \\
&\quad \left. - \frac{\rho ErlC(N, \rho)}{(N - \rho)^2} \left(\frac{\rho}{N}\right)^{k-N} - \frac{ErlC(N, \rho)}{N - \rho} (k - N) \left(\frac{\rho}{N}\right)^{k-N} + \frac{k - N}{N} \left(\frac{\rho}{N}\right)^{k-N} ErlC(N, \rho) \right] \\
&= \frac{I^2(\mu, \mu)}{\mu} \left[\frac{\rho}{N - \rho} + \frac{\rho ErlC(N, \rho)}{(N - \rho)^2} - \frac{\rho ErlC(N, \rho)}{N - \rho} \left(\frac{\rho}{N}\right)^{k-N} \left(1 + \frac{1}{N - \rho} + \frac{k - N}{N}\right) \right] \\
&= \frac{\rho}{(N - \rho)^2} \frac{I^2(\mu, \mu)}{\mu} \left[N - \rho + ErlC(N, \rho) \left(1 - \left(\frac{\rho}{N}\right)^{k-N} \left(1 + \frac{k}{N}(N - \rho)\right)\right) \right].
\end{aligned}$$

Substituting for $I(\mu, \mu)$ using (A.20) in the above display yields

$$\begin{aligned}
& \frac{\partial I(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} = \frac{\rho}{(N - \rho)^2} \frac{1}{\mu} \frac{\left(1 - \frac{\rho}{N}\right)^2}{\left(1 - \left(\frac{\rho}{N}\right)^{k-N+1} ErlC(N, \rho)\right)^2} \\
&\quad \left[N - \rho + ErlC(N, \rho) \left(1 - \left(\frac{\rho}{N}\right)^{k-N} \left(1 + \frac{k}{N}(N - \rho)\right)\right) \right] \\
&= \frac{\rho}{N \mu} \frac{1 - \frac{\rho}{N} + ErlC(N, \rho) \left(\frac{1}{N} - \left(\frac{\rho}{N}\right)^{k-N} \left(\frac{1}{N} + \frac{k}{N} \left(1 - \frac{\rho}{N}\right)\right)\right)}{\left(1 - \left(\frac{\rho}{N}\right)^{k-N+1} ErlC(N, \rho)\right)^2}.
\end{aligned}$$

■

A.4.1.3 Proof of Lemma 23

(a): From (A.20) in Corollary 2,

$$\begin{aligned}
I(\mu, \mu) &= \frac{1 - \frac{\lambda}{N\mu}}{1 - ErlC\left(N, \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{N\mu}\right)^{k-N+1}} \\
&\stackrel{(i)}{=} \frac{1 - \frac{\lambda}{N\mu}}{1 - ErlC\left(N, \frac{\lambda}{\mu}\right) \left[\left(\frac{\lambda}{N\mu}\right) - \left(1 - \frac{\lambda}{N\mu}\right) \sum_{i=1}^{k-N} \left(\frac{\lambda}{N\mu}\right)^i\right]} \\
&= \frac{ErlB\left(N, \frac{\lambda}{\mu}\right)}{ErlB\left(N, \frac{\lambda}{\mu}\right) \frac{1 - ErlC\left(N, \frac{\lambda}{\mu}\right) \frac{\lambda}{N\mu}}{1 - \frac{\lambda}{N\mu}} + ErlB\left(N, \frac{\lambda}{\mu}\right) ErlC\left(N, \frac{\lambda}{\mu}\right) \sum_{i=1}^{k-N} \left(\frac{\lambda}{N\mu}\right)^i} \\
&\stackrel{(ii)}{=} \frac{ErlB\left(N, \frac{\lambda}{\mu}\right)}{ErlC\left(N, \frac{\lambda}{\mu}\right) \left(1 + ErlB\left(N, \frac{\lambda}{\mu}\right) \sum_{i=1}^{k-N} \left(\frac{\lambda}{N\mu}\right)^i\right)},
\end{aligned}$$

where (i) follows from the finite summation formula, and (ii) follows from the relationship between $ErlB$ and $ErlC$ (from Lemma 19 (a)). Substitution for $I(\mu, \mu)$ using the above display yields

$$\begin{aligned}
&I(\mu, \mu)^2 ErlC\left(N, \frac{\lambda}{\mu}\right) \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu}\right)^i \\
&= \left(\frac{ErlB\left(N, \frac{\lambda}{\mu}\right)}{ErlC\left(N, \frac{\lambda}{\mu}\right) \left(1 + ErlB\left(N, \frac{\lambda}{\mu}\right) \sum_{i=1}^{k-N} \left(\frac{\lambda}{N\mu}\right)^i\right)} \right)^2 ErlC\left(N, \frac{\lambda}{\mu}\right) \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu}\right)^i \\
&= \frac{ErlB\left(N, \frac{\lambda}{\mu}\right)^2}{ErlC\left(N, \frac{\lambda}{\mu}\right)} \frac{\sum_{i=0}^{k-N} i \left(\frac{\lambda}{N\mu}\right)^i}{\left(1 - ErlB\left(N, \frac{\lambda}{\mu}\right) + ErlB\left(N, \frac{\lambda}{\mu}\right) \sum_{i=0}^{k-N} \left(\frac{\lambda}{N\mu}\right)^i\right)^2} \\
&\stackrel{(iii)}{\leq} \frac{1}{ErlC\left(N, \frac{\lambda}{\mu}\right)} \frac{\sum_{i=0}^{k-N} i \left(\frac{\lambda}{N\mu}\right)^i}{\left(\sum_{i=0}^{k-N} \left(\frac{\lambda}{N\mu}\right)^i\right)^2} \\
&\stackrel{(iv)}{<} \frac{1}{ErlC\left(N, \frac{\lambda}{\mu}\right)}, \tag{A.31}
\end{aligned}$$

where (iii) follows by noting that $ErlB\left(N, \frac{\lambda}{\mu}\right) \leq 1$ and (iv) follows because

$$\begin{aligned}
& \frac{\sum_{i=0}^{k-N} i \left(\frac{\lambda}{N\mu}\right)^i}{\left(\sum_{i=0}^{k-N} \left(\frac{\lambda}{N\mu}\right)^i\right)^2} = \frac{\sum_{i=0}^{k-N} i \left(\frac{\lambda}{N\mu}\right)^i}{\sum_{i=0}^{k-N} \sum_{j=0}^{k-N} \left(\frac{\lambda}{N\mu}\right)^{i+j}} = \frac{\sum_{i=0}^{k-N} i \left(\frac{\lambda}{N\mu}\right)^i}{\sum_{\ell=0}^{2(k-N)} \sum_{i=\max\{0, \ell-(k-N)\}}^{\min\{\ell, k-N\}} \left(\frac{\lambda}{N\mu}\right)^\ell} \\
&= \frac{\sum_{i=0}^{k-N} i \left(\frac{\lambda}{N\mu}\right)^i}{\sum_{\ell=0}^{k-N} \sum_{i=0}^{\ell} \left(\frac{\lambda}{N\mu}\right)^\ell + \sum_{\ell=k-N+1}^{2(k-N)} \sum_{i=\ell-(k-N)}^{k-N} \left(\frac{\lambda}{N\mu}\right)^\ell} \\
&= \frac{\sum_{i=0}^{k-N} i \left(\frac{\lambda}{N\mu}\right)^i}{\sum_{\ell=0}^{k-N} (\ell+1) \left(\frac{\lambda}{N\mu}\right)^\ell + \sum_{\ell=k-N+1}^{2(k-N)} (2(k-N)-\ell) \left(\frac{\lambda}{N\mu}\right)^\ell} \tag{A.33} \\
&< 1,
\end{aligned}$$

noting that $\sum_{i=0}^{k-N} i \left(\frac{\lambda}{N\mu}\right)^i < \sum_{\ell=0}^{k-N} (\ell+1) \left(\frac{\lambda}{N\mu}\right)^\ell$.

Case (I): When $0 < \mu \leq \frac{\lambda}{N}$, i.e., $\frac{\lambda}{N\mu} \geq 1$, we have $ErlC\left(N, \frac{\lambda}{\mu}\right) \geq 1$ (from Lemma 19 (b)).

Then, from (A.31),

$$\begin{aligned}
I(\mu, \mu)^2 ErlC\left(N, \frac{\lambda}{\mu}\right) \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu}\right)^i &\leq \frac{1}{ErlC\left(N, \frac{\lambda}{\mu}\right)} \frac{\sum_{i=0}^{k-N} i \left(\frac{\lambda}{N\mu}\right)^i}{\left(\sum_{i=0}^{k-N} \left(\frac{\lambda}{N\mu}\right)^i\right)^2} \\
&\leq \frac{\sum_{i=0}^{k-N} i \left(\frac{\lambda}{N\mu}\right)^i}{\left(\sum_{i=0}^{k-N} \left(\frac{\lambda}{N\mu}\right)^i\right)^2} \\
&\stackrel{(*)}{\leq} \frac{\sum_{i=0}^{k-N} i \left(\frac{\lambda}{N\frac{\lambda}{N}}\right)^i}{\left(\sum_{i=0}^{k-N} \left(\frac{\lambda}{N\frac{\lambda}{N}}\right)^i\right)^2} \\
&= \frac{1}{2} \frac{k-N}{k-N+1} < \frac{1}{2} < 2\sqrt{N}, \quad \forall k \geq N \geq 2,
\end{aligned}$$

where (*) follows from the next claim, whose proof appears at the end.

Claim 2. For any $\lambda > 0$ and $k \geq N \geq 2$, $\frac{\sum_{i=0}^{k-N} i \left(\frac{\lambda}{N\mu}\right)^i}{\left(\sum_{i=0}^{k-N} \left(\frac{\lambda}{N\mu}\right)^i\right)^2}$ is increasing in μ for $\mu \in (0, \frac{\lambda}{N}]$.

Case (II): When $\mu > \frac{\lambda}{N}$, i.e., $0 < \frac{\lambda}{N\mu} < 1$, we have $0 < ErlC(N, \frac{\lambda}{\mu}) < 1$ (from Lemma 19 (b)). Using the finite summation formula,

$$\begin{aligned} & I(\mu, \mu)^2 ErlC\left(N, \frac{\lambda}{\mu}\right) \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu}\right)^i \\ &= I(\mu, \mu)^2 ErlC\left(N, \frac{\lambda}{\mu}\right) \frac{\frac{\lambda}{N\mu}}{\left(1 - \frac{\lambda}{N\mu}\right)^2} \left[1 - \left(\frac{\lambda}{N\mu}\right)^{k-N} \left(1 + (k-N)\left(1 - \frac{\lambda}{N\mu}\right)\right) \right] \\ &\leq ErlC\left(N, \frac{\lambda}{\mu}\right) \frac{\frac{\lambda}{N\mu}}{\left(1 - \frac{\lambda}{N\mu}\right)^2}, \end{aligned} \tag{A.34}$$

where the last inequality follows because $I(\mu, \mu) \leq 1$ and $1 - \left(\frac{\lambda}{N\mu}\right)^{k-N} \left(1 + (k-N)\left(1 - \frac{\lambda}{N\mu}\right)\right) < 1$.

Combining (A.32) and (A.34) yields

$$\begin{aligned} I(\mu, \mu)^2 ErlC\left(N, \frac{\lambda}{\mu}\right) \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu}\right)^i &\leq \min \left\{ \frac{1}{ErlC\left(N, \frac{\lambda}{\mu}\right)}, ErlC\left(N, \frac{\lambda}{\mu}\right) \frac{\frac{\lambda}{N\mu}}{\left(1 - \frac{\lambda}{N\mu}\right)^2} \right\} \\ &= 2\sqrt{N} \cdot \min \left\{ \frac{1}{2\sqrt{N}ErlC\left(N, \frac{\lambda}{\mu}\right)}, \frac{ErlC\left(N, \frac{\lambda}{\mu}\right)}{2\sqrt{N}} \frac{\frac{\lambda}{N\mu}}{\left(1 - \frac{\lambda}{N\mu}\right)^2} \right\} \\ &\stackrel{(*)}{<} 2\sqrt{N}, \end{aligned}$$

where $(*)$ follows from the next claim, whose proof appears at the end.

Claim 3. *The following holds for all λ , $k \geq N \geq 2$ and $\mu > \frac{\lambda}{N}$:*

$$\min \left\{ \frac{1}{2\sqrt{N}ErlC\left(N, \frac{\lambda}{\mu}\right)}, \frac{ErlC\left(N, \frac{\lambda}{\mu}\right)}{2\sqrt{N}} \frac{\frac{\lambda}{N\mu}}{\left(1 - \frac{\lambda}{N\mu}\right)^2} \right\} < 1.$$

Therefore, together the above two cases establish that

$$I(\mu, \mu)^2 ErlC\left(N, \frac{\lambda}{\mu}\right) \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu}\right)^i < 2\sqrt{N}.$$

(b): From (A.31),

$$\begin{aligned}
I(\mu, \mu)^2 ErlC \left(N, \frac{\lambda}{\mu} \right) \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^{i+1} &\leq \frac{1}{ErlC \left(N, \frac{\lambda}{\mu} \right)} \frac{\sum_{i=0}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^{i+1}}{\left(\sum_{i=0}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i \right)^2} \\
&\stackrel{(*)}{<} \frac{1}{ErlC \left(N, \frac{\lambda}{\mu} \right)}, \tag{A.35}
\end{aligned}$$

where (*) follows by noting, from (A.33), that

$$\begin{aligned}
\frac{\sum_{i=0}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^{i+1}}{\left(\sum_{i=0}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i \right)^2} &= \frac{\sum_{i=0}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^{i+1}}{\sum_{\ell=0}^{k-N} (\ell+1) \left(\frac{\lambda}{N\mu} \right)^\ell + \sum_{\ell=k-N+1}^{2(k-N)} (2(k-N)-\ell) \left(\frac{\lambda}{N\mu} \right)^\ell} \\
&= \frac{\sum_{i=0}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^{i+1}}{\sum_{\ell=0}^{k-N} (\ell+1) \left(\frac{\lambda}{N\mu} \right)^\ell + \left(\frac{\lambda}{N\mu} \right)^{k-N+1} \sum_{\ell=0}^{k-N} (k-N-\ell) \left(\frac{\lambda}{N\mu} \right)^\ell} \\
&= \frac{\sum_{i=0}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^{i+1}}{1 + \sum_{i=0}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^{i+1} + 2 \sum_{i=1}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i + \left(\frac{\lambda}{N\mu} \right)^{k-N+1} \sum_{i=1}^{k-N} (k-N-i) \left(\frac{\lambda}{N\mu} \right)^i} < 1.
\end{aligned}$$

Case (I): When $0 < \mu \leq \frac{\lambda}{N}$, i.e., $\frac{\lambda}{N\mu} \geq 1$, we have $ErlC \left(N, \frac{\lambda}{\mu} \right) \geq 1$ (from Lemma 19 (b)).

Then, from (A.35),

$$I(\mu, \mu)^2 ErlC \left(N, \frac{\lambda}{\mu} \right) \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^{i+1} \leq \frac{1}{ErlC \left(N, \frac{\lambda}{\mu} \right)} \leq 1 < 2\sqrt{N}, \quad \forall k \geq N \geq 2,$$

Case (II): When $\mu > \frac{\lambda}{N}$, i.e., $0 < \frac{\lambda}{N\mu} < 1$, we have $0 < ErlC \left(N, \frac{\lambda}{\mu} \right) < 1$ (from

Lemma 19 (b)). Using the finite summation formula,

$$\begin{aligned}
& I(\mu, \mu)^2 \text{ErlC} \left(N, \frac{\lambda}{\mu} \right) \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^{i+1} \\
& = I(\mu, \mu)^2 \text{ErlC} \left(N, \frac{\lambda}{\mu} \right) \frac{\frac{\lambda}{N\mu}}{\left(1 - \frac{\lambda}{N\mu} \right)^2} \frac{\lambda}{N\mu} \left[1 - \left(\frac{\lambda}{N\mu} \right)^{k-N} \left(1 + (k-N) \left(1 - \frac{\lambda}{N\mu} \right) \right) \right] \\
& \leq \text{ErlC} \left(N, \frac{\lambda}{\mu} \right) \frac{\frac{\lambda}{N\mu}}{\left(1 - \frac{\lambda}{N\mu} \right)^2},
\end{aligned} \tag{A.36}$$

where the last inequality follows because $I(\mu, \mu) \leq 1$, $\frac{\lambda}{N\mu} < 1$ and

$$1 - \left(\frac{\lambda}{N\mu} \right)^{k-N} \left(1 + (k-N) \left(1 - \frac{\lambda}{N\mu} \right) \right) < 1.$$

Combining (A.35) and (A.36) yields

$$\begin{aligned}
I(\mu, \mu)^2 \text{ErlC} \left(N, \frac{\lambda}{\mu} \right) \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^{i+1} & \leq \min \left\{ \frac{1}{\text{ErlC} \left(N, \frac{\lambda}{\mu} \right)}, \text{ErlC} \left(N, \frac{\lambda}{\mu} \right) \frac{\frac{\lambda}{N\mu}}{\left(1 - \frac{\lambda}{N\mu} \right)^2} \right\} \\
& = 2\sqrt{N} \cdot \min \left\{ \frac{1}{2\sqrt{N} \text{ErlC} \left(N, \frac{\lambda}{\mu} \right)}, \frac{\text{ErlC} \left(N, \frac{\lambda}{\mu} \right)}{2\sqrt{N}} \frac{\frac{\lambda}{N\mu}}{\left(1 - \frac{\lambda}{N\mu} \right)^2} \right\} \\
& \stackrel{(*)}{<} 2\sqrt{N},
\end{aligned}$$

where $(*)$ follows from Claim 3. ■

Proof of Claim 2

Let $x := \frac{\lambda}{N\mu}$ and define $f(x) := \frac{\sum_{i=0}^k ix^i}{\left(\sum_{i=0}^k x^i \right)^2}$ for $x \geq 0$. It suffices to show $f(x)$ is decreasing in $x \in [1, \infty)$. Using the finite summation formula in $f(x)$ yields

$$f(x) = \frac{x + kx^{k+2} - (k+1)x^{k+1}}{(1 - x^{k+1})^2}.$$

Consider the first-order derivative of the function $f(x)$:

$$f'(x) = \frac{1 - (k+1)^2 x^k + (k^2 + 4k + 1)x^{k+1} - (k+1)^2 x^{2k+1} + k^2 x^{2(k+1)}}{(1 - x^{k+1})^3} =: \frac{g(x)}{(1 - x^{k+1})^3}.$$

We first evaluate $f'(x)$ at $x = 1$ by applying L'Hôpital's rule for three times and algebra:

$$\begin{aligned} f'(1) &= \left. \frac{[1 - (k+1)^2 x^k + (k^2 + 4k + 1)x^{k+1} - (k+1)^2 x^{2k+1} + k^2 x^{2(k+1)}]'''}{[(1 - x^{k+1})^3]'''} \right|_{x=1} \\ &= \frac{k(k+1)^2(k-1)}{-k(k+1)^3} = -\frac{k(k-1)}{k(k+1)} \leq 0, \quad \forall k \geq 1. \end{aligned}$$

This means that $f(x)$ is decreasing at $x = 1$. In what follows, we focus on $x \in (1, \infty)$.

Consider the first-order derivative of the function $g(x)$:

$$\begin{aligned} g'(x) &= (k+1)x^{k-1} \left[(2k^2 x^{k+1} + k(k+1))(x-1) - (3k+1)x(x^k - 1) \right] \\ &= (k+1)x^{k-1}(x-1) \left[2k^2 x^{k+1} - (3k+1) \sum_{i=1}^k x^i + k(k+1) \right] \\ &= : (k+1)x^{k-1}(x-1) \cdot h(x), \end{aligned} \tag{A.37}$$

where

$$h(x) := \begin{cases} 2x^2 - 4x + 2, & \text{if } k = 1, \\ 2k^2 x^{k+1} - (3k+1) \sum_{i=1}^k x^i + k(k+1), & \text{if } k \geq 2. \end{cases}$$

If $k = 1$, then it is clear that $h(x) > 0$ for $x > 1$. If $k \geq 2$, then consider the first-order derivative of the function $h(x)$:

$$h'(x) = 2k^2(k+1)x^k - (3k+1) \left(1 + \sum_{i=2}^k i x^{i-1} \right),$$

which changes the sign of the coefficients exactly once. Descartes' rule of signs for polynomials implies that $h'(x)$ has at most one positive real root. Note that $h'(0) = -(3k+1) < 0$

and $h'(1) = \frac{k(k-1)(k+1)}{2} > 0$ for all $k \geq 2$, then the intermediate value theorem implies that $h'(x)$ has at least one root in $(0, 1)$. Therefore, $h'(x)$ has one unique root, denoted by $x_0 \in (0, 1)$, with $h'(x) < 0$ for all $x \in (0, x_0)$ and $h'(x) > 0$ for all $x \in (x_0, \infty)$. This implies that $h(x)$ is strictly increasing in x for $x \in (1, \infty)$. Note that $h(1) = 0$, thus $h(x) > h(1) = 0$ for all $x > 1$.

Then, from (A.37), $h(x) > 0$ for all $x > 1$ implies that $g'(x) > 0$ for all $x > 1$, which means that the function $g(x)$ is strictly increasing in x for $(1, \infty)$. Note that $g(1) = 0$, thus $g(x) > g(0) = 0$ for all $x > 1$, implying that $f'(x) < 0$ for all $x > 1$ (since $1 - x^{k+1} < 0$ for $x > 1$); that is, $f(x)$ is strictly decreasing in x for $x \in (1, \infty)$.

Therefore, we conclude that $f(x)$ is decreasing in $[1, \infty)$. ■

Proof of Claim 3

Let $f(x) := \frac{1}{ErlC(N, Nx)}$ and $g(x) := \frac{x \cdot ErlC(N, Nx)}{(1-x)^2}$ for $x \in (0, 1)$. From Lemma 18 (b) and Lemma 19 (b), it is clear that $f(x)$ is strictly decreasing in x with $\lim_{x \rightarrow 0} f(x) = \infty$ and $\lim_{x \rightarrow 1} f(x) = 1$; and $g(x)$ is strictly increasing in x with $\lim_{x \rightarrow 0} g(x) = 0$ and $\lim_{x \rightarrow 1} g(x) = \infty$. This implies that $f(x)$ and $g(x)$ intersect once in $(0, 1)$. We denote the unique solution to $f(x) = g(x)$ by $x^* \in (0, 1)$, i.e.,

$$ErlC(N, Nx^*) = \frac{1}{\sqrt{x^*}} - \sqrt{x^*}, \quad (\text{A.38})$$

and it is straightforward that $\min\{f(x), g(x) : x \in (0, 1)\} \leq f(x^*) = g(x^*)$. Thus, to prove Claim 3, it suffices to show that

$$\frac{f(x^*)}{2\sqrt{N}} < 1, \text{ or, equivalently, } 2\sqrt{N}ErlC(N, Nx^*) > 1.$$

Substituting for $ErlC(N, Nx^*)$ using (A.38) into the above display shows

$$\frac{1}{\sqrt{x^*}} - \sqrt{x^*} > \frac{1}{2\sqrt{N}},$$

which after algebra implies

$$\sqrt{x^*} < \sqrt{1 + \frac{1}{16N}} - \frac{1}{4\sqrt{N}} =: \sqrt{\bar{x}}. \quad (\text{A.39})$$

To show (A.39) is true, we introduce an auxiliary function $h(x) := ErlC(N, Nx) - \frac{1}{\sqrt{x}} + \sqrt{x}$ for $x \in (0, 1)$. It is clear that $h(x)$ is strictly increasing in x (from Lemma 18 (b)). Thus, showing $\bar{x} > x^*$ (i.e., (A.39)) is equivalent to showing $h(\bar{x}) > h(x^*)$. Note that, by definition of x^* in (A.38), $h(x^*) = 0$, it suffices to show $h(\bar{x}) > 0$, i.e.,

$$ErlC(N, N\bar{x}) - \frac{1}{\sqrt{\bar{x}}} + \sqrt{\bar{x}} > 0,$$

or, equivalently,

$$ErlC(N, N\bar{x}) > \frac{1}{\sqrt{\bar{x}}} - \sqrt{\bar{x}} \stackrel{(*)}{=} \frac{1}{2\sqrt{N}}, \quad (\text{A.40})$$

where $(*)$ follows by definition of \bar{x} . The remainder of the proof is devoted to verifying (A.40). To show (A.40), we consider $N = 1$, $N = 2$ and $N \geq 3$ separately.

- When $N = 1$, $\bar{x} = 0.7808$ and (A.40) becomes to $ErlC(1, 0.7808) > \frac{1}{2}$, which is true.
- When $N = 2$, $\bar{x} = 0.8387$ and (A.40) becomes to $ErlC(2, 2 \times 0.8387) > \frac{1}{2\sqrt{2}}$, which is true.
- When $N \geq 3$, from Proposition 2 in Harel (2010),

$$ErlC(N, N\bar{x}) > 1 - (1 - \bar{x}^2)\sqrt{\frac{\pi N}{8}} > 1 - (1 - \bar{x}^2)\sqrt{\frac{N}{2}}.$$

To verify (A.40) stands, it is sufficient to show

$$1 - (1 - \bar{x}^2)\sqrt{\frac{N}{2}} > \frac{1}{2\sqrt{N}},$$

which after algebra implies

$$\bar{x}^2 > 1 - \sqrt{\frac{2}{N}} + \frac{1}{\sqrt{2}N}.$$

Plugging in the expression for \bar{x} from (A.39):

$$\left(\sqrt{1 + \frac{1}{16N}} - \frac{1}{4\sqrt{N}} \right)^2 > 1 - \sqrt{\frac{2}{N}} + \frac{1}{\sqrt{2}N},$$

which can be equivalently written as

$$\sqrt{N} \left(\sqrt{2} - \frac{1}{2} \sqrt{1 + \frac{1}{16N}} \right) > \frac{1}{\sqrt{2}} - \frac{1}{8}. \quad (\text{A.41})$$

It is clear that the left-hand side of the above display is strictly increasing in N and the right-hand side of the above display is a constant. Note that, when $N = 3$, the left-hand side evaluates to 1.5745, which is greater than the right-hand side $\frac{1}{\sqrt{2}} - \frac{1}{8} = 0.5821$.

Hence, (A.41) is true for all $N \geq 3$.

Therefore, combining the analysis in the above three bullet points establishes that (A.40) holds for all $N \geq 1$, and thus the Claim is proved. ■

A.4.1.4 Proof of Corollary 3

From (A.21) in Corollary 2,

$$\mu \frac{\partial I(\mu_1, \mu)}{\partial \mu_1} \Bigg|_{\mu_1=\mu} = I(\mu, \mu)(1 - I(\mu, \mu)) + I(\mu, \mu)^2 \frac{ErlC(N, \frac{\lambda}{\mu})}{N} \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^i.$$

(a): Using the bound in Lemma 23 (a),

$$\mu \frac{\partial I(\mu_1, \mu)}{\partial \mu_1} \Bigg|_{\mu_1=\mu} < I(\mu, \mu)(1 - I(\mu, \mu)) + \frac{2\sqrt{N}}{N} = I(\mu, \mu)(1 - I(\mu, \mu)) + \frac{2}{\sqrt{N}}, \quad \forall \mu > 0.$$

(b): Using the bound in Lemma 23 (b),

$$\mu \frac{\partial I(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} < I(\mu, \mu)(1 - I(\mu, \mu)) + \frac{1}{N} \frac{N\mu}{\lambda} \cdot 2\sqrt{N} < I(\mu, \mu) + \frac{2\mu\sqrt{N}}{\lambda},$$

which implies that

$$\frac{\partial I(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} < \frac{I(\mu, \mu)}{\mu} + \frac{2\sqrt{N}}{\lambda}, \quad \forall \mu > 0.$$

■

A.4.1.5 Proof of Lemma 24

From (A.18) in Corollary 2,

$$\frac{1}{I(\mu, \mu)} = 1 + \frac{\lambda}{\mu} \left(\frac{1 - ErlC(N, \frac{\lambda}{\mu})}{N - \frac{\lambda}{\mu}} \right) + ErlC(N, \frac{\lambda}{\mu}) \sum_{i=1}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i. \quad (\text{A.42})$$

Differentiating the above display on both sides yields

$$\begin{aligned} & -\frac{1}{I(\mu, \mu)^2} \frac{dI(\mu, \mu)}{d\mu} \\ &= \left(-\frac{\lambda}{\mu^2} \right) \frac{1 - ErlC(N, \frac{\lambda}{\mu})}{N - \frac{\lambda}{\mu}} + \frac{\lambda}{\mu} \frac{-\frac{\partial ErlC(N, \frac{\lambda}{\mu})}{\partial \mu} \left(N - \frac{\lambda}{\mu} \right) - \left(1 - ErlC(N, \frac{\lambda}{\mu}) \right) \frac{\lambda}{\mu^2}}{\left(N - \frac{\lambda}{\mu} \right)^2} \\ & \quad + \frac{\partial ErlC(N, \frac{\lambda}{\mu})}{\partial \mu} \sum_{i=1}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i - ErlC(N, \frac{\lambda}{\mu}) \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^{i-1} \left(-\frac{\lambda}{N\mu^2} \right) \\ &= -\frac{\lambda}{\mu^2} \frac{1 - ErlC(N, \frac{\lambda}{\mu})}{N - \frac{\lambda}{\mu}} - \left(1 - ErlC(N, \frac{\lambda}{\mu}) \right) \frac{\frac{\lambda}{\mu} \frac{\lambda}{\mu^2}}{\left(N - \frac{\lambda}{\mu} \right)^2} \\ & \quad + \frac{\partial ErlC(N, \rho)}{\partial \rho} \left(-\frac{\lambda}{\mu^2} \right) \left[\sum_{i=1}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i - \frac{\lambda}{N - \frac{\lambda}{\mu}} \right] - \frac{ErlC(N, \frac{\lambda}{\mu})}{\mu} \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^i \\ &= -\frac{\lambda}{\mu^2} \frac{1 - ErlC(N, \frac{\lambda}{\mu})}{N - \frac{\lambda}{\mu}} \frac{N}{N - \frac{\lambda}{\mu}} + \frac{\partial ErlC(N, \rho)}{\partial \rho} \left(-\frac{\lambda}{\mu^2} \right) \left[\sum_{i=1}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i - \frac{\lambda}{N - \frac{\lambda}{\mu}} \right] - \frac{ErlC(N, \frac{\lambda}{\mu})}{\mu} \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^i. \end{aligned} \quad (\text{A.43})$$

On the other hand, note that

$$\begin{aligned}
& \frac{N}{\mu} \frac{1 - I(\mu, \mu)}{I(\mu, \mu)} + \frac{1}{\mu} ErlC \left(N, \frac{\lambda}{\mu} \right) \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^i - \frac{\lambda}{\mu^2} \frac{1}{I(\mu, \mu)} \left(-\frac{1}{ErlC \left(N, \frac{\lambda}{\mu} \right)} \frac{\partial ErlC(N, \rho)}{\partial \rho} + \frac{N\mu}{\lambda} \right) \\
&= \frac{1}{I(\mu, \mu)} \cdot \frac{\lambda}{\mu^2} \frac{1}{ErlC \left(N, \frac{\lambda}{\mu} \right)} \frac{\partial ErlC(N, \rho)}{\partial \rho} - \frac{N}{\mu} + \frac{ErlC \left(N, \frac{\lambda}{\mu} \right)}{\mu} \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^i \\
&\stackrel{(i)}{=} \left[1 + \frac{\lambda}{\mu} \frac{1 - ErlC \left(N, \frac{\lambda}{\mu} \right)}{N - \frac{\lambda}{\mu}} + ErlC \left(N, \frac{\lambda}{\mu} \right) \sum_{i=1}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i \right] \frac{\lambda}{\mu^2} \frac{1}{ErlC \left(N, \frac{\lambda}{\mu} \right)} \frac{\partial ErlC(N, \rho)}{\partial \rho} \\
&\quad - \frac{N}{\mu} + \frac{ErlC \left(N, \frac{\lambda}{\mu} \right)}{\mu} \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^i \\
&= \frac{\lambda}{\mu^2} \frac{\partial ErlC(N, \rho)}{\partial \rho} \left[\sum_{i=1}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i - \frac{\frac{\lambda}{\mu}}{N - \frac{\lambda}{\mu}} + \frac{N}{ErlC \left(N, \frac{\lambda}{\mu} \right) \left(N - \frac{\lambda}{\mu} \right)} \right] - \frac{N}{\mu} + \frac{ErlC \left(N, \frac{\lambda}{\mu} \right)}{\mu} \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^i \\
&\stackrel{(ii)}{=} \frac{\lambda}{\mu^2} \frac{\partial ErlC(N, \rho)}{\partial \rho} \left[\sum_{i=1}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i - \frac{\frac{\lambda}{\mu}}{N - \frac{\lambda}{\mu}} \right] + \frac{ErlC \left(N, \frac{\lambda}{\mu} \right)}{\mu} \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^i \\
&\quad + \frac{\lambda}{\mu^2} ErlC \left(N, \frac{\lambda}{\mu} \right) \left(\frac{1 - ErlC \left(N, \frac{\lambda}{\mu} \right)}{N - \frac{\lambda}{\mu}} + \frac{N - \frac{\lambda}{\mu}}{\frac{\lambda}{\mu}} \right) \frac{N}{ErlC \left(N, \frac{\lambda}{\mu} \right) \left(N - \frac{\lambda}{\mu} \right)} - \frac{N}{\mu} \\
&= \frac{\lambda}{\mu^2} \frac{\partial ErlC(N, \rho)}{\partial \rho} \left[\sum_{i=1}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i - \frac{\frac{\lambda}{\mu}}{N - \frac{\lambda}{\mu}} \right] + \frac{ErlC \left(N, \frac{\lambda}{\mu} \right)}{\mu} \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^i + \frac{\lambda}{\mu^2} \frac{N}{N - \frac{\lambda}{\mu}} \frac{1 - ErlC \left(N, \frac{\lambda}{\mu} \right)}{N - \frac{\lambda}{\mu}},
\end{aligned} \tag{A.44}$$

where (i) follows from (A.42), and (ii) follows from Lemma 20.

Comparing (A.43) and (A.44) finds that

$$\begin{aligned}
\frac{1}{I(\mu, \mu)^2} \frac{dI(\mu, \mu)}{d\mu} &= \frac{N}{\mu} \frac{1 - I(\mu, \mu)}{I(\mu, \mu)} + \frac{1}{\mu} ErlC \left(N, \frac{\lambda}{\mu} \right) \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^i \\
&\quad - \frac{\lambda}{\mu^2} \frac{1}{I(\mu, \mu)} \left(-\frac{1}{ErlC \left(N, \frac{\lambda}{\mu} \right)} \frac{\partial ErlC(N, \rho)}{\partial \rho} + \frac{N\mu}{\lambda} \right),
\end{aligned}$$

which implies that

$$\begin{aligned}
\frac{dI(\mu, \mu)}{d\mu} &= \frac{N}{\mu} I(\mu, \mu)(1 - I(\mu, \mu)) + \frac{1}{\mu} I(\mu, \mu)^2 ErlC\left(N, \frac{\lambda}{\mu}\right) \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu}\right)^i \\
&\quad - \frac{\lambda}{\mu^2} I(\mu, \mu) \left(-\frac{1}{ErlC\left(N, \frac{\lambda}{\mu}\right)} \frac{\partial ErlC(N, \rho)}{\partial \rho} + \frac{N\mu}{\lambda} \right) \\
&\stackrel{(i)}{=} N \frac{\partial I(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} - \frac{I(\mu, \mu)}{\mu} \frac{\lambda}{\mu} \left(-\frac{1}{ErlC\left(N, \frac{\lambda}{\mu}\right)} \frac{\partial ErlC(N, \rho)}{\partial \rho} + \frac{N\mu}{\lambda} \right) \\
&\stackrel{(ii)}{=} N \frac{\partial I(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} - \frac{I(\mu, \mu)}{\mu} \frac{\lambda}{N\mu} \left(N - \frac{1 - ErlC\left(N, \frac{\lambda}{\mu}\right)}{1 - \frac{\lambda}{N\mu}} \right),
\end{aligned}$$

where (i) follows from (A.21) in Corollary 2 and (ii) follows from Lemma 20. \blacksquare

A.4.1.6 Proof of Lemma 25

Letting $k = \infty$ in (A.20) in Corollary 2 yields

$$I(\mu, \mu; \lambda, \infty, N) = \lim_{k \rightarrow \infty} I(\mu, \mu; \lambda, k, N) = 1 - \frac{\lambda}{N\mu}.$$

This implies that the difference between the idle time in an $M/M/N/k$ system and that in an $M/M/N$ system is given by $I(\mu, \mu) - \left(1 - \frac{\lambda}{N\mu}\right)$. For any $\mu \in (0, \infty)$,

$$\begin{aligned}
&\mu \cdot \frac{d}{d\mu} \left(I(\mu, \mu) - \left(1 - \frac{\lambda}{N\mu}\right) \right) = \mu \frac{dI(\mu, \mu)}{d\mu} - \frac{\lambda}{N\mu} \\
&\stackrel{(i)}{=} N \left(\mu \frac{\partial I(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} \right) - I(\mu, \mu) \frac{\lambda}{\mu} \left(1 - \frac{1 - ErlC\left(N, \frac{\lambda}{\mu}\right)}{N - \frac{\lambda}{\mu}} \right) - \frac{\lambda}{N\mu} \\
&\stackrel{(ii)}{=} N \left(I(\mu, \mu)(1 - I(\mu, \mu)) + I(\mu, \mu)^2 \frac{ErlC\left(N, \frac{\lambda}{\mu}\right)}{N} \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu}\right)^i \right) - I(\mu, \mu) \left(\frac{\lambda}{\mu} - \left(\frac{ErlC\left(N, \frac{\lambda}{\mu}\right)}{ErlB\left(N, \frac{\lambda}{\mu}\right)} - 1 \right) \right) - \frac{\lambda}{N\mu} \\
&\stackrel{(iii)}{=} I(\mu, \mu) \left\{ -NI(\mu, \mu) + I(\mu, \mu) ErlC\left(N, \frac{\lambda}{\mu}\right) \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu}\right)^i + N - \frac{\lambda}{N\mu} + \frac{ErlC\left(N, \frac{\lambda}{\mu}\right)}{ErlB\left(N, \frac{\lambda}{\mu}\right)} - 1 \right. \\
&\quad \left. - \frac{\lambda}{N\mu} \left(\frac{ErlC\left(N, \frac{\lambda}{\mu}\right)}{ErlB\left(N, \frac{\lambda}{\mu}\right)} + ErlC\left(N, \frac{\lambda}{\mu}\right) \sum_{i=1}^{k-N} \left(\frac{\lambda}{N\mu}\right)^i \right) \right\}
\end{aligned}$$

$$\begin{aligned}
&= I(\mu, \mu) \left\{ I(\mu, \mu) ErlC \left(N, \frac{\lambda}{\mu} \right) \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^i + N(1 - I(\mu, \mu)) - \frac{\lambda}{\mu} - ErlC \left(N, \frac{\lambda}{\mu} \right) \frac{\lambda}{N\mu} \sum_{i=0}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i \right\} \\
&= I(\mu, \mu) \left\{ N \left(\frac{ErlC \left(N, \frac{\lambda}{\mu} \right)}{ErlB \left(N, \frac{\lambda}{\mu} \right)} - 1 \right) I(\mu, \mu) + I(\mu, \mu) ErlC \left(N, \frac{\lambda}{\mu} \right) \sum_{i=1}^{k-N} (i+N) \left(\frac{\lambda}{N\mu} \right)^i \right. \\
&\quad \left. - \frac{\lambda}{\mu} - \frac{\lambda}{N\mu} ErlC \left(N, \frac{\lambda}{\mu} \right) \sum_{i=0}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i \right\} \\
&= \frac{\lambda}{N\mu} I(\mu, \mu) \left\{ \frac{1 - ErlC \left(N, \frac{\lambda}{\mu} \right)}{1 - \frac{\lambda}{N\mu}} NI(\mu, \mu) - N + I(\mu, \mu) ErlC \left(N, \frac{\lambda}{\mu} \right) \sum_{i=0}^{k-N-1} (N+i+1) \left(\frac{\lambda}{N\mu} \right)^i \right. \\
&\quad \left. - ErlC \left(N, \frac{\lambda}{\mu} \right) \sum_{i=0}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i \right\} \\
&\stackrel{(iv)}{=} \frac{\lambda}{N\mu} I(\mu, \mu) \left\{ \left[N \frac{1 - ErlC \left(N, \frac{\lambda}{\mu} \right)}{1 - \frac{\lambda}{N\mu}} - N - \frac{\lambda}{\mu} \frac{1 - ErlC \left(N, \frac{\lambda}{\mu} \right)}{1 - \frac{\lambda}{N\mu}} - N ErLC \left(N, \frac{\lambda}{\mu} \right) \sum_{i=1}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i \right] I(\mu, \mu) \right. \\
&\quad \left. + I(\mu, \mu) ErlC \left(N, \frac{\lambda}{\mu} \right) \sum_{i=0}^{k-N-1} (N+i+1) \left(\frac{\lambda}{N\mu} \right)^i - ErlC \left(N, \frac{\lambda}{\mu} \right) \sum_{i=0}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i \right\} \\
&= \frac{\lambda}{N\mu} I(\mu, \mu) ErlC \left(N, \frac{\lambda}{\mu} \right) \left\{ \left[\sum_{i=0}^{k-N} (i+1) \left(\frac{\lambda}{N\mu} \right)^i - (k+1) \left(\frac{\lambda}{N\mu} \right)^{k-N} \right] I(\mu, \mu) - \sum_{i=0}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i \right\} \\
&= \frac{\lambda}{N\mu} I(\mu, \mu)^2 ErlC \left(N, \frac{\lambda}{\mu} \right) \left\{ \sum_{i=0}^{k-N} (i+1) \left(\frac{\lambda}{N\mu} \right)^i - (k+1) \left(\frac{\lambda}{N\mu} \right)^{k-N} \right. \\
&\quad \left. - \left[N \frac{1 - ErlC \left(N, \frac{\lambda}{\mu} \right)}{N - \frac{\lambda}{\mu}} + ErlC \left(N, \frac{\lambda}{\mu} \right) \sum_{i=0}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i \right] \sum_{i=0}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i \right\} \\
&= \frac{\lambda}{N\mu} I(\mu, \mu)^2 ErlC \left(N, \frac{\lambda}{\mu} \right) \left\{ -N(k-N+1) \frac{1 - ErlC \left(N, \frac{\lambda}{\mu} \right)}{N - \frac{\lambda}{\mu}} \left(\frac{\lambda}{N\mu} \right)^{k-N+1} - (k+1) \left(\frac{\lambda}{N\mu} \right)^{k-N} \right. \\
&\quad \left. - ErlC \left(N, \frac{\lambda}{\mu} \right) \left(\frac{\lambda}{N\mu} \right)^{k-N+1} \sum_{i=0}^{k-N} (k-N-i) \left(\frac{\lambda}{N\mu} \right)^i \right\} \\
&\stackrel{(v)}{=} - \left(\frac{\lambda}{N\mu} \right)^{k-N+1} I(\mu, \mu)^2 ErlC \left(N, \frac{\lambda}{\mu} \right) \left(N + ErlC \left(N, \frac{\lambda}{\mu} \right) \left(\frac{k-N+1}{ErlB \left(N, \frac{\lambda}{\mu} \right)} + \frac{\lambda}{N\mu} \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^{k-N+i} \right) \right) < 0,
\end{aligned}$$

where (i) follows from Lemma 24, (ii) follows from (A.21) in Corollary 2 and $\rho \frac{1 - ErlC(N, \rho)}{N - \rho} = \frac{ErlC(N, \rho)}{ErlB(N, \rho)} - 1$ (using the relationship between $ErlB$ and $ErlC$ from Lemma 19 (a)), (iii) follows from (A.19) in Corollary 2 and the definition of $ErlB$ in (A.1), (iv) follows from (A.18) in Corollary 2, and (v) follows from $\rho \frac{1 - ErlC(N, \rho)}{N - \rho} = \frac{ErlC(N, \rho)}{ErlB(N, \rho)} - 1$.

Hence, $I(\mu, \mu; \lambda, k, N) - \left(1 - \frac{\lambda}{N\mu} \right)$ is strictly decreasing in μ for $\mu \in (0, \infty)$. Moreover,

note that $\lim_{\mu \rightarrow \infty} I(\mu, \mu; \lambda, k, N) - \left(1 - \frac{\lambda}{N\mu}\right) = 1 - 1 = 0$ (from Lemma 19 (b)). Thus, $I(\mu, \mu; \lambda, k, N) - \left(1 - \frac{\lambda}{N\mu}\right) > 0$ for all $\mu \in (0, \infty)$. Hence, the difference between the idle time in an $M/M/N/k$ system and that in an $M/M/N$ system satisfies $I(\mu, \mu; \lambda, k, N) - \left(1 - \frac{\lambda}{N\mu}\right) \geq 0$ with equality holding only when $k = \infty$.

■

A.4.2 Proof of Theorem 1

Since $\mu \neq 0$, the FOC in (1.6) for a symmetric equilibrium is equivalent to

$$\frac{\mu c'(\mu)}{I(\mu, \mu)} = p \left(\frac{\mu}{I(\mu, \mu)} - \mu \right) + (v - p\mu) \frac{\mu I'(\mu, \mu)}{I(\mu, \mu)},$$

where $I'(\mu, \mu)$ denotes $\frac{\partial I(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu}$. Set $LHS(\mu)$ and $RHS(\mu)$ equal to the left-hand side and the right-hand side of the above display, respectively. Since $LHS(\mu)$ and $RHS(\mu)$ are both continuous functions of $\mu \in (0, \infty)$, a sufficient condition for the FOC to admit a solution for $\mu \in (0, \infty)$ is

$$\lim_{\mu \downarrow 0} LHS(\mu) < \lim_{\mu \downarrow 0} RHS(\mu) \text{ and } \lim_{\mu \uparrow \infty} \mu \cdot LHS(\mu) > \lim_{\mu \uparrow \infty} \mu \cdot RHS(\mu). \quad (\text{A.45})$$

Claim 4. *The following hold:*

$$\begin{aligned} \lim_{\mu \downarrow 0} LHS(\mu) &= \lim_{\mu \downarrow 0} \frac{c'(\mu)}{\mu^{k-N}} N \left(\frac{\lambda}{N} \right)^{k-N+1}, \quad \lim_{\mu \downarrow 0} RHS(\mu) = \lim_{\mu \downarrow 0} p \cdot \frac{1}{\mu^{k-N}} N \left(\frac{\lambda}{N} \right)^{k-N+1} + v \frac{k}{N}, \\ \lim_{\mu \uparrow \infty} \mu \cdot LHS(\mu) &= \lim_{\mu \uparrow \infty} \mu^2 c'(\mu), \quad \lim_{\mu \uparrow \infty} \mu \cdot RHS(\mu) = -\frac{1}{2} p \lambda^2 + v \frac{\lambda}{N} - p \mu \lambda \left(1 + \frac{1}{N} \right). \end{aligned}$$

By Claim 4, (A.45) becomes

$$\begin{aligned} \lim_{\mu \downarrow 0} \frac{c'(\mu)}{\mu^{k-N}} N \left(\frac{\lambda}{N} \right)^{k-N+1} &< \lim_{\mu \downarrow 0} p \cdot \frac{1}{\mu^{k-N}} N \left(\frac{\lambda}{N} \right)^{k-N+1} + v \frac{k}{N}, \text{ and} \\ \lim_{\mu \uparrow \infty} \mu^2 c'(\mu) &> \lim_{\mu \uparrow \infty} -\frac{1}{2} p \lambda^2 + v \frac{\lambda}{N} - p \mu \lambda \left(1 + \frac{1}{N} \right), \end{aligned}$$

which can be equivalently written as

$$\lim_{\mu \downarrow 0} \frac{c'(\mu) - p}{\mu^{k-N}} < v \frac{k}{N^2} \left(\frac{N}{\lambda} \right)^{k-N+1}, \text{ and } \lim_{\mu \uparrow \infty} \mu^2 c'(\mu) + p\mu\lambda \left(1 + \frac{1}{N} \right) > v \frac{\lambda}{N} - \frac{1}{2} p\lambda^2,$$

where the second condition is always true, and the first condition establishes (1.7).

The second part of the theorem is established by applying (1.7) to the Taylor expansion of $c(\mu)$ and $c'(\mu)$ at $\mu = 0$, recalling that an entire function is equal to the sum of its Taylor series everywhere. We first express $c'(\mu)$ as a Taylor series at 0:

$$c'(\mu) = \sum_{i=1}^{\infty} \frac{c^{(i)}(0)}{(i-1)!} \mu^{i-1} \Leftrightarrow \frac{c'(\mu)}{\mu^{k-N}} = \sum_{i=1}^{\infty} \frac{c^{(i)}(0)}{(i-1)!} \mu^{i-(k-N+1)}.$$

Then, (1.7) can be rewritten as

$$\lim_{\mu \downarrow 0} \sum_{i=1}^{\infty} \frac{c^{(i)}(0)}{(i-1)!} \mu^{i-(k-N+1)} < v \frac{k}{N^2} \left(\frac{N}{\lambda} \right)^{k-N+1} + \lim_{\mu \downarrow 0} \frac{p}{\mu^{k-N}}.$$

This inequality holds if and only if

$$\begin{cases} c^{(1)}(0) < p, \text{ and } c^{(i)}(0) \text{ any for } i \geq 2, & \text{if } k > N, \\ c^{(1)}(0) < \frac{v}{\lambda} + p, \text{ and } c^{(i)}(0) \text{ any for } i \geq 2, & \text{if } k = N. \end{cases} \quad (\text{A.46})$$

Next, we express $c(\mu)$ as a Taylor series at 0:

$$c(\mu) = c(0) + \sum_{i=1}^{\infty} \frac{c^{(i)}(0)}{i!} \mu^i,$$

Let integer $q > 0$ be such that $c^{(i)}(0) = 0$ for all $1 \leq i < q$ and $c^{(q)}(0) \neq 0$. Then, the above display can be rewritten as

$$\begin{aligned} c(\mu) &= c(0) + \sum_{i=q}^{\infty} \frac{c^{(i)}(0)}{i!} \mu^i = c(0) + \frac{c^{(q)}(0)}{q!} \mu^q \left(1 + \sum_{i=q+1}^{\infty} \frac{c^{(i)}(0)}{c^{(q)}(0)} \frac{q!}{i!} \mu^{i-q} \right) \\ &=: b_E + c_E \mu^q \cdot h(\mu), \end{aligned}$$

where $b_E := c(0) = 0$ (recalling that $c(0) = 0$), $c_E := \frac{c^{(q)}(0)}{q!} \neq 0$, and $h(\mu) := 1 + \sum_{i=q+1}^{\infty} \frac{c^{(i)}(0)}{c^{(q)}(0)} \frac{q!}{i!} \mu^{i-q} = h(0) + \sum_{i=q+1}^{\infty} \frac{h^{(i-q)}(0)}{(i-q)!} \mu^{i-q}$ is an entire function with $h(0) = 1$.

- If $q \geq 2$, then $c^{(1)}(0) = 0$. Thus, condition (A.46) is satisfied.
- If $q = 1$, then condition (A.46) holds if and only if

$$c_E = \frac{c^{(1)}(0)}{1!} = c^{(1)}(0) < \begin{cases} p, & \text{if } k > N, \\ \frac{v}{\lambda} + p, & \text{if } k = N. \end{cases}$$

Therefore, an entire function c satisfies (1.7) if and only if it is of the form $c(\mu) = c_E \mu^q \cdot h(\mu)$, where $c_E \in \mathbb{R}$, $c_E \neq 0$, $q \in \mathbb{Z}_+$; h is an entire function with $h(0) = 1$; and, either (a) $q = 1$ and $c_E < p + \frac{v}{\lambda} \mathbf{1}\{k = N\}$, or (b) $q \geq 2$. ■

To complete the proof, we verify Claim 4 below.

A.4.2.1 Proof of Claim 4

Note that the building blocks for evaluating these limits are $\frac{\mu}{I(\mu, \mu)}$ and $\frac{\mu I'(\mu, \mu)}{I(\mu, \mu)}$, so we first investigate their associated limits as $\mu \downarrow 0$ and as $\mu \uparrow \infty$.

From (A.19) in Corollary 2,

$$\begin{aligned} \frac{\mu}{I(\mu, \mu)} &= \mu \left(1 + \sum_{i=0}^{N-1} \frac{N!}{i!} \left(\frac{\mu}{\lambda} \right)^{N-i} + \sum_{i=1}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i \right) \text{ErlC} \left(N, \frac{\lambda}{\mu} \right) \\ &= \mu \left(\left(\frac{N\mu}{\lambda} \right)^{k-N} + \sum_{i=0}^{N-1} \frac{N!}{i!} \left(\frac{\mu}{\lambda} \right)^{N-i} \left(\frac{N\mu}{\lambda} \right)^{k-N} + \sum_{i=1}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i \left(\frac{N\mu}{\lambda} \right)^{k-N} \right) \left(\frac{\lambda}{N\mu} \right)^{k-N} \text{ErlC} \left(N, \frac{\lambda}{N\mu} \right) \\ &= \mu \left(1 + \sum_{i=1}^{k-N} \left(\frac{N\mu}{\lambda} \right)^i + N^{k-N} \sum_{i=k-N+1}^k \frac{N!}{(k-i)!} \left(\frac{\mu}{\lambda} \right)^i \right) \left(\frac{\lambda}{N\mu} \right)^{k-N} \text{ErlC} \left(N, \frac{\lambda}{\mu} \right) \\ &= \frac{1}{\mu^{k-N}} N \left(\frac{\lambda}{N} \right)^{k-N+1} \frac{\text{ErlC} \left(N, \frac{\lambda}{\mu} \right)}{\frac{\lambda}{\mu}} \left(1 + \sum_{i=1}^{k-N} \left(\frac{N\mu}{\lambda} \right)^i + N^{k-N} \sum_{i=k-N+1}^k \frac{N!}{(k-i)!} \left(\frac{\mu}{\lambda} \right)^i \right), \end{aligned}$$

which implies that

$$\lim_{\mu \downarrow 0} \frac{\mu}{I(\mu, \mu)} = \lim_{\mu \downarrow 0} \frac{1}{\mu^{k-N}} N \left(\frac{\lambda}{N} \right)^{k-N+1} \frac{\text{ErlC} \left(N, \frac{\lambda}{\mu} \right)}{\frac{\lambda}{\mu}} \stackrel{(*)}{=} \lim_{\mu \downarrow 0} \frac{1}{\mu^{k-N}} N \left(\frac{\lambda}{N} \right)^{k-N+1}, \quad (\text{A.47})$$

where $(*)$ follows from $\lim_{\mu \downarrow 0} \frac{ErlC(N, \frac{\lambda}{\mu})}{\frac{\lambda}{\mu}} = 1$ (Lemma 19 (b)), and

$$\begin{aligned} \lim_{\mu \uparrow \infty} \mu \cdot \frac{\mu}{I(\mu, \mu)} &= \lim_{\mu \uparrow \infty} \frac{\mu}{\mu^{k-N}} N \left(\frac{\lambda}{N} \right)^{k-N+1} \frac{ErlC(N, \frac{\lambda}{\mu})}{\frac{\lambda}{\mu}} N^{k-N} N! \left(\frac{\mu}{\lambda} \right)^k \\ &= \lim_{\mu \uparrow \infty} \mu^2 \left(\frac{\mu}{\lambda} \right)^N N! ErlC \left(N, \frac{\lambda}{\mu} \right) \stackrel{(**)}{=} \lim_{\mu \uparrow \infty} \mu^2 \frac{\frac{N}{N-\lambda/\mu}}{\sum_{i=0}^{N-1} \left(\frac{\lambda}{\mu} \right)^i \frac{1}{i!} + \left(\frac{\lambda}{\mu} \right)^N \frac{1}{N!} \frac{N}{N-\lambda/\mu}}, \end{aligned} \quad (\text{A.48})$$

where $(**)$ follows from the expression for $ErlC$ in (A.3).

From (A.19) and (A.21) in Corollary 2,

$$\begin{aligned} \frac{\mu I'(\mu, \mu)}{I(\mu, \mu)} &= 1 - I(\mu, \mu) \left(1 - \frac{ErlC(N, \frac{\lambda}{\mu})}{N} \sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^i \right) \\ &= 1 - \frac{1 - \frac{ErlC(N, \frac{\lambda}{\mu})}{N} \left(\sum_{i=1}^{k-N} i \left(\frac{\lambda}{N\mu} \right)^i \left(\frac{N\mu}{\lambda} \right)^{k-N} \right) \left(\frac{\lambda}{N\mu} \right)^{k-N}}{\left(\left(\frac{N\mu}{\lambda} \right)^{k-N} + \sum_{i=0}^{N-1} \frac{N!}{i!} \left(\frac{\mu}{\lambda} \right)^{N-i} \left(\frac{N\mu}{\lambda} \right)^{k-N} + \sum_{i=1}^{k-N} \left(\frac{\lambda}{N\mu} \right)^i \left(\frac{N\mu}{\lambda} \right)^{k-N} \right) \left(\frac{\lambda}{N\mu} \right)^{k-N} ErlC \left(N, \frac{\lambda}{\mu} \right)} \\ &= 1 - \frac{1 - \left(k - N + \sum_{i=1}^{k-N} (k - N - i) \left(\frac{N\mu}{\lambda} \right)^i \right) \left(\frac{\lambda}{N\mu} \right)^{k-N} \frac{ErlC(N, \frac{\lambda}{\mu})}{N}}{\left(1 + \sum_{i=1}^{k-N} \left(\frac{N\mu}{\lambda} \right)^i + N^{k-N} \sum_{i=k-N+1}^k \frac{N!}{(k-i)!} \left(\frac{\mu}{\lambda} \right)^i \right) \left(\frac{\lambda}{N\mu} \right)^{k-N} ErlC \left(N, \frac{\lambda}{\mu} \right)} \\ &= \frac{\frac{k}{N} + \sum_{i=1}^{k-N-1} \frac{k-i}{N} \left(\frac{N\mu}{\lambda} \right)^i + N^{k-N} \sum_{i=k-N+1}^k \frac{(N-1)!}{(k-i)!} \left(\frac{\mu}{\lambda} \right)^{i-1}}{1 + \sum_{i=1}^{k-N} \left(\frac{N\mu}{\lambda} \right)^i + N^{k-N} \sum_{i=k-N+1}^k \frac{N!}{(k-i)!} \left(\frac{\mu}{\lambda} \right)^i}, \end{aligned}$$

which implies that

$$\lim_{\mu \downarrow 0} \frac{\mu I'(\mu, \mu)}{I(\mu, \mu)} = \frac{k}{N}, \quad (\text{A.49})$$

and

$$\lim_{\mu \uparrow \infty} \mu \cdot \frac{\mu I'(\mu, \mu)}{I(\mu, \mu)} = \lim_{\mu \uparrow \infty} \mu \cdot \frac{N^{k-N} (N-1)! \left(\frac{\mu}{\lambda} \right)^{k-1}}{N^{k-N} N! \left(\frac{\mu}{\lambda} \right)^k} = \frac{\lambda}{N}. \quad (\text{A.50})$$

When $\lim_{\mu \downarrow 0} \frac{c'(\mu)}{\mu^{k-N}}$ exists and is finite, (A.47) implies that

$$\lim_{\mu \downarrow 0} LHS(\mu) = \lim_{\mu \downarrow 0} \frac{c'(\mu)}{\mu^{k-N}} N \left(\frac{\lambda}{N} \right)^{k-N+1}, \quad (\text{A.51})$$

and (A.48) implies that

$$\lim_{\mu \uparrow \infty} \mu \cdot LHS(\mu) = \lim_{\mu \uparrow \infty} \mu^2 c'(\mu) \frac{\frac{N}{N-\lambda/\mu}}{\sum_{i=0}^{N-1} \left(\frac{\lambda}{\mu} \right)^i \frac{1}{i!} + \left(\frac{\lambda}{\mu} \right)^N \frac{1}{N!} \frac{N}{N-\lambda/\mu}} \stackrel{(*)}{=} \lim_{\mu \uparrow \infty} \mu^2 c'(\mu), \quad (\text{A.52})$$

where $(*)$ follows because $\sum_{i=0}^{N-1} \left(\frac{\lambda}{\mu} \right)^i \frac{1}{i!} \rightarrow 1$, $\left(\frac{\lambda}{\mu} \right)^N \frac{1}{N!} \frac{N}{N-\lambda/\mu} \rightarrow 0$, and $\frac{N}{N-\lambda/\mu} \rightarrow 1$, as $\mu \rightarrow \infty$.

From (A.47) and (A.49),

$$\lim_{\mu \downarrow 0} RHS(\mu) = p \cdot \lim_{\mu \downarrow 0} \frac{\mu}{I(\mu, \mu)} + \lim_{\mu \downarrow 0} (v - p\mu) \frac{\mu I'(\mu, \mu)}{I(\mu, \mu)} = \lim_{\mu \downarrow 0} p \cdot \frac{1}{\mu^{k-N}} N \left(\frac{\lambda}{N} \right)^{k-N+1} + v \frac{k}{N}, \quad (\text{A.53})$$

and from (A.48) and (A.50),

$$\begin{aligned} \lim_{\mu \uparrow \infty} \mu \cdot RHS(\mu) &= \lim_{\mu \uparrow \infty} p \cdot \left(\mu \cdot \frac{\mu}{I(\mu, \mu)} - \mu^2 \right) + v \cdot \lim_{\mu \uparrow \infty} \mu \cdot \frac{\mu I'(\mu, \mu)}{I(\mu, \mu)} - p \cdot \lim_{\mu \uparrow \infty} \frac{\mu^3 I'(\mu, \mu)}{I(\mu, \mu)} \\ &= \lim_{\mu \uparrow \infty} p \cdot \mu^2 \frac{\frac{N}{N-\lambda/\mu}}{\sum_{i=0}^{N-1} \left(\frac{\lambda}{\mu} \right)^i \frac{1}{i!} + \left(\frac{\lambda}{\mu} \right)^N \frac{1}{N!} \frac{N}{N-\lambda/\mu}} - \mu^2 + v \cdot \frac{\lambda}{N} - p \cdot \frac{\lambda \mu}{N} \\ &= \lim_{\mu \uparrow \infty} p \cdot \mu^2 \frac{\frac{N}{N-\lambda/\mu} \left(1 - \left(\frac{\lambda}{\mu} \right)^N \frac{1}{N!} \right) - \sum_{i=0}^{N-1} \left(\frac{\lambda}{\mu} \right)^i \frac{1}{i!}}{\sum_{i=0}^{N-1} \left(\frac{\lambda}{\mu} \right)^i \frac{1}{i!} + \left(\frac{\lambda}{\mu} \right)^N \frac{1}{N!} \frac{N}{N-\lambda/\mu}} + (v - p\mu) \frac{\lambda}{N} \\ &= \lim_{\mu \uparrow \infty} p \cdot \frac{\frac{N}{N-\lambda/\mu} \left(\mu^2 - \left(\frac{\lambda}{\mu} \right)^{N-2} \frac{\lambda^2}{N!} \right) - \left(\mu^2 + \lambda\mu + \frac{1}{2!} \lambda^2 + \frac{1}{3!} \frac{\lambda^3}{\mu} + \dots + \frac{1}{(N-1)!} \frac{\lambda^{N-1}}{\mu^{N-3}} \right)}{\sum_{i=0}^{N-1} \left(\frac{\lambda}{\mu} \right)^i \frac{1}{i!} + \left(\frac{\lambda}{\mu} \right)^N \frac{1}{N!} \frac{N}{N-\lambda/\mu}} + (v - p\mu) \frac{\lambda}{N} \\ &\stackrel{(*)}{=} \lim_{\mu \uparrow \infty} p \cdot \left(\mu^2 - \left(\mu^2 + \lambda\mu + \frac{1}{2} \lambda^2 \right) \right) + (v - p\mu) \frac{\lambda}{N} \\ &= \lim_{\mu \uparrow \infty} -\frac{1}{2} p\lambda^2 + v \frac{\lambda}{N} - p\mu\lambda \left(1 + \frac{1}{N} \right), \end{aligned} \quad (\text{A.54})$$

where $(*)$ follows by noting that $\sum_{i=0}^{N-1} \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!} \rightarrow 1$, $\left(\frac{\lambda}{\mu}\right)^N \frac{1}{N!} \frac{N}{N-\lambda/\mu} \rightarrow 0$, $\frac{N}{N-\frac{\lambda}{\mu}} \rightarrow 1$, $\left(\frac{\lambda}{\mu}\right)^{N-2} \frac{\lambda^2}{N!} \rightarrow 0$, and $\frac{1}{3!} \frac{\lambda^3}{\mu} + \dots + \frac{1}{(N-1)!} \frac{\lambda^{N-1}}{\mu^{N-3}} \rightarrow 0$, as $\mu \rightarrow \infty$. \blacksquare

A.5 Proofs from Section 1.4.1

A.5.1 Preliminaries

The expressions for $I(\mu_1, \mu)$ and its derivatives simplify significantly for a loss system, i.e., when $k = N$. We provide the expressions and their properties in the following lemmas.

Lemma 26. *In an $M/M/N/N$ loss system, the steady-state probability that the tagged server is idle, and its first two partial derivatives with respect to μ_1 satisfy the following expressions:*

- (a) $I(\mu_1, \mu) = \frac{1 - \frac{\rho}{N}(1 - ErlB(N, \rho))}{1 - \left(1 - \frac{\mu}{\mu_1}\right) \frac{\rho}{N}(1 - ErlB(N, \rho))}$.
- (b) $\frac{\partial I(\mu_1, \mu)}{\partial \mu_1} = \frac{I(\mu_1, \mu)^2}{\mu_1} \left(\frac{1}{I(\mu_1, \mu)} - 1 \right) > 0, \quad \forall \mu_1, \mu > 0$.
- (c) $\frac{\partial^2 I(\mu_1, \mu)}{\partial \mu_1^2} = -2 \frac{I(\mu_1, \mu)^2}{\mu_1} (1 - I(\mu_1, \mu)) < 0, \quad \forall \mu_1, \mu > 0$.

Corollary 4 ($\mu_1 = \mu$ in Lemma 26).

- (a) $I(\mu, \mu) = 1 - \frac{\rho}{N}(1 - ErlB(N, \rho))$.
- (b) $\frac{\partial I(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} = \frac{I(\mu, \mu)^2}{\mu} \left(\frac{1}{I(\mu, \mu)} - 1 \right) > 0, \quad \forall \mu > 0$.
- (c) $\frac{\partial^2 I(\mu_1, \mu)}{\partial \mu_1^2} \Big|_{\mu_1=\mu} = -2 \frac{I(\mu, \mu)^2}{\mu} (1 - I(\mu, \mu)) < 0, \quad \forall \mu > 0$.

Lemma 27. *In an $M/M/N/N$ loss system, $\frac{I(\mu, \mu; \lambda, N)}{\mu}$ is strictly increasing in μ for $\mu \in (0, \frac{5\lambda}{4N}]$ for all $\lambda > 0$ and $N \geq 6$.*

Lemma 28. *In an $M/M/N/N$ loss system, for any $x \in (0, 1)$, $\mu > 0$, and $N > 0$, there exists a unique $\lambda > 0$ such that $I(\mu, \mu; \lambda, N) = x$. Similarly, for any $x \in (0, 1)$, $\mu > 0$, and $\lambda > 0$, there exists a unique (real-valued) $N > 0$ such that $I(\mu, \mu; \lambda, N) = x$.*

A.5.1.1 Proof of Lemma 26

(a): From Lemma 2, substituting $k = N$ into (1.1) yields

$$I(\mu_1, \mu) = \left(1 + \rho \frac{\mu}{\mu_1} \left(\frac{1 - ErlC(N, \rho)}{N - \rho} \right) \right)^{-1}.$$

Using the relationship between $ErlB$ and $ErlC$ from Lemma 19 (a),

$$\begin{aligned} I(\mu_1, \mu) &= \left(1 + \rho \frac{\mu}{\mu_1} \left(\frac{1 - \frac{N}{ErlB(N, \rho)} + \rho}{N - \rho} \right) \right)^{-1} = \left(1 + \frac{\rho}{N} \frac{\mu}{\mu_1} \frac{1 - ErlB(N, \rho)}{1 - \frac{\rho}{N}(1 - ErlB(N, \rho))} \right)^{-1} \\ &= \frac{1 - \frac{\rho}{N}(1 - ErlB(N, \rho))}{1 - \frac{\rho}{N}(1 - ErlB(N, \rho)) + \frac{\rho}{N} \frac{\mu}{\mu_1} (1 - ErlB(N, \rho))} \\ &= \frac{1 - \frac{\rho}{N}(1 - ErlB(N, \rho))}{1 - \left(1 - \frac{\mu}{\mu_1} \right) \frac{\rho}{N} (1 - ErlB(N, \rho))}. \end{aligned}$$

(b): From Lemma 22, substituting $k = N$ into (A.16) yields

$$\frac{\partial I(\mu_1, \mu)}{\partial \mu_1} = \frac{1}{\mu_1} I(\mu_1, \mu) (1 - I(\mu_1, \mu)) = \frac{I(\mu_1, \mu)^2}{\mu_1} \left(\frac{1}{I(\mu_1, \mu)} - 1 \right) > 0, \quad \forall \mu_1, \mu > 0,$$

because $I(\mu_1, \mu) \in (0, 1)$ by definition.

(c): From Lemma 22, substituting $k = N$ into (A.17) yields

$$\frac{\partial^2 I(\mu_1, \mu)}{\partial \mu_1^2} = -2 \frac{I(\mu_1, \mu)^2}{\mu_1} (1 - I(\mu_1, \mu)) < 0, \quad \forall \mu_1, \mu > 0,$$

because $I(\mu_1, \mu) \in (0, 1)$ by definition.

■

A.5.1.2 Proof of Corollary 4

This immediately follows from Lemma 26 by substituting $\mu_1 = \mu$.

■

A.5.1.3 Proof of Lemma 27

Note that

$$\frac{d}{d\mu} \left(\frac{I(\mu, \mu)}{\mu} \right) = \frac{1}{\mu} \frac{dI(\mu, \mu)}{d\mu} - \frac{I(\mu, \mu)}{\mu^2} = \frac{I(\mu, \mu)}{\mu^2} \left(\frac{\mu}{I(\mu, \mu)} \frac{dI(\mu, \mu)}{d\mu} - 1 \right).$$

From Lemma 24,

$$\begin{aligned} \frac{\mu}{I(\mu, \mu)} \frac{dI(\mu, \mu)}{d\mu} - 1 &= \frac{N\mu}{I(\mu, \mu)} \left. \frac{\partial I(\mu_1, \mu)}{\partial \mu_1} \right|_{\mu_1=\mu} - \frac{\lambda}{\mu} \left(1 - \frac{1 - ErlC(N, \frac{\lambda}{\mu})}{N - \frac{\lambda}{\mu}} \right) - 1 \\ &\stackrel{(i)}{=} N\mu \frac{1}{\mu} (1 - I(\mu, \mu)) - \frac{\lambda}{\mu} \left(1 - \frac{1 - ErlC(N, \frac{\lambda}{\mu})}{N - \frac{\lambda}{\mu}} \right) - 1 \\ &\stackrel{(ii)}{=} N \frac{\lambda}{N\mu} \left(1 - ErlB(N, \frac{\lambda}{\mu}) \right) - \frac{\lambda}{\mu} \left(1 - \frac{1 - ErlC(N, \frac{\lambda}{\mu})}{N - \frac{\lambda}{\mu}} \right) - 1 \\ &\stackrel{(iii)}{=} \frac{\lambda}{\mu} \left(1 - ErlB(N, \frac{\lambda}{\mu}) \right) - \frac{\lambda}{\mu} + \frac{ErlC(N, \frac{\lambda}{\mu})}{ErlB(N, \frac{\lambda}{\mu})} - 1 - 1 \\ &= \frac{ErlC(N, \frac{\lambda}{\mu})}{ErlB(N, \frac{\lambda}{\mu})} - \frac{\lambda}{\mu} ErlB(N, \frac{\lambda}{\mu}) - 2, \end{aligned}$$

where (i) follows from Corollary 4 (b), (ii) follows from Corollary 4 (a), and (iii) follows by noting that $\frac{\lambda}{\mu} \left(\frac{1 - ErlC(N, \frac{\lambda}{\mu})}{N - \frac{\lambda}{\mu}} \right) = \frac{ErlC(N, \frac{\lambda}{\mu})}{ErlB(N, \frac{\lambda}{\mu})} - 1$ (using the relationship between $ErlB$ and $ErlC$ from Lemma 19 (a)).

Note that we can use Lemma 19 (a) to evaluate

$$\begin{aligned} &\frac{ErlC(N, \frac{\lambda}{\mu})}{ErlB(N, \frac{\lambda}{\mu})} - \frac{\lambda}{\mu} ErlB(N, \frac{\lambda}{\mu}) - 2 \\ &= \frac{\frac{N}{\frac{N-\lambda}{\mu} + \frac{\lambda}{\mu}}}{ErlB(N, \frac{\lambda}{\mu})} - \frac{\lambda}{\mu} ErlB(N, \frac{\lambda}{\mu}) - 2 \end{aligned}$$

$$\begin{aligned}
&= \frac{N}{N - \frac{\lambda}{\mu} + \frac{\lambda}{\mu} ErlB(N, \frac{\lambda}{\mu})} - \frac{\lambda}{\mu} ErlB(N, \frac{\lambda}{\mu}) - 2 \\
&= \frac{\frac{\lambda}{\mu} \left(2 + \frac{\lambda}{\mu} ErlB(N, \frac{\lambda}{\mu})\right) \left(1 - ErlB(N, \frac{\lambda}{\mu})\right) - N \left(1 + \frac{\lambda}{\mu} ErlB(N, \frac{\lambda}{\mu})\right)}{N - \frac{\lambda}{\mu} \left(1 - ErlB(N, \frac{\lambda}{\mu})\right)},
\end{aligned}$$

where the denominator is strictly positive because, from Lemma 19 (a), $ErlC(N, \frac{\lambda}{\mu}) = \frac{NErlB(N, \frac{\lambda}{\mu})}{N - \frac{\lambda}{\mu} + \frac{\lambda}{\mu} ErlB(N, \frac{\lambda}{\mu})} > 0$ implies $N - \frac{\lambda}{\mu} + \frac{\lambda}{\mu} ErlB(N, \frac{\lambda}{\mu}) > 0$. Thus, the above display implies that, to show $\frac{ErlC(N, \frac{\lambda}{\mu})}{ErlB(N, \frac{\lambda}{\mu})} - \frac{\lambda}{\mu} ErlB(N, \frac{\lambda}{\mu}) - 2 > 0$, it suffices to show

$$\begin{aligned}
&\frac{\lambda}{\mu} \left(2 + \frac{\lambda}{\mu} ErlB(N, \frac{\lambda}{\mu})\right) \left(1 - ErlB(N, \frac{\lambda}{\mu})\right) - N \left(1 + \frac{\lambda}{\mu} ErlB(N, \frac{\lambda}{\mu})\right) > 0 \\
\Leftrightarrow &\left(\frac{2\lambda}{\mu} - N\right) \left(\frac{\mu}{\lambda ErlB(N, \frac{\lambda}{\mu})}\right)^2 + \left(\frac{\lambda}{\mu} - N - 2\right) \left(\frac{\mu}{\lambda ErlB(N, \frac{\lambda}{\mu})}\right) - 1 > 0 \\
\Leftrightarrow &\left(\frac{\mu}{\lambda ErlB(N, \frac{\lambda}{\mu})}\right)^2 + \frac{\frac{\lambda}{\mu} - N - 2}{\frac{2\lambda}{\mu} - N} \left(\frac{\mu}{\lambda ErlB(N, \frac{\lambda}{\mu})}\right) - \frac{1}{\frac{2\lambda}{\mu} - N} > 0 \\
\Leftrightarrow &\left(\frac{\mu}{\lambda ErlB(N, \frac{\lambda}{\mu})} + \frac{\frac{\lambda}{\mu} - N - 2}{2\left(\frac{2\lambda}{\mu} - N\right)}\right)^2 > \frac{\left(\frac{\lambda}{\mu} - N\right)^2 + 4\left(1 + \frac{\lambda}{\mu}\right)}{4\left(\frac{2\lambda}{\mu} - N\right)^2} \\
\Leftrightarrow &\frac{\mu}{\lambda ErlB(N, \frac{\lambda}{\mu})} + \frac{\frac{\lambda}{\mu} - N - 2}{2\left(\frac{2\lambda}{\mu} - N\right)} > \frac{\sqrt{\left(\frac{\lambda}{\mu} - N\right)^2 + 4\left(1 + \frac{\lambda}{\mu}\right)}}{2\left(\frac{2\lambda}{\mu} - N\right)} \\
\Leftrightarrow &ErlB\left(N, \frac{\lambda}{\mu}\right) < \frac{\sqrt{\left(\frac{\lambda}{\mu} - N\right)^2 + 4\left(1 + \frac{\lambda}{\mu}\right) + \left(\frac{\lambda}{\mu} - N - 2\right)}}{\frac{2\lambda}{\mu}}. \tag{A.55}
\end{aligned}$$

We prove (A.55) by induction. Suppose (A.55) holds, then we want to show that it holds

with N replaced by $N + 1$; that is,

$$\begin{aligned}
& \text{ErlB} \left(N+1, \frac{\lambda}{\mu} \right) < \frac{\sqrt{\left(\frac{\lambda}{\mu} - (N+1) \right)^2 + 4 \left(1 + \frac{\lambda}{\mu} \right)} + \left(\frac{\lambda}{\mu} - (N+1) - 2 \right)}{\frac{2\lambda}{\mu}} \\
\Leftrightarrow & \left(1 + \frac{(N+1)\mu}{\lambda} \frac{1}{\text{ErlB} \left(N, \frac{\lambda}{\mu} \right)} \right)^{-1} < \frac{\sqrt{\left(\frac{\lambda}{\mu} - (N+1) \right)^2 + 4 \left(1 + \frac{\lambda}{\mu} \right)} + \left(\frac{\lambda}{\mu} - (N+1) - 2 \right)}{\frac{2\lambda}{\mu}} \\
\Leftrightarrow & 1 + \frac{(N+1)\mu}{\lambda} \frac{1}{\text{ErlB} \left(N, \frac{\lambda}{\mu} \right)} > \frac{\frac{\lambda}{\mu} \sqrt{\left(\frac{\lambda}{\mu} - (N+1) \right)^2 + 4 \left(1 + \frac{\lambda}{\mu} \right)} - \frac{\lambda}{\mu} \left(\frac{\lambda}{\mu} - (N+1) - 2 \right)}{2 \left(\frac{2\lambda}{\mu} - (N+1) \right)} \\
\Leftrightarrow & \text{ErlB} \left(N, \frac{\lambda}{\mu} \right) < \frac{\frac{\lambda}{\mu} \sqrt{\left(\frac{\lambda}{\mu} - (N+1) \right)^2 + 4 \left(1 + \frac{\lambda}{\mu} \right)} + \left(\frac{\lambda}{\mu} - (N+1) \right) \left(2 + \frac{\lambda}{\mu} \right)}{\frac{2\lambda}{\mu} \left(\frac{\lambda}{\mu} + 1 \right)}, \tag{A.56}
\end{aligned}$$

where $(*)$ follows from (A.2). Suppose we can establish that, for all $\lambda > 0$, $N \geq 6$ and $\mu \in (0, \frac{5\lambda}{4N}]$,

$$\frac{\sqrt{\left(\frac{\lambda}{\mu} - N \right)^2 + 4 \left(1 + \frac{\lambda}{\mu} \right)} + \left(\frac{\lambda}{\mu} - N - 2 \right)}{\frac{2\lambda}{\mu}} < \frac{\frac{\lambda}{\mu} \sqrt{\left(\frac{\lambda}{\mu} - (N+1) \right)^2 + 4 \left(1 + \frac{\lambda}{\mu} \right)} + \left(\frac{\lambda}{\mu} - (N+1) \right) \left(2 + \frac{\lambda}{\mu} \right)}{\frac{2\lambda}{\mu} \left(\frac{\lambda}{\mu} + 1 \right)}, \tag{A.57}$$

then the induction hypothesis (A.55) would imply that (A.56) holds, as desired. To see (A.57), it is equivalent to show

$$\left(\frac{\lambda}{\mu} + 1 \right) \sqrt{\left(\frac{\lambda}{\mu} - N \right)^2 + 4 \left(1 + \frac{\lambda}{\mu} \right)} - \frac{\lambda}{\mu} \sqrt{\left(\frac{\lambda}{\mu} - (N+1) \right)^2 + 4 \left(1 + \frac{\lambda}{\mu} \right)} < \frac{2\lambda}{\mu} - N. \tag{A.58}$$

Note that, when $\mu \in (0, \frac{5\lambda}{4N}]$, the right-hand side of (A.58) satisfies

$$\frac{2\lambda}{\mu} - N \geq \frac{2\lambda}{\frac{5\lambda}{4N}} - N = \frac{3N}{5} > 0.$$

- If the left-hand side of (A.58) is non-positive, then (A.58) trivially holds.
- If the left-hand side of (A.58) is strictly positive, then squaring both sides of (A.58) and

algebra yields

$$\left(\frac{\lambda}{\mu} - N\right) \left(\frac{\lambda}{\mu} - N - 1\right) + 4 \left(\frac{\lambda}{\mu} + 1\right) < \sqrt{\left[\left(\frac{\lambda}{\mu} - N\right)^2 + 4 \left(1 + \frac{\lambda}{\mu}\right)\right] \left[\left(\frac{\lambda}{\mu} - (N+1)\right)^2 + 4 \left(1 + \frac{\lambda}{\mu}\right)\right]}. \quad (\text{A.59})$$

Note that the left-hand side of (A.59) satisfies

$$\left(\frac{\lambda}{\mu} - N\right) \left(\frac{\lambda}{\mu} - N - 1\right) + 4 \left(\frac{\lambda}{\mu} + 1\right) = \left(\frac{\lambda}{\mu} - N\right)^2 + N + 3\frac{\lambda}{\mu} + 4 > 0.$$

Then, squaring both sides of (A.59) and algebra implies

$$\begin{aligned} 2 \left(\frac{\lambda}{\mu} - N\right) \left(\frac{\lambda}{\mu} - N - 1\right) &< \left(\frac{\lambda}{\mu} - N\right)^2 + \left(\frac{\lambda}{\mu} - N\right)^2 + \left(\frac{\lambda}{\mu} - N - 1\right)^2 \\ \Leftrightarrow \quad \left(\frac{\lambda}{\mu} - N - 1\right)^2 - 1 &< \left(\frac{\lambda}{\mu} - N - 1\right)^2, \end{aligned}$$

which is clearly true. This concludes the induction step. Hence, (A.55) holds as desired. ■

We conclude that $\frac{I(\mu, \mu; \lambda, N)}{\mu}$ is strictly increasing in μ for $\mu \in (0, \frac{5\lambda}{4N}]$ for all $\lambda > 0$ and $N \geq 6$.

A.5.1.4 Proof of Lemma 28

From Corollary 4 (a), $I(\mu, \mu; \lambda, N) = 1 - \frac{\lambda}{N\mu}(1 - ErlB(N, \frac{\lambda}{\mu}))$, and from Lemma 18 (c), $I(\mu, \mu; \lambda, N)$ is strictly decreasing in λ . In particular,

$$\lim_{\lambda \downarrow 0} I(\mu, \mu; \lambda, N) = 1,$$

because $ErlB(N, \frac{\lambda}{\mu}) \stackrel{(i)}{=} \frac{N - \frac{\lambda}{\mu}}{\frac{N}{ErlC(N, \frac{\lambda}{\mu})} - \frac{\lambda}{\mu}} \stackrel{(ii)}{\rightarrow} 0$ as $\lambda \rightarrow 0$ (where (i) follows from the relationship between $ErlB$ and $ErlC$ in Lemma 19 (a), and (ii) follows from Lemma 19 (b))), and

$$\lim_{\lambda \uparrow \infty} I(\mu, \mu; \lambda, N) = 0,$$

because $\frac{\lambda}{N\mu}(1 - ErlB(N, \frac{\lambda}{\mu})) \stackrel{(iii)}{=} \frac{\lambda}{N\mu} \frac{\frac{N}{ErlC(N, \frac{\lambda}{\mu})} - N}{\frac{N}{ErlC(N, \frac{\lambda}{\mu})} - \frac{\lambda}{\mu}} \stackrel{(iv)}{\rightarrow} 1$ as $\lambda \rightarrow \infty$ (where (iii) follows from the relationship between $ErlB$ and $ErlC$ in Lemma 19 (a), and (iv) follows from Lemma 19 (b)). Since $I(\mu, \mu; \lambda, N)$ is a monotonic and continuous function of λ , it follows that for any $x \in (0, 1)$, there exists a unique $\lambda > 0$ such that $I(\mu, \mu; \lambda, N) = x$.

Using the continuous extension of the Erlang B formula (Jagerman (1974)), the definition of $I(\mu, \mu; \lambda, N)$ can be extended to all real-valued $N > 0$. Then, using the same technique as above,

$$\lim_{N \downarrow 0} I(\mu, \mu; \lambda, N) = 0,$$

because $\frac{\lambda}{N\mu}(1 - ErlB(N, \frac{\lambda}{\mu})) = \frac{\lambda}{N\mu} \frac{\frac{N}{ErlC(N, \frac{\lambda}{\mu})} - N}{\frac{N}{ErlC(N, \frac{\lambda}{\mu})} - \frac{\lambda}{\mu}} = \frac{\lambda}{\mu} \frac{\frac{1}{ErlC(N, \frac{\lambda}{\mu})} - 1}{\frac{N}{ErlC(N, \frac{\lambda}{\mu})} - \frac{\lambda}{\mu}} \rightarrow 1$ as $N \rightarrow 0$ (noting

that $ErlC(N, \frac{\lambda}{\mu}) \rightarrow \infty$ as $N \rightarrow 0$, from Problem 2 in Whitt, 2002, p.8), and

$$\lim_{N \uparrow \infty} I(\mu, \mu; \lambda, N) = 1,$$

because $ErlB(N, \frac{\lambda}{\mu}) = \left[\frac{\lambda}{\mu} \int_0^\infty e^{-\frac{\lambda}{\mu}y} (1+y)^N dy \right]^{-1} \rightarrow 0$ as $N \rightarrow \infty$ (from (1.5) and (1.16) in Whitt (2002)). Since (the extended) $I(\mu, \mu; \lambda, N)$ is a monotonic and continuous function of N , it follows that for any $x \in (0, 1)$, there exists a unique (real-valued) $N > 0$ such that $I(\mu, \mu; \lambda, N) = x$.

■

A.5.2 Proof of Theorem 2

Using Corollary 4 (b), the FOC (1.6) simplifies to

$$c'(\mu) = p(1 - I(\mu, \mu; \lambda, N)) + (v - p\mu) \frac{I(\mu, \mu; \lambda, N)}{\mu} (1 - I(\mu, \mu; \lambda, N)), \quad (\text{A.60})$$

which can be equivalently written as

$$c'(\mu) = \left(p(1 - I(\mu, \mu; \lambda, N)) + \frac{v}{\mu} I(\mu, \mu; \lambda, N) \right) (1 - I(\mu, \mu; \lambda, N)),$$

which establishes (1.10).

Next, note that by definition, a solution $\mu^* > 0$ to the symmetric FOC (1.10) is a symmetric equilibrium if and only if $\mu_1 = \mu^*$ is a global maximizer of the utility function $U(\mu_1, \mu^*)$ (obtained by setting $k = N$ in (1.4)). To establish the latter, we show that for the loss system, $U(\mu_1, \mu^*)$ is strictly concave; that is, $\frac{\partial^2 U(\mu_1, \mu)}{\partial \mu_1^2} < 0$, for all $\mu_1, \mu > 0$. Note that

$$\begin{aligned} \frac{\partial^2 U(\mu_1, \mu)}{\partial \mu_1^2} &= \frac{\partial^2}{\partial \mu_1^2} (p\mu_1 + (v - p\mu_1)I(\mu_1, \mu) - c(\mu_1)) \\ &= \frac{\partial}{\partial \mu_1} \left(p(1 - I(\mu_1, \mu)) - p\mu_1 \frac{\partial I(\mu_1, \mu)}{\partial \mu_1} + v \frac{\partial I(\mu_1, \mu)}{\partial \mu_1} - c'(\mu_1) \right) \\ &\stackrel{(i)}{=} \frac{\partial}{\partial \mu_1} \left(p(1 - I(\mu_1, \mu)) - p\mu_1 \frac{I(\mu_1, \mu)}{\mu_1} (1 - I(\mu_1, \mu)) + v \frac{\partial I(\mu_1, \mu)}{\partial \mu_1} - c'(\mu_1) \right) \\ &= \frac{\partial}{\partial \mu_1} \left(p(1 - I(\mu_1, \mu))^2 + v \frac{\partial I(\mu_1, \mu)}{\partial \mu_1} - c'(\mu_1) \right) \\ &= -2p(1 - I(\mu_1, \mu)) \frac{\partial I(\mu_1, \mu)}{\partial \mu_1} + v \frac{\partial^2 I(\mu_1, \mu)}{\partial \mu_1^2} - c''(\mu_1) \\ &\stackrel{(ii)}{<} 0, \quad \forall \mu_1, \mu > 0, \end{aligned}$$

where (i) follows from Lemma 26 (b), and (ii) follows because $\frac{\partial I(\mu_1, \mu)}{\partial \mu_1} > 0$ for all $\mu_1, \mu > 0$ (from Lemma 26 (b)), $\frac{\partial^2 I(\mu_1, \mu)}{\partial \mu_1^2} < 0$ for all $\mu_1, \mu > 0$ (from Lemma 26 (c)), and $c'' > 0$ (recalling that c is strictly convex). Hence, the utility function for the loss system is concave, as desired.

Finally, when $k = N$, the sufficient condition (1.7) of Theorem 1 that guarantees the existence of a solution to the symmetric FOC (1.10) simplifies to $c'(0) < p + v/\lambda$.

■

A.5.3 Proof of Proposition 2

Suppose μ_1^* , μ_2^* are symmetric equilibrium service rates such that $\mu_1^* > \mu_2^* > 0$. Setting $\mu_1 = \mu$ in (1.4) yields the server utility at any symmetric service rate μ . Therefore,

$$U(\mu, \mu) = p\mu(1 - I(\mu, \mu)) + vI(\mu, \mu) - c(\mu), \text{ for } \mu = \mu_1^*, \mu_2^*. \quad (\text{A.61})$$

The service rates μ_1^* and μ_2^* satisfy the symmetric FOC (1.10) from Theorem 2, that is,

$$\mu c'(\mu) = (U(\mu, \mu) + c(\mu))(1 - I(\mu, \mu)), \text{ for } \mu = \mu_1^*, \mu_2^*. \quad (\text{A.62})$$

Moreover, strict convexity of c implies that

$$c(\mu_1^*) - c(\mu_2^*) < (\mu_1^* - \mu_2^*)c'(\mu_1^*) < \mu_1^*c'(\mu_1^*) - \mu_2^*c'(\mu_2^*),$$

where the second inequality follows from $c'(\mu_1^*) > c'(\mu_2^*)$ for $\mu_1^* > \mu_2^*$. Thus,

$$c(\mu_1^*) - c(\mu_2^*) < \mu_1^*c'(\mu_1^*) - \mu_2^*c'(\mu_2^*). \quad (\text{A.63})$$

Substituting for $\mu_1^*c'(\mu_1^*)$ and $\mu_2^*c'(\mu_2^*)$ using (A.62), (A.63) becomes

$$c(\mu_1^*) - c(\mu_2^*) < (U(\mu_1^*, \mu_1^*) + c(\mu_1^*))(1 - I(\mu_1^*, \mu_1^*)) - (U(\mu_2^*, \mu_2^*) + c(\mu_2^*))(1 - I(\mu_2^*, \mu_2^*)),$$

which, after algebra, can be rewritten as

$$U(\mu_1^*, \mu_1^*) - U(\mu_2^*, \mu_2^*) > (U(\mu_1^*, \mu_1^*) + c(\mu_1^*))I(\mu_1^*, \mu_1^*) - (U(\mu_2^*, \mu_2^*) + c(\mu_2^*))I(\mu_2^*, \mu_2^*). \quad (\text{A.64})$$

Finally, we show that $(U(\mu, \mu) + c(\mu))I(\mu, \mu)$ is a strictly increasing function of $\mu > 0$, so that (A.64) implies $U(\mu_1^*, \mu_1^*) - U(\mu_2^*, \mu_2^*) > 0$ for $\mu_1^* > \mu_2^*$.

Let $f(\mu) := (U(\mu, \mu) + c(\mu)) I(\mu, \mu)$ for $\mu > 0$. From (A.61) and Corollary 4 (a),

$$f(\mu) = (p\mu(1 - I(\mu, \mu)) + vI(\mu, \mu)) I(\mu, \mu) = \left(p\frac{\lambda}{N} \left(1 - ErlB \left(N, \frac{\lambda}{\mu} \right) \right) + vI(\mu, \mu) \right) I(\mu, \mu),$$

where $ErlB(N, \rho)$ is strictly increasing in ρ (from Lemma 18 (a)), and thus is strictly decreasing in μ (since $\rho = \lambda/\mu$), and $I(\mu, \mu)$ is strictly increasing in μ (either by directly calculating $\frac{dI(\mu, \mu)}{d\mu}$ using Corollary 4 (a) or by applying Lemma 21 (a) twice). Hence, using the fact that the product of two strictly positive and strictly increasing functions is strictly increasing, $f(\mu)$ is strictly increasing in μ , as desired. \blacksquare

A.5.4 Proof of Proposition 3

From Theorem 2, obtaining $\mu_{\max}(p, v)$ (defined in (1.5)) for loss systems is equivalent to maximizing a solution to the FOC (1.10) over all λ, N . Let $x = I(\mu, \mu; \lambda, N)$. For $\mu \neq 0$, the FOC (1.10), after multiplying by μ on both sides, can be written as

$$\mu c'(\mu) = (p\mu(1 - x) + vx) \cdot (1 - x). \quad (\text{A.65})$$

Given any $x \in (0, 1)$, $v > 0$, and $p \geq 0$, we observe that (i) the right-hand side of (A.65) is linear in μ with a strictly positive y -intercept, $vx(1 - x)$, and a non-negative slope, $p(1 - x)^2$, and (ii) the left-hand side of (A.65) is a strictly increasing and strictly convex function (from Assumption 1) with zero y -intercept, thereby resulting in the existence of a unique solution $\mu(x; p, v) > 0$. From Lemma 28, for any $\mu > 0$ and $x \in (0, 1)$, there exists a pair (λ, N) such that $x = I(\mu, \mu; \lambda, N)$. As a result, since we are optimizing over all (λ, N) , (1.5) becomes equivalent to $\mu_{\max}^*(p, v) = \sup_{x \in (0, 1)} \mu(x; p, v)$.

We study $\mu'(x)$ and $\mu''(x)$ by differentiating (A.65) twice with respect to x , and obtain the following observations: (i) $\mu'(x) < 0$ when $\mu'(x) = 0$, and (ii) $\mu'(x) = 0$ has at most one solution in $(0, \frac{1}{2}]$ and no solutions in $(\frac{1}{2}, 1)$. To see this, we first differentiate (A.65) with

respect to x and after algebra we obtain

$$(\mu c''(\mu) + c'(\mu)) \mu'(x) = v(1 - 2x) - 2p\mu(1 - x) + p(1 - x)^2 \mu'(x),$$

which implies

$$\mu'(x) = \frac{v(1 - 2x) - 2p\mu(1 - x)}{\mu c''(\mu) + c'(\mu) - p(1 - x)^2} \stackrel{(*)}{=} \frac{v(1 - 2x) - 2p\mu(1 - x)}{\mu c''(\mu) + \frac{v}{\mu}x(1 - x)},$$

where $(*)$ follows from (A.65). Note that the above display can be rewritten as

$$\frac{1}{2}\mu'(x) = \frac{v\left(\frac{1-x}{1-x}\right) - p\mu(x)}{\frac{\mu(x)c''(\mu(x))}{1-x} + v\frac{x}{\mu(x)}}, \quad (\text{A.66})$$

which is strictly negative when $x > \frac{1}{2}$. Thus, $\mu'(x) = 0$ has no solution in $(\frac{1}{2}, 1)$ and, if it has a solution in $(0, 1)$, then that solution must lie in $(0, \frac{1}{2}]$. To see this, we take the derivative of (A.66):

$$\begin{aligned} & \frac{1}{2}\mu''(x) \\ &= \frac{\left(\frac{\mu(x)c''(\mu(x))}{1-x} + v\frac{x}{\mu(x)}\right)\left(-\frac{v}{2(1-x)^2} - p\mu'(x)\right) - \left(v\left(\frac{1-x}{1-x}\right) - p\mu(x)\right)\frac{d}{dx}\left(\frac{\mu(x)c''(\mu(x))}{1-x} + v\frac{x}{\mu(x)}\right)}{\left(\frac{\mu(x)c''(\mu(x))}{1-x} + v\frac{x}{\mu(x)}\right)^2}, \end{aligned}$$

which is strictly negative when $\mu'(x) = 0$, by noting that

$$v\left(\frac{\frac{1}{2}-x}{1-x}\right) - p\mu(x) = \frac{1}{2(1-x)}(v(1-2x) - 2p\mu(1-x)) = 0.$$

Hence, $\mu'(x) = 0$ has at most one solution in $(0, \frac{1}{2}]$ because $\mu''(x)$ evaluated at two consecutive solutions to $\mu'(x) = 0$ must have opposite signs, using the fact that the derivative of a continuous function evaluated at its two consecutive roots have opposite signs. From the

above, the two properties (i) and (ii) are proved.

Based on (i) and (ii), we can argue that if a solution to $\mu'(x) = 0$ exists, denoted by $x^* \in (0, \frac{1}{2}]$, then $\mu_{\max}^* = \mu(x^*)$; if not, then $\mu(x)$ is strictly decreasing in $x \in (0, 1)$, implying that $\mu_{\max}^* = \lim_{x \downarrow 0} \mu(x)$. Substituting $x = 0$ into (A.65), we define

$$\mu(0) := \lim_{x \downarrow 0} \mu(x) = \begin{cases} 0, & \text{if } 0 \leq p \leq c'(0), \\ (c')^{-1}(p), & \text{otherwise.} \end{cases} \quad (\text{A.67})$$

Substituting $x = 0$ into (A.66), we define

$$\mu'(0) := \lim_{x \downarrow 0} \mu'(x) = \frac{v - 2p\mu(0)}{\mu(0)c''(\mu(0)) + v \lim_{x \downarrow 0} \frac{x}{\mu(x)}}$$

Dividing (A.65) by μ on both sides and substituting for $x = 0$ yields $\lim_{x \downarrow 0} \frac{x}{\mu(x)} = \frac{c'(\mu(0))-p}{v}$.

Then, substitution into the above display using this and (A.67) yields

$$\mu'(0) = \frac{v - 2p\mu(0)}{\mu(0)c''(\mu(0)) + c'(\mu(0)) - p} = \begin{cases} \frac{v}{c'(0)-p}, & \text{if } 0 \leq p < c'(0), \\ +\infty, & \text{if } p = c'(0), \\ \frac{v-2p(c')^{-1}(p)}{(c')^{-1}(p)c''((c')^{-1}(p))}, & \text{if } p > c'(0). \end{cases} \quad (\text{A.68})$$

Next, the intermediate value theorem implies that $\mu'(x) = 0$ has a unique solution $x^* \in (0, \frac{1}{2}]$ if and only if $\mu'(0) > 0$, since $\mu'(\frac{1}{2}) < 0$ from (A.66). This condition, from (A.68), is equivalent to

$$\frac{v}{2p} > (c')^{-1}(p). \quad (\text{A.69})$$

Given $v > 0$, let $p^\dagger(v)$ be the unique solution for $p > c'(0)$ to $\frac{v}{2p} = (c')^{-1}(p)$, or, equivalently, $c'(\frac{v}{2p}) = p$. Note that $\frac{v}{2p}$ is a strictly decreasing function of p for $p > c'(0)$, with value $\frac{v}{2c'(0)}$ at $p = c'(0)$ and value 0 at $p = \infty$; $(c')^{-1}(p)$ is a strictly increasing function of p for $p > c'(0)$,

with value 0 at $p = c'(0)$ and ∞ at $p = \infty$. This implies that $p^\ddagger(v)$ must exist and is unique. Then, (A.69) is equivalent to $0 \leq p < p^\ddagger(v)$. This allows us to characterize μ_{\max}^* and $I(\mu_{\max}^*, \mu_{\max}^*)$ by considering the two cases $0 \leq p < p^\ddagger(v)$ and $p \geq p^\ddagger(v)$ separately.

Case (I): If $0 \leq p < p^\ddagger(v)$, then $\mu'(x) = 0$ has a unique solution $x^* \in (0, \frac{1}{2}]$. From (A.66),

$$v \left(\frac{\frac{1}{2} - x^*}{1 - x^*} \right) = p\mu(x^*).$$

Note that $\frac{\frac{1}{2} - x^*}{1 - x^*}$ is a strictly decreasing function of $x^* \in (0, \frac{1}{2}]$, with value $\frac{1}{2}$ at $x^* = 0$ and value 0 at $x^* = \frac{1}{2}$; that is, $\frac{\frac{1}{2} - x^*}{1 - x^*} < \frac{1}{2}$ for all $x^* \in (0, \frac{1}{2}]$. Thus, the above equation implies that $\mu_{\max}^* = \mu(x^*) < \frac{v}{2p}$. Additionally, the above equation implies

$$I(\mu_{\max}^*, \mu_{\max}^*) = x^* = \frac{v - 2p\mu_{\max}^*}{2v - 2p\mu_{\max}^*},$$

recalling that $\mu_{\max}^* = \mu(x^*)$ and $x^* = I(\mu_{\max}^*, \mu_{\max}^*)$. Substituting the expression for x^* into (A.65), we obtain

$$\mu_{\max}^* c'(\mu_{\max}^*) = \frac{v^2}{4(v - p\mu_{\max}^*)}, \quad (\text{A.70})$$

equivalently, μ_{\max}^* solves

$$(v - p\mu) \mu c'(\mu) = \frac{v^2}{4}. \quad (\text{A.71})$$

In what follows, we verify that (A.71) has a unique solution in $(0, \frac{v}{2p})$. Note that the right-hand side of (A.71) is a constant, and the left-hand side of (A.71), denoted by $LHS(\mu)$, is a strictly increasing function of μ for $\mu \in (0, \frac{v}{2p})$. To see this, differentiating $LHS(\mu)$ with

respect to μ :

$$\begin{aligned} LHS'(\mu) &= (v - p\mu)(\mu c''(\mu) + c'(\mu)) - p\mu c'(\mu) \\ &= (v - p\mu)\mu c''(\mu) + (v - 2p\mu)c'(\mu) > 0, \end{aligned}$$

when $\mu \in (0, \frac{v}{2p})$, noting that $c' > 0$ and $c'' > 0$. Moreover, note that $LHS(0) = 0 < \frac{v^2}{4}$, and $LHS(\frac{v}{2p}) = \frac{v^2}{4} \frac{1}{p} c'(\frac{v}{2p}) > \frac{v^2}{4}$ because $p < p^\dagger(v)$; therefore, it follows that (A.71) admits a unique solution in $(0, \frac{v}{2p})$; that is, there exists a unique μ_{\max}^* in $(0, \frac{v}{2p})$.

When p increases, the left-hand side of (A.71) (which is a strictly increasing function of μ) decreases for all μ , and the right-hand side of (A.71) remains unchanged for each μ , so μ_{\max}^* (which is the solution to (A.71)) is strictly increasing in p .

Case (II): If $p \geq p^\dagger(v)$, (A.67) implies that $\mu_{\max}^* = \mu(0) = (c')^{-1}(p)$ (since $p \geq c'(\frac{v}{2p}) > c'(0)$ where the first inequality follows from $p \geq p^\dagger(v)$), which is strictly increasing in p (since c' is strictly increasing by strict convexity of c , and the inverse of a strictly monotone function is strictly monotone). Hence, $c'(\mu_{\max}^*) = p$ and $I(\mu_{\max}^*, \mu_{\max}^*) = 0$. ■

A.6 Proofs From Section 1.4.2

From (1.11) and (1.12), the server's utility function is given by

$$U(\mu; \lambda, k, p, v) = \frac{v - p\mu}{1 + \sum_{i=1}^k \left(\frac{\lambda}{\mu}\right)^i} + p\mu - c(\mu). \quad (\text{A.72})$$

A.6.1 Proof of Lemma 4

The server's utility function (A.72), after algebra, can be equivalently written as

$$U(\mu; \lambda, k, p, v) = \frac{v}{1 + \sum_{i=1}^k \left(\frac{\lambda}{\mu}\right)^i} + \frac{p\lambda}{\frac{\lambda}{\mu} + \left(1 + \sum_{i=1}^{k-1} \left(\frac{\lambda}{\mu}\right)^i\right)^{-1}} - c(\mu).$$

Then, it is clear that $\lim_{\mu \downarrow 0} U(\mu; \lambda, k, p, v) = 0$ and $\lim_{\mu \rightarrow \infty} U(\mu; \lambda, k, p, v) = -\infty$. Since $U(\mu; \lambda, k, p, v)$ is a continuous function of $\mu > 0$, if $U(\mu; \lambda, k, p, v)$ is non-negative for some μ , it must attain a global maximum for some $\mu \in (0, \infty)$, i.e., there exists an equilibrium.

■

A.6.2 Proof of Proposition 4

When $p \leq c'(0)$, $c'(\mu) > c'(0) \geq p$ for all $\mu > 0$, because c' is a strictly increasing function (recalling strict convexity of c). Thus, $c'(\mu^*) > p$ for any equilibrium $\mu^* > 0$. By definition, any equilibrium $\mu^* > 0$ satisfies the FOC (1.13), where $c'(\mu^*) > p$ ensures that the left-hand side of (1.13) is strictly positive, implying that the right-hand side of (1.13) is also strictly positive. Hence, $\mu^* < \frac{v}{p}$.

Suppose that $\mu^*(k) > 0$ is a server equilibrium for an $M/M/1/k$ queueing system for some $k \in \mathbb{Z}_+ \cup \{\infty\}$. Then, $U(\mu^*(k); \lambda, k, p, v) \geq 0$ (from Proposition 1). Note that, when $\mu < \frac{v}{p}$, the server's utility function (A.72) is strictly decreasing in k for all $k \in \mathbb{Z}_+ \cup \{\infty\}$; in particular, $\lim_{k \rightarrow \infty} U(\mu; \lambda, k, p, v) < U(\mu; \lambda, k, p, v)$ for all $k \in \mathbb{Z}_+$ when $\mu < \frac{v}{p}$. Thus, for any $1 \leq k' \leq k$, $U(\mu^*(k); \lambda, k', p, v) \geq U(\mu^*(k); \lambda, k, p, v) \geq 0$ (since $\mu^*(k) < \frac{v}{p}$). Therefore, there exists some $\mu > 0$ (namely, $\mu^*(k)$) for which $U(\mu; \lambda, k', p, v) \geq 0$. From Lemma 4, we conclude that there exists an equilibrium $\mu^*(k') > 0$ in the $M/M/1/k'$ system. ■

A.6.3 Proof of Theorem 3

Differentiating $U(\mu; \lambda, k, p, v)$ in (A.72) with respect to μ yields

$$U'(\mu; \lambda, k, p, v) = \left(\frac{v}{\mu} - p \right) \frac{\sum_{i=0}^k i \left(\frac{\lambda}{\mu} \right)^i}{\left(\sum_{i=0}^k \left(\frac{\lambda}{\mu} \right)^i \right)^2} - \left(c'(\mu) - p + \frac{p}{\sum_{i=0}^k \left(\frac{\lambda}{\mu} \right)^i} \right) \quad (\text{A.73})$$

$$= \left(\frac{v - p\mu}{\lambda} \right) \frac{k + \sum_{i=1}^k (k-i) \left(\frac{\mu}{\lambda} \right)^i}{\left(1 + \sum_{i=1}^k \left(\frac{\mu}{\lambda} \right)^i \right)^2} \left(\frac{\mu}{\lambda} \right)^{k-1} - \frac{p}{1 + \sum_{i=1}^k \left(\frac{\lambda}{\mu} \right)^i} + p - c'(\mu). \quad (\text{A.74})$$

(a): When $k = 1$, from (A.73),

$$U'(\mu; \lambda, 1, p, v) = \left(\frac{v}{\mu} - p \right) \frac{\frac{\lambda}{\mu}}{\left(1 + \frac{\lambda}{\mu} \right)^2} - \left(c'(\mu) - p + \frac{p}{1 + \frac{\lambda}{\mu}} \right) = \left(\frac{\lambda}{\lambda + \mu} \right)^2 \left(\frac{v}{\lambda} + p \right) - c'(\mu),$$

which is strictly decreasing in $\mu \in (0, \infty)$, with value $\frac{v}{\lambda} + p - c'(0)$ when $\mu \rightarrow 0$ and value $-\infty$ when $\mu \rightarrow \infty$. Therefore, $\mu^*(1)$ exists if and only if $U'(\mu; \lambda, 1, p, v)$ at $\mu = 0$ is strictly positive; that is, $p > c'(0) - \frac{v}{\lambda}$. Hence, when $p \leq c'(0) - \frac{v}{\lambda}$, $\mu^*(1) > 0$ does not exist. Then, Proposition 4 implies that $\mu^*(k) > 0$ does not exist for any $k \in \mathbb{Z}_+ \cup \{\infty\}$, when $p \leq c'(0) - \frac{v}{\lambda}$.

(b): When $c'(0) - \frac{v}{\lambda} < p \leq c'(0)$, we first note that $\mu^*(1)$ exists from the proof of (a) since $c'(0) - \frac{v}{\lambda} < p$. Then, it suffices to derive the existence conditions for an $M/M/1/k$ system when $k \rightarrow \infty$, i.e., an infinite-buffer $M/M/1$ system. Then, Proposition 4 guarantees that an equilibrium exists for all $k \in \mathbb{Z}_+ \cup \{\infty\}$.

We begin by evaluating the derivative of $\lim_{k \rightarrow \infty} U(\mu; \lambda, k, p, v)$ with respect to μ , denoted by $(\lim_{k \rightarrow \infty} U(\mu; \lambda, k, p, v))'$. From (A.72),

$$\lim_{k \rightarrow \infty} U(\mu; \lambda, k, p, v) = \begin{cases} p\mu - c(\mu), & \mu \leq \lambda, \\ v + p\lambda - \frac{v\lambda}{\mu} - c(\mu), & \mu > \lambda. \end{cases}$$

Differentiating the above display with respect to μ yields

$$\left(\lim_{k \rightarrow \infty} U(\mu; \lambda, k, p, v) \right)' = \begin{cases} p - c'(\mu), & \mu \leq \lambda, \\ \frac{v\lambda}{\mu^2} - c'(\mu), & \mu > \lambda. \end{cases}$$

It is clear that

- $(\lim_{k \rightarrow \infty} U(\mu; \lambda, k, p, v))'$ is strictly decreasing in $\mu \in (0, \lambda]$, with value $p - c'(0)$ at $\mu = 0$ and $p - c'(\lambda)$ at $\mu = \lambda$;
- $(\lim_{k \rightarrow \infty} U(\mu; \lambda, k, p, v))'$ is strictly decreasing in $\mu \in (\lambda, \infty)$, with value $\frac{v}{\lambda} - c'(\lambda)$ as

$\mu \rightarrow \lambda+$ and $-\infty$ as $\mu \rightarrow \infty$;

- $\lim_{k \rightarrow \infty} U(\mu; \lambda, k, p, v)$ is not differentiable at $\mu = \lambda$ when $p \neq \frac{v}{\lambda}$.

When $p \leq c'(0)$, $p - c'(\mu) < p - c'(0) \leq 0$, i.e., $(\lim_{k \rightarrow \infty} U(\mu; \lambda, k, p, v))'$ is strictly negative for all $\mu \in (0, \lambda]$. This implies that $\lim_{k \rightarrow \infty} U(\mu; \lambda, k, p, v)$ is strictly negative for all $\mu \in (0, \lambda]$ (recalling that $\lim_{k \rightarrow \infty} U(0; \lambda, k, p, v) = 0$). Therefore, an equilibrium, if exists, must be underloaded. Recall that $(\lim_{k \rightarrow \infty} U(\mu; \lambda, k, p, v))'$ is strictly decreasing in $\mu \in (\lambda, \infty)$, the solution to $(\lim_{k \rightarrow \infty} U(\mu; \lambda, k, p, v))' = 0$ in (λ, ∞) , if exists, is unique, and a local maximizer. Hence, when $p \leq c'(0)$, there exists an equilibrium $\mu^*(k) > \lambda$ as $k \rightarrow \infty$ if and only if

- $(\lim_{k \rightarrow \infty} U(\mu; \lambda, k, p, v))' > 0$ when $\mu \rightarrow \lambda+$ (given that $\lim_{k \rightarrow \infty} U(\mu; \lambda, k, p, v)$ is continuous at $\mu = \lambda$), and
- The solution $\mu^{*,?} \in (\lambda, \infty)$ that satisfies $(\lim_{k \rightarrow \infty} U(\mu^{*,?}; \lambda, k, p, v))' = 0$ also satisfies $\lim_{k \rightarrow \infty} U(\mu^{*,?}; \lambda, k, p, v) \geq \lim_{k \rightarrow \infty} U(0; \lambda, k, p, v)$.

Condition (i) can be equivalently written as

$$\frac{v}{\lambda} - c'(\lambda) > 0 \Leftrightarrow v > \lambda c'(\lambda).$$

For condition (ii), note that the solution $\mu^{*,?} \in (\lambda, \infty)$ to $(\lim_{k \rightarrow \infty} U(\mu^{*,?}; \lambda, k, p, v))' = 0$ satisfies

$$\frac{v\lambda}{(\mu^{*,?})^2} - c'(\mu^{*,?}) = 0. \quad (\text{A.75})$$

Then, $\lim_{k \rightarrow \infty} U(\mu^{*,?}; \lambda, k, p, v) \geq \lim_{k \rightarrow \infty} U(0; \lambda, k, p, v) = 0$ can be equivalently written as

$$\frac{v\lambda}{\mu^{*,?}} + c(\mu^{*,?}) \leq v + p\lambda.$$

Substitution using (A.75) yields

$$\mu^{\star,?} c'(\mu^{\star,?}) + c(\mu^{\star,?}) \leq v + p\lambda.$$

(c): When $p > c'(0)$. From (A.74), note that

$$\begin{aligned} \lim_{\mu \downarrow 0} U'(\mu; \lambda, k, p, v) &= \lim_{\mu \downarrow 0} \left[\left(\frac{v - p\mu}{\lambda} \right) \frac{k + \sum_{i=1}^k (k-i) \left(\frac{\mu}{\lambda} \right)^i}{\left(1 + \sum_{i=1}^k \left(\frac{\mu}{\lambda} \right)^i \right)^2} \left(\frac{\mu}{\lambda} \right)^{k-1} - \frac{p}{1 + \sum_{i=1}^k \left(\frac{\lambda}{\mu} \right)^i} + p - c'(\mu) \right] \\ &= \frac{v}{\lambda} \cdot \lim_{\mu \downarrow 0} \frac{k}{\left(1 - \left(\frac{\mu}{\lambda} \right)^{k+1} \right)^2} \cdot \lim_{\mu \downarrow 0} \left(\frac{\mu}{\lambda} \right)^{k-1} + p - c'(0) \\ &= \frac{vk}{\lambda} \cdot \lim_{\mu \downarrow 0} \left(\frac{\mu}{\lambda} \right)^{k-1} + p - c'(0) > 0, \quad \text{for all } k \geq 1, \end{aligned}$$

noting that $\lim_{\mu \downarrow 0} \left(\frac{\mu}{\lambda} \right)^{k-1} \geq 0$, and $p - c'(0) > 0$ by assumption. Recalling that $\lim_{\mu \downarrow 0} U(\mu; \lambda, k, p, v) = 0$, $\lim_{\mu \downarrow 0} U'(\mu; \lambda, k, p, v) > 0$ implies that there must exist some $\mu > 0$ for which $U(\mu; \lambda, k, p, v) > 0$. Lemma 4 then guarantees an equilibrium. \blacksquare

A.7 Proofs From Section 1.5

A.7.1 Preliminaries A: Asymptotic Properties of Erlang Formulae Under Linear Staffing

We present the following asymptotic properties of the Erlang B and Erlang C Formulae, which are useful for the proofs from Section 1.5.

Lemma 29 (Asymptotic Properties of Erlang Formulae). *The following hold under linear staffing (1.14).*

(a)

$$\lim_{\lambda \rightarrow \infty} \text{ErlB} \left(N^\lambda, \frac{\lambda}{\mu} \right) = \left(1 - \frac{\mu}{a} \right)^+ = \begin{cases} 1 - \frac{\mu}{a}, & \mu < a, \\ 0, & \mu \geq a. \end{cases}$$

(b)

$$\lim_{\lambda \rightarrow \infty} ErlC \left(N^\lambda, \frac{\lambda}{\mu} \right) = \begin{cases} \infty, & \mu < a, \\ 0, & \mu > a, \end{cases}$$

and, when $\mu = a$,

$$\lim_{\lambda \rightarrow \infty} ErlC \left(N^\lambda, \frac{\lambda}{a} \right) = \begin{cases} \infty, & 0 < \frac{\lambda}{a} - N^\lambda \in \omega(\sqrt{\lambda}) \cap o(\lambda), \\ \left(1 - \frac{z\Phi^c(z)}{\phi(z)} \right)^{-1} \in (1, \infty), & 0 < \frac{\lambda}{a} - N^\lambda \in \Theta(\sqrt{\lambda}), \\ 1, & |N^\lambda - \frac{\lambda}{a}| \in o(\sqrt{\lambda}), \\ \left(1 - \frac{z\Phi^c(z)}{\phi(z)} \right)^{-1} \in (0, 1), & 0 < N^\lambda - \frac{\lambda}{a} \in \Theta(\sqrt{\lambda}), \\ 0, & 0 < N^\lambda - \frac{\lambda}{a} \in \omega(\sqrt{\lambda}) \cap o(\lambda), \end{cases}$$

where $z = \lim_{\lambda \rightarrow \infty} \frac{\frac{\lambda}{a} - N^\lambda}{\sqrt{N^\lambda}}$, $\Phi^c(z) = \frac{1}{\sqrt{2\pi}} \int_z^\infty e^{-\frac{t^2}{2}} dt$ and $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$.

Lemma 30 (Asymptotic Properties of Functions of Erlang Formulae). *The following hold under linear staffing (1.14).*

(a) If $\mu < a$, then $\lim_{\lambda \rightarrow \infty} \frac{ErlC(N^\lambda, \frac{\lambda}{\mu})}{\lambda} = \frac{(a-\mu)^2}{a^2\mu}$. That is, $ErlC(N^\lambda, \frac{\lambda}{\mu})$ converges to ∞ linearly fast as $\lambda \rightarrow \infty$.

(b) If $\mu > a$, then $\lim_{\lambda \rightarrow \infty} P(\lambda)ErlC(N^\lambda, \frac{\lambda}{\mu}) = 0$, where $P(\lambda)$ represents any polynomial in λ . That is, $ErlC(N^\lambda, \frac{\lambda}{\mu})$ converges to zero super-polynomially fast as $\lambda \rightarrow \infty$.

(c) If $|N^\lambda - \frac{\lambda}{a}| \in \mathcal{O}(\sqrt{\lambda})$, then $\lim_{\lambda \rightarrow \infty} \sqrt{\lambda} \left(\frac{1 - ErlC(N^\lambda, \frac{\lambda}{a})}{N^\lambda - \frac{\lambda}{a}} \right) \in (0, \infty)$ and $\lim_{\lambda \rightarrow \infty} \sqrt{\lambda} ErlB(N^\lambda, \frac{\lambda}{a}) \in (0, \infty)$.

(d) $\lim_{\lambda \rightarrow \infty} \lambda \left(\frac{1 - ErlC(N^\lambda, \frac{\lambda}{a})}{N^\lambda - \frac{\lambda}{a}} \right) = \infty$.

A.7.1.1 Proof of Lemma 29

For simplicity, we use B , C and ρ^λ to denote $ErlB$, $ErlC$ and λ/μ . The proof relies on the following properties of the Erlang-B formula (given in (A.1)), whose proofs are delayed to the end of this subsection.

Lemma 31. *If $|N^\lambda - \rho^\lambda| \in o(\sqrt{\lambda})$, then $\lim_{\lambda \rightarrow \infty} \sqrt{\rho^\lambda} B(N^\lambda, \rho^\lambda) \in \left[\frac{1}{\sqrt{2}}, 1\right]$.*

Lemma 32. *If $N^\lambda = \rho^\lambda - z\sqrt{\rho^\lambda} + o(\sqrt{\rho^\lambda})$ for real $z \neq 0$, then $\lim_{\lambda \rightarrow \infty} \sqrt{\rho^\lambda} B(N^\lambda, \rho^\lambda) = \frac{\phi(z)}{\Phi^c(z)}$.*

- **When** $|N^\lambda - \rho^\lambda| \in o(\sqrt{\lambda})$: From Lemma 19 (a),

$$\lim_{\lambda \rightarrow \infty} C(N^\lambda, \rho^\lambda) = \lim_{\lambda \rightarrow \infty} \frac{N^\lambda}{\frac{N^\lambda - \rho^\lambda}{B(N^\lambda, \rho^\lambda)} + \rho^\lambda} = \lim_{\lambda \rightarrow \infty} \frac{\frac{N^\lambda}{\rho^\lambda}}{\frac{N^\lambda - \rho^\lambda}{\sqrt{\rho^\lambda}} + 1} = 1,$$

from Lemma 31 and since $\lim_{\lambda \rightarrow \infty} \frac{N^\lambda}{\rho^\lambda} = 1$ and $\lim_{\lambda \rightarrow \infty} \frac{N^\lambda - \rho^\lambda}{\sqrt{\rho^\lambda}} = 0$.

- **When** $|N^\lambda - \rho^\lambda| \in \Theta(\sqrt{\lambda})$: Without loss of generality, let $N^\lambda = \rho^\lambda - z\sqrt{\rho^\lambda} + o(\sqrt{\rho^\lambda})$ for real $z \neq 0$. Note that $z = \lim_{\lambda \rightarrow \infty} \frac{\rho^\lambda - N^\lambda}{\sqrt{N^\lambda}}$.

From Lemma 19 (a),

$$\begin{aligned} \lim_{\lambda \rightarrow \infty} C(N^\lambda, \rho^\lambda) &= \lim_{\lambda \rightarrow \infty} \frac{N^\lambda}{\frac{N^\lambda - \rho^\lambda}{B(N^\lambda, \rho^\lambda)} + \rho^\lambda} \\ &= \lim_{\lambda \rightarrow \infty} \frac{N^\lambda}{\rho^\lambda + \left(-z\sqrt{\rho^\lambda} + o(\sqrt{\rho^\lambda})\right) B(N^\lambda, \rho^\lambda)^{-1}} \\ &= \lim_{\lambda \rightarrow \infty} \frac{\frac{N^\lambda}{\rho^\lambda}}{1 + \left(-z + \frac{o(\sqrt{\rho^\lambda})}{\sqrt{\rho^\lambda}}\right) \frac{1}{\sqrt{\rho^\lambda}} B(N^\lambda, \rho^\lambda)^{-1}} \\ &= \frac{1}{1 - \frac{z\Phi^c(z)}{\phi(z)}}, \end{aligned} \tag{A.76}$$

where the last equality follows from Lemma 32. Furthermore,

- When $0 < \rho^\lambda - N^\lambda \in \Theta(\sqrt{\lambda})$, $z > 0$ and $\frac{z\Phi^c(z)}{\phi(z)} \in (0, 1)$; thus, $\left(1 - \frac{z\Phi^c(z)}{\phi(z)}\right)^{-1} \in (1, \infty)$.
- When $0 < N^\lambda - \rho^\lambda \in \Theta(\sqrt{\lambda})$, $z < 0$; thus, $\left(1 - \frac{z\Phi^c(z)}{\phi(z)}\right)^{-1} \in (0, 1)$.
- **When** $0 < \rho^\lambda - N^\lambda \in \omega(\sqrt{\lambda})$: Let $f(\lambda) = \rho^\lambda - N^\lambda$. Let $g(\lambda; z) = \rho^\lambda - z\sqrt{\rho^\lambda} + o(\sqrt{\rho^\lambda})$ for $z > 0$, so that $0 < \rho^\lambda - g(\lambda; z) \in \Theta(\sqrt{\lambda})$.

By definition of the ω and Θ notations, for any $z > 0$, there exists $\Lambda(z)$ such that $f(\lambda) \geq \rho^\lambda - g(\lambda; z) > 0$ for all $\lambda \geq \Lambda(z)$. Since $C(N, \rho)$ is strictly decreasing in N (from Lemma 18 (b)), it follows that, for all $\lambda \geq \Lambda(z)$,

$$C(N^\lambda, \rho^\lambda) = C(\rho^\lambda - f(\lambda), \rho^\lambda) \geq C(g(\lambda; z), \rho^\lambda),$$

implying that

$$\lim_{\lambda \rightarrow \infty} C(N^\lambda, \rho^\lambda) \geq \lim_{\lambda \rightarrow \infty} C(g(\lambda; z), \rho^\lambda) = \left(1 - \frac{z\Phi^c(z)}{\phi(z)}\right)^{-1},$$

from (A.76). Since $z > 0$ is chosen arbitrarily, we have

$$\lim_{\lambda \rightarrow \infty} C(N^\lambda, \rho^\lambda) \geq \sup_{z > 0} \left(1 - \frac{z\Phi^c(z)}{\phi(z)}\right)^{-1} = \infty.$$

Therefore,

$$\lim_{\lambda \rightarrow \infty} C(N^\lambda, \rho^\lambda) = \infty.$$

- **When** $0 < N^\lambda - \rho^\lambda \in \omega(\sqrt{\lambda})$. Let $f(\lambda) = N^\lambda - \rho^\lambda$. Let $g(\lambda; z) = \rho^\lambda - z\sqrt{\rho^\lambda} + o(\sqrt{\rho^\lambda})$ for $z < 0$, so that $0 < g(\lambda; z) - \rho^\lambda \in \Theta(\sqrt{\lambda})$.

By definition of the ω and Θ notations, for any $z < 0$, there exists $\Lambda(z)$ such that $f(\lambda) \geq g(\lambda; z) - \rho^\lambda > 0$ for all $\lambda \geq \Lambda(z)$. Since $C(N, \rho)$ is strictly decreasing in N (from Lemma 18 (b)), it follows that, for all $\lambda \geq \Lambda(z)$,

$$C(N^\lambda, \rho^\lambda) = C(\rho^\lambda + f(\lambda), \rho^\lambda) \leq C(g(\lambda; z), \rho^\lambda),$$

implying that

$$\lim_{\lambda \rightarrow \infty} C(N^\lambda, \rho^\lambda) \leq \lim_{\lambda \rightarrow \infty} C(g(\lambda; z), \rho^\lambda) = \left(1 - \frac{z\Phi^c(z)}{\phi(z)}\right)^{-1},$$

from (A.76). Since z is chosen arbitrarily, we have

$$\lim_{\lambda \rightarrow \infty} C(N^\lambda, \rho^\lambda) \leq \inf_{z < 0} \left(1 - \frac{z\Phi^c(z)}{\phi(z)}\right)^{-1} = 0.$$

Since $C(N, \rho) \geq 0$ for all N and ρ , we have

$$\lim_{\lambda \rightarrow \infty} C(N^\lambda, \rho^\lambda) = 0.$$

■

To complete the proof, we prove Lemmas 31 and 32 as follows.

Proof of Lemma 31

From Proposition 1 in Harel (2010), when $N \geq 2$, the upper and lower bounds for the Erlang-B formula are given by

$$B(N, \rho) \leq \frac{-\frac{2}{N}\rho + \left(\frac{\rho}{N} - 1\right) + \sqrt{\left(\frac{\rho}{N} - 1\right)^2 + \frac{4}{N}\rho}}{2\left(1 - \frac{1}{N}\right)\frac{\rho}{N}}, \quad (\text{A.77})$$

and

$$B(N, \rho) \geq \frac{-\frac{2}{N} + 3\left(\frac{\rho}{N} - 1\right) + \sqrt{\left(\frac{\rho}{N} - 1\right)^2 + \frac{4}{N} + \frac{4\rho}{N^2}}}{4\frac{\rho}{N}}. \quad (\text{A.78})$$

It is clear that, when $|N^\lambda - \rho^\lambda| \in o(\sqrt{\lambda})$, the denominators of the upper and lower bounds of $B(N^\lambda, \rho^\lambda)$ given above converge to 2 and 4, respectively, as $\lambda \rightarrow \infty$. Therefore, it suffices to evaluate the limits of their numerators multiplied by $\sqrt{\rho^\lambda}$ as $\lambda \rightarrow \infty$.

Multiplying the numerator of the upper bound of $B(N^\lambda, \rho^\lambda)$ (from (A.77)) by $\sqrt{\rho^\lambda}$ and evaluating the limit of the resulting expression as $\lambda \rightarrow \infty$ yields

$$\begin{aligned} & \lim_{\lambda \rightarrow \infty} \sqrt{\rho^\lambda} \left(-\frac{2}{N^\lambda} \frac{\rho^\lambda}{N^\lambda} + \left(\frac{\rho^\lambda}{N^\lambda} - 1 \right) + \sqrt{\left(\frac{\rho^\lambda}{N^\lambda} - 1 \right)^2 + \frac{4}{N^\lambda} \frac{\rho^\lambda}{N^\lambda}} \right) \\ &= \lim_{\lambda \rightarrow \infty} -\frac{\frac{2}{\sqrt{\rho^\lambda}}}{\left(\frac{N^\lambda}{\rho^\lambda} \right)^2} + \frac{\frac{\rho^\lambda - N^\lambda}{\sqrt{\rho^\lambda}}}{\sqrt{\mu} \frac{N^\lambda}{\rho^\lambda}} + \sqrt{\frac{\left(\frac{\rho^\lambda - N^\lambda}{\sqrt{\rho^\lambda}} \right)^2 + 4}{\left(\frac{N^\lambda}{\rho^\lambda} \right)^2}} = 2, \end{aligned}$$

recalling that $|N^\lambda - \rho^\lambda| \in o(\sqrt{\lambda}) = o(\sqrt{\rho^\lambda})$. Hence, $\lim_{\lambda \rightarrow \infty} \sqrt{\rho^\lambda} B(N^\lambda, \rho^\lambda)$ has an upper bound equal to 1.

Similarly, multiplying the numerator of the lower bound of $B(N^\lambda, \rho^\lambda)$ (from (A.78)) by $\sqrt{\rho^\lambda}$ and evaluating the limit of the resulting expression as $\lambda \rightarrow \infty$ yields

$$\begin{aligned} & \lim_{\lambda \rightarrow \infty} \sqrt{\rho^\lambda} \left(-\frac{2}{N^\lambda} + 3\left(\frac{\rho^\lambda}{N^\lambda} - 1\right) \right. \\ & \quad \left. + \sqrt{\left(\frac{\rho^\lambda}{N^\lambda} - 1\right)^2 + \frac{4}{N^\lambda} + \frac{4}{N^\lambda} \frac{\rho^\lambda}{N^\lambda} + \frac{4}{(N^\lambda)^2}} \right) \\ &= \lim_{\lambda \rightarrow \infty} -\frac{\frac{2}{\sqrt{\rho^\lambda}}}{\frac{N^\lambda}{\rho^\lambda}} + 3\frac{\frac{\rho^\lambda - N^\lambda}{\sqrt{\rho^\lambda}}}{\frac{N^\lambda}{\rho^\lambda}} + \sqrt{\frac{\left(\frac{\rho^\lambda - N^\lambda}{\sqrt{\rho^\lambda}} \right)^2 + 4 \frac{N^\lambda}{\rho^\lambda} + 4 + \frac{4}{\rho^\lambda}}{\left(\frac{N^\lambda}{\rho^\lambda} \right)^2}} = 2\sqrt{2}, \end{aligned}$$

recalling that $|N^\lambda - \rho^\lambda| \in o(\sqrt{\lambda}) = o(\sqrt{\rho^\lambda})$. Hence, $\lim_{\lambda \rightarrow \infty} \sqrt{\rho^\lambda} B(N^\lambda, \rho^\lambda)$ has a lower bound equal to $\frac{1}{\sqrt{2}}$.

Together, the upper and lower bounds of $\lim_{\lambda \rightarrow \infty} \sqrt{\rho^\lambda} B(N^\lambda, \rho^\lambda)$ imply

$$\lim_{\lambda \rightarrow \infty} \sqrt{\rho^\lambda} B(N^\lambda, \rho^\lambda) \in \left[\frac{1}{\sqrt{2}}, 1 \right].$$

■

Proof of Lemma 32

We adapt the proof of Theorem 14 in Jagerman (1974). Using the integral representation of the Erlang-B formula from Theorem 3 in Jagerman (1974), it follows that

$$\begin{aligned} B(N^\lambda, \rho^\lambda)^{-1} &= \rho^\lambda \int_0^\infty e^{-\rho^\lambda u} (1+u)^{N^\lambda} du \\ &= \rho^\lambda \int_0^\infty e^{-\left(N^\lambda + (\rho^\lambda - N^\lambda)\right)u} (1+u)^{N^\lambda} du. \end{aligned}$$

Let $v = \sqrt{N^\lambda} u$. Then, the above display implies that

$$\left(\sqrt{\rho^\lambda} B(N^\lambda, \rho^\lambda) \right)^{-1} = \frac{1}{\sqrt{\rho^\lambda}} \int_0^\infty e^{-\left(\frac{1}{2}v^2 + \left(\frac{\rho^\lambda - N^\lambda}{\sqrt{N^\lambda}}\right)v\right)} h(v, N^\lambda, \rho^\lambda) dv, \quad (\text{A.79})$$

where

$$h(v, N^\lambda, \rho^\lambda) := \varphi(v, N^\lambda) \frac{\rho^\lambda}{\sqrt{N^\lambda}}, \quad (\text{A.80})$$

and $\varphi(v, N^\lambda) := e^{\frac{1}{2}v^2 - \sqrt{N^\lambda}v} \left(1 + \frac{v}{\sqrt{N^\lambda}}\right)^{N^\lambda}$. We begin by deriving an asymptotic expansion for $\varphi(v, N^\lambda)$ in $\sqrt{N^\lambda}$.

$$\varphi(v, N^\lambda) \sim \sum_{j=0}^{\infty} a_j(v) (N^\lambda)^{-j/2}, \quad (\text{A.81})$$

where $a_0(v) = 1$, $a_1(v) = \frac{v^3}{3}$, $a_2(v) = -\frac{v^4}{4} + \frac{v^6}{18}$, $a_3(v) = \frac{v^5}{5} - \frac{v^7}{12} + \frac{v^9}{162}$, ...

Suppose $N^\lambda = \rho^\lambda - z\sqrt{\rho^\lambda} + o(\sqrt{\rho^\lambda})$ for real $z \neq 0$, which implies that:

$$\rho^\lambda = N^\lambda + z\sqrt{\rho^\lambda} - o(\sqrt{\rho^\lambda}). \quad (\text{A.82})$$

Using (A.82) to substitute for ρ^λ and (A.81) to substitute for $\varphi(v, N^\lambda)$ in (A.80), we obtain the following asymptotic expansion for $h(v, N^\lambda, \rho^\lambda)$:

$$\begin{aligned} h(v, N^\lambda, \rho^\lambda) &= \varphi(v, N^\lambda) \left(\sqrt{N^\lambda} + \left(z - \frac{o(\sqrt{\rho^\lambda})}{\sqrt{\rho^\lambda}} \right) \sqrt{\frac{\rho^\lambda}{N^\lambda}} \right) \\ &=: \varphi(v, N^\lambda) \left(\sqrt{N^\lambda} + t^\lambda \right) \\ &\sim \sum_{j=0}^{\infty} b_j(v, t^\lambda) (N^\lambda)^{-(j-1)/2}, \end{aligned} \quad (\text{A.83})$$

where $t^\lambda = \left(z - \frac{o(\sqrt{\rho^\lambda})}{\sqrt{\rho^\lambda}} \right) \sqrt{\frac{\rho^\lambda}{N^\lambda}}$, $b_0(v, t^\lambda) = 1$, $b_j(v, t^\lambda) = a_{j-1}(v) t^\lambda + a_j(v)$ for all $j \geq 1$.

Using (A.83) in (A.79), factoring the term $\sqrt{N^\lambda}$ out of the integral, and using (A.82) to substitute for ρ^λ in the exponents, we obtain:

$$\begin{aligned} &\left(\sqrt{\rho^\lambda} B(N^\lambda, \rho^\lambda) \right)^{-1} \\ &\sim \sqrt{\frac{N^\lambda}{\rho^\lambda}} \sum_{j=0}^{\infty} (N^\lambda)^{-j/2} \int_0^{\infty} b_j(v, t^\lambda) e^{-\left(\frac{1}{2}v^2 + t^\lambda v\right)} dv. \end{aligned} \quad (\text{A.84})$$

Finally, taking the limit as $\lambda \rightarrow \infty$, only the first term in the summation survives on the right-hand side of (A.84), because $\int_0^{\infty} b_j(v, t^\lambda) e^{-\left(\frac{1}{2}v^2 + t^\lambda v\right)} dv \in (0, \infty)$ for all $j \geq 1$ (since $b_j(v, t^\lambda)$ are polynomials in v). Hence, noting that $\lim_{\lambda \rightarrow \infty} \sqrt{\frac{N^\lambda}{\rho^\lambda}} = 1$ and $\lim_{\lambda \rightarrow \infty} t^\lambda = z$,

it follows that

$$\begin{aligned}
& \lim_{\lambda \rightarrow \infty} \left(\sqrt{\rho^\lambda} B(N^\lambda, \rho^\lambda) \right)^{-1} \\
&= \lim_{\lambda \rightarrow \infty} \sqrt{\frac{N^\lambda}{\rho^\lambda}} \sum_{j=0}^{\infty} (N^\lambda)^{-j/2} \int_0^\infty b_j(v, t^\lambda) e^{-\left(\frac{1}{2}v^2 + t^\lambda v\right)} dv \\
&= \lim_{\lambda \rightarrow \infty} \int_0^\infty e^{-\left(\frac{1}{2}v^2 + t^\lambda v\right)} dv \\
&= \lim_{\lambda \rightarrow \infty} e^{\frac{1}{2}(t^\lambda)^2} \int_0^\infty e^{-\frac{1}{2}(v+t^\lambda)^2} dv \\
&= \lim_{\lambda \rightarrow \infty} \frac{\int_{t^\lambda}^\infty e^{-\frac{1}{2}v^2} dv}{e^{-\frac{1}{2}(t^\lambda)^2}} = \lim_{\lambda \rightarrow \infty} \frac{\Phi^c(t^\lambda)}{\phi(t^\lambda)} = \frac{\Phi^c(z)}{\phi(z)}.
\end{aligned}$$

■

A.7.1.2 Proof of Lemma 30

(a): From Propositions 3 and 4 in Harel (1988), when $\frac{\lambda}{\mu} \geq N^\lambda$, a lower and upper bound for the Erlang-C formula are given by

$$ErlC\left(N^\lambda, \frac{\lambda}{\mu}\right) \geq \frac{(N^\lambda)^2}{2\frac{\lambda}{\mu}} \left[\left(\frac{\lambda}{N^\lambda \mu} - 1\right)^2 + \frac{2}{N^\lambda} \frac{\lambda}{N^\lambda \mu} + \left(\frac{\lambda}{N^\lambda \mu} - 1\right) \sqrt{\left(\frac{\lambda}{N^\lambda \mu} - 1\right)^2 + \frac{4}{N^\lambda} \frac{\lambda}{N^\lambda \mu}} \right],$$

and

$$ErlC\left(N^\lambda, \frac{\lambda}{\mu}\right) \leq \frac{(N^\lambda)^2}{2\frac{\lambda}{\mu}} \left[\left(\frac{\lambda}{N^\lambda \mu} - 1\right)^2 + \frac{2}{N^\lambda} \frac{\lambda}{N^\lambda \mu} + \left(\frac{\lambda}{N^\lambda \mu} - 1\right) \sqrt{\left(\frac{\lambda}{N^\lambda \mu} - 1\right)^2 + \frac{4}{N^\lambda} \left(\frac{\lambda}{N^\lambda \mu} + \frac{1}{N^\lambda} + 1\right)} \right].$$

When $\mu < a$, $\frac{\lambda}{\mu} > N^\lambda$ for all large enough λ , which satisfies the condition for the above bounds. Then, dividing both sides of the above inequalities by λ and letting $\lambda \rightarrow \infty$ implies

$$\lim_{\lambda \rightarrow \infty} \frac{ErlC\left(N^\lambda, \frac{\lambda}{\mu}\right)}{\lambda} \geq \frac{\mu}{2a^2} \left[\left(\frac{a}{\mu} - 1\right) + 0 + \left(\frac{a}{\mu} - 1\right) \sqrt{\left(\frac{a}{\mu} - 1\right)^2 + 0} \right] = \frac{\mu}{a^2} \left(\frac{a}{\mu} - 1\right)^2,$$

and

$$\lim_{\lambda \rightarrow \infty} \frac{ErlC(N^\lambda, \frac{\lambda}{\mu})}{\lambda} \leq \frac{\mu}{2a^2} \left[\left(\frac{a}{\mu} - 1 \right)^2 + 0 + \left(\frac{a}{\mu} - 1 \right) \sqrt{\left(\frac{a}{\mu} - 1 \right)^2 + 0} \right] = \frac{\mu}{a^2} \left(\frac{a}{\mu} - 1 \right)^2.$$

Together the above two inequalities imply

$$\lim_{\lambda \rightarrow \infty} \frac{ErlC(N^\lambda, \frac{\lambda}{\mu})}{\lambda} = \frac{\mu}{a^2} \left(\frac{a}{\mu} - 1 \right)^2 = \frac{(a - \mu)^2}{a^2 \mu},$$

meaning that $ErlC(N^\lambda, \frac{\lambda}{\mu})$ converges to ∞ linearly fast as $\lambda \rightarrow \infty$.

(b): From Proposition 2 of Harel (2010), when $N^\lambda \geq 2$ and $\frac{\lambda}{\mu} < N^\lambda$, an upper bound for the Erlang-C formula is given by

$$ErlC(N^\lambda, \frac{\lambda}{\mu}) < \left(\frac{\lambda}{N^\lambda \mu} \right)^{\sqrt{N^\lambda}}. \quad (\text{A.85})$$

Under linear staffing (1.14), $\mu > a$ implies $\frac{\lambda}{N^\lambda \mu} < 1$ for all large enough λ , which satisfies the condition for bound (A.85). Therefore, for any polynomial $P(\lambda)$, under (1.14),

$$P(\lambda)ErlC(N^\lambda, \frac{\lambda}{\mu}) < P(\lambda) \left(\frac{\lambda}{N^\lambda \mu} \right)^{\sqrt{N^\lambda}} \rightarrow 0, \quad \text{as } \lambda \rightarrow \infty,$$

noting that exponential decay dominates polynomial growth. Therefore, by non-negativity, $\lim_{\lambda \rightarrow \infty} P(\lambda)ErlC(N^\lambda, \frac{\lambda}{\mu}) = 0$, meaning that $ErlC(N^\lambda, \frac{\lambda}{\mu})$ asymptotically decays to zero in a super-polynomial fashion, when $\mu > a$.

(c): We first evaluate $\lim_{\lambda \rightarrow \infty} \sqrt{\lambda} ErlB(N^\lambda, \frac{\lambda}{a})$ when $|N^\lambda - \frac{\lambda}{a}| \in \mathcal{O}(\sqrt{\lambda})$. From Proposition 1 in Harel (2010), when $N^\lambda \geq 2$, an upper and lower bound for the Erlang-B formula

are given by

$$\begin{aligned} ErlB\left(N^\lambda, \frac{\lambda}{a}\right) &\leq \frac{-2\frac{\lambda}{N^\lambda a} - (N^\lambda - \frac{\lambda}{a}) + \sqrt{(N^\lambda - \frac{\lambda}{a})^2 + 4\frac{\lambda}{a}}}{2(1 - \frac{1}{N^\lambda}) \frac{\lambda}{a}} \\ &= \frac{-\frac{2}{N^\lambda} \frac{\lambda}{N^\lambda a} - (1 - \frac{\lambda}{N^\lambda a}) + \sqrt{(1 - \frac{\lambda}{N^\lambda a})^2 + \frac{4}{N^\lambda} \frac{\lambda}{N^\lambda a}}}{2(1 - \frac{1}{N^\lambda}) \frac{\lambda}{N^\lambda a}}, \end{aligned} \quad (\text{A.86})$$

and

$$\begin{aligned} ErlB\left(N^\lambda, \frac{\lambda}{a}\right) &\geq \frac{-2 - 3(N^\lambda - \frac{\lambda}{a}) + \sqrt{(N^\lambda - \frac{\lambda}{a})^2 + 4N^\lambda + 4\frac{\lambda}{a} + 4}}{4\frac{\lambda}{a}} \\ &= \frac{-\frac{2}{N^\lambda} - 3(1 - \frac{\lambda}{N^\lambda a}) + \sqrt{(1 - \frac{\lambda}{N^\lambda a})^2 + \frac{4}{N^\lambda} + \frac{4}{N^\lambda} \frac{\lambda}{N^\lambda a} + \frac{4}{(N^\lambda)^2}}}{4\frac{\lambda}{N^\lambda a}}. \end{aligned} \quad (\text{A.87})$$

It is clear that, when $|N^\lambda - \frac{\lambda}{a}| \in \mathcal{O}(\sqrt{\lambda})$, the denominators of the upper and lower bounds in (A.86) and (A.87) converge to 2 and 4, respectively, as $\lambda \rightarrow \infty$. It suffices to evaluate the limits of the numerators of the upper and lower bounds multiplied by $\sqrt{\lambda}$.

Multiplying the numerator of (A.86) by $\sqrt{\lambda}$ and evaluating the limit yields

$$\begin{aligned} &\lim_{\lambda \rightarrow \infty} \sqrt{\lambda} \left(-\frac{2}{N^\lambda} \frac{\lambda}{N^\lambda a} + \left(\frac{\lambda}{N^\lambda a} - 1 \right) + \sqrt{\left(\frac{\lambda}{N^\lambda a} - 1 \right)^2 + \frac{4}{N^\lambda} \frac{\lambda}{N^\lambda a}} \right) \\ &= \sqrt{a} \left\{ \lim_{\lambda \rightarrow \infty} -\frac{\frac{1}{a} \frac{2}{\sqrt{\lambda}}}{\sqrt{a} \left(\frac{N^\lambda}{\lambda} \right)^2} + \frac{\frac{\lambda}{a} - N^\lambda}{\sqrt{a} \frac{N^\lambda}{\lambda}} + \sqrt{\frac{\left(\frac{\lambda}{a} - N^\lambda \right)^2 + \frac{4}{a}}{a \left(\frac{N^\lambda}{\lambda} \right)^2}} \right\} = 2\sqrt{a} \in (0, \infty), \end{aligned}$$

recalling that $|N^\lambda - \frac{\lambda}{a}| \in \mathcal{O}(\sqrt{\lambda})$. Hence, $\lim_{\lambda \rightarrow \infty} \sqrt{\lambda} ErlB\left(N^\lambda, \frac{\lambda}{a}\right)$ has an upper bound in $(0, \infty)$.

Similarly, multiplying the numerator of (A.87) by $\sqrt{\lambda}$ and evaluating the limit as $\lambda \rightarrow \infty$

yields

$$\begin{aligned} & \lim_{\lambda \rightarrow \infty} \sqrt{\lambda} \cdot \left(-\frac{2}{N^\lambda} - 3 \left(1 - \frac{\lambda}{N^\lambda a} \right) + \sqrt{\left(1 - \frac{\lambda}{N^\lambda a} \right)^2 + \frac{4}{N^\lambda} + \frac{4}{N^\lambda} \frac{\lambda}{N^\lambda a} + \frac{4}{(N^\lambda)^2}} \right) \\ &= \sqrt{a} \left\{ \lim_{\lambda \rightarrow \infty} -\frac{-\frac{2}{\sqrt{\lambda}}}{\sqrt{a} \frac{N^\lambda}{\lambda}} + 3 \frac{\frac{\lambda-N^\lambda}{\sqrt{\lambda}}}{\sqrt{a} \frac{N^\lambda}{\lambda}} + \sqrt{\frac{\left(\frac{\lambda-N^\lambda}{\sqrt{a}} \right)^2 + 4 \frac{N^\lambda}{\lambda} + \frac{4}{a} + \frac{4}{\lambda}}{a \left(\frac{N^\lambda}{\lambda} \right)^2}} \right\} = 2\sqrt{2a} \in (0, \infty), \end{aligned}$$

recalling that $|N^\lambda - \frac{\lambda}{a}| \in \mathcal{O}(\sqrt{\lambda})$. Hence, $\lim_{\lambda \rightarrow \infty} \sqrt{\lambda} \text{ErlB} \left(N^\lambda, \frac{\lambda}{a} \right)$ has a lower bound in $(0, \infty)$.

Combining the upper and lower bounds for $\lim_{\lambda \rightarrow \infty} \sqrt{\lambda} \text{ErlB} \left(N^\lambda, \frac{\lambda}{a} \right)$ implies

$$\lim_{\lambda \rightarrow \infty} \sqrt{\lambda} \text{ErlB} \left(N^\lambda, \frac{\lambda}{a} \right) \in (0, \infty).$$

Now, using the relationship between the Erlang-B and C formulae (from Lemma 19 (a)):

$$\sqrt{\lambda} \left(\frac{1 - \text{ErlC}(N^\lambda, \frac{\lambda}{a})}{N^\lambda - \frac{\lambda}{a}} \right) = \frac{\sqrt{\lambda}(1 - \text{ErlB}(N^\lambda, \frac{\lambda}{a}))}{\frac{\lambda}{a} \text{ErlB}(N^\lambda, \frac{\lambda}{a}) + N^\lambda - \frac{\lambda}{a}} = \frac{1 - \text{ErlB}(N^\lambda, \frac{\lambda}{a})}{\frac{\sqrt{\lambda}}{a} \text{ErlB}(N^\lambda, \frac{\lambda}{a}) + \frac{N^\lambda - \frac{\lambda}{a}}{\sqrt{\lambda}}} \in (0, \infty),$$

recalling that $\lim_{\lambda \rightarrow \infty} \text{ErlB}(N^\lambda, \frac{\lambda}{a}) = 0$ (from Lemma 29 (a)), and $\lim_{\lambda \rightarrow \infty} \sqrt{\lambda} \text{ErlB}(N^\lambda, \frac{\lambda}{a}) \in (0, \infty)$ when $|N^\lambda - \frac{\lambda}{a}| = \mathcal{O}(\sqrt{\lambda})$ (as shown in the first part of the proof).

(d): Using the relationship between the Erlang-B and C formulae (from Lemma 19 (a)):

$$\begin{aligned} \lim_{\lambda \rightarrow \infty} \lambda \left(\frac{1 - \text{ErlC}(N^\lambda, \frac{\lambda}{a})}{N^\lambda - \frac{\lambda}{a}} \right) &= \lim_{\lambda \rightarrow \infty} \frac{\lambda(1 - \text{ErlB}(N^\lambda, \frac{\lambda}{a}))}{\frac{\lambda}{a} \text{ErlB}(N^\lambda, \frac{\lambda}{a}) + N^\lambda - \frac{\lambda}{a}} \\ &= \lim_{\lambda \rightarrow \infty} \frac{1 - \text{ErlB}(N^\lambda, \frac{\lambda}{a})}{\frac{1}{a} \text{ErlB}(N^\lambda, \frac{\lambda}{a}) + \frac{N^\lambda - \frac{\lambda}{a}}{\lambda}} = \infty, \end{aligned}$$

recalling that $\lim_{\lambda \rightarrow \infty} \text{ErlB}(N^\lambda, \frac{\lambda}{a}) = 0$ (from Lemma 29 (a)). ■

A.7.2 Preliminaries B: Limiting Idle Time and Derivative of Idle Time

Building on the asymptotic properties of the Erlang formulae in Section A.7.1, under linear staffing (1.14), we derive the limiting values of $I^\lambda(\mu_1, \mu) := I(\mu, \mu; \lambda, k^\lambda, N^\lambda)$ and $\frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1}$ as $\lambda \rightarrow \infty$ for any $\mu_1 > 0$ and $\mu > 0$.

Proposition 16. Fix $\mu_1 > 0$. Under linear staffing (1.14),

$$\lim_{\lambda \rightarrow \infty} I^\lambda(\mu_1, \mu) = \frac{\left(1 - \frac{a}{\mu}\right)^+}{1 - \frac{a}{\mu} + \frac{a}{\mu_1}} = \begin{cases} 0, & \mu \leq a, \\ \frac{1 - \frac{a}{\mu}}{1 - \frac{a}{\mu} + \frac{a}{\mu_1}}, & \mu > a. \end{cases}$$

Proposition 17. Fix $\mu_1 > 0$. Under linear staffing (1.14),

$$\lim_{\lambda \rightarrow \infty} \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} = \frac{\frac{a}{\mu_1^2} \left(1 - \frac{a}{\mu}\right)^+}{\left(1 - \frac{a}{\mu} + \frac{a}{\mu_1}\right)^2} = \begin{cases} 0, & \mu \leq a, \\ \frac{\frac{a}{\mu_1^2} \left(1 - \frac{a}{\mu}\right)}{\left(1 - \frac{a}{\mu} + \frac{a}{\mu_1}\right)^2}, & \mu > a. \end{cases}$$

For ease of presentation, we denote $I^\lambda(\mu_1, \mu)$ and $\frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1}$ simply by I^λ and $\frac{\partial I^\lambda}{\partial \mu_1}$, respectively; and let $d_1^\lambda := N^\lambda - \left(1 - \frac{\mu_1}{\mu}\right)$, $d_2^\lambda := d_1^\lambda - \rho^\lambda = (N^\lambda - \rho^\lambda) - \left(1 - \frac{\mu_1}{\mu}\right)$, and $C^\lambda := ErlC(N^\lambda, \rho^\lambda)$, where $\rho^\lambda = \frac{\lambda}{\mu}$.

A.7.2.1 Proof of Proposition 16

From (1.1) in Lemma 2,

$$I^\lambda = \left(1 + \frac{\mu}{\mu_1} \left(\rho^\lambda \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} + \frac{\rho^\lambda C^\lambda}{d_2^\lambda} \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda}\right)\right)\right)^{-1}. \quad (\text{A.88})$$

Observe that $\frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \geq 0$ (from Lemma 19 (c)) and $\frac{1}{d_2^\lambda} \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda}\right) \geq 0$ (recalling the definition $d_2^\lambda := d_1^\lambda - \rho^\lambda$) for all λ , resulting in all the summands in (A.88) being non-

negative for all λ . Therefore, I^λ would vanish in the limit as $\lambda \rightarrow \infty$ even if one of them grows unboundedly with λ .

Case (I): If $\mu < a$, then $\lim_{\lambda \rightarrow \infty} \frac{\rho^\lambda}{N^\lambda - \rho^\lambda} = (\frac{\mu}{a} - 1)^{-1} \in (-\infty, 0)$ and $\lim_{\lambda \rightarrow \infty} C^\lambda = \infty$ (from Lemma 29 (b)). Using these facts, (A.88) implies that $\lim_{\lambda \rightarrow \infty} I^\lambda = 0$.

Case (II): If $\mu > a$, then $\lim_{\lambda \rightarrow \infty} \frac{\rho^\lambda}{N^\lambda - \rho^\lambda} = (\frac{\mu}{a} - 1)^{-1}$ and $\lim_{\lambda \rightarrow \infty} C^\lambda = 0$ (from Lemma 29 (b)). Moreover, $\lim_{\lambda \rightarrow \infty} \frac{\rho^\lambda}{d_2^\lambda} = (\frac{\mu}{a} - 1)^{-1} \in (0, \infty)$ and $\left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda} \in (0, 1]$ for all large enough λ . Using these facts, (A.88) implies that

$$\lim_{\lambda \rightarrow \infty} I^\lambda = \left(1 + \frac{\mu}{\mu_1} \left(\frac{\mu}{a} - 1\right)^{-1}\right)^{-1} = \frac{1 - \frac{a}{\mu}}{1 - \frac{a}{\mu} + \frac{a}{\mu_1}}.$$

Case (III): If $\mu = a$, then $\lim_{\lambda \rightarrow \infty} \rho^\lambda \frac{1-C^\lambda}{N^\lambda - \rho^\lambda} = \infty$ (from Lemma 30 (d), recalling that when $\mu = a$, $\rho^\lambda = \frac{\lambda}{a}$). Thus, (A.88) implies that $\lim_{\lambda \rightarrow \infty} I^\lambda = 0$.

■

The proof of Proposition 17 will rely on the asymptotic behavior of k^λ in relation to N^λ . This requires the following setup. Consider a subsequence λ' on which $b := \lim_{\lambda' \rightarrow \infty} d_2^{\lambda'} \frac{k^{\lambda'} - N^{\lambda'}}{d_1^{\lambda'}} \in \mathbb{R} \cup \{-\infty, \infty\}$. If $b = 0$, then consider a further subsequence λ'' on which $d_2^{\lambda''} \frac{k^{\lambda''} - N^{\lambda''}}{d_1^{\lambda''}} > 0$ for all large enough λ'' or $d_2^{\lambda''} \frac{k^{\lambda''} - N^{\lambda''}}{d_1^{\lambda''}} < 0$ for all large enough λ'' . Simply using λ rather than λ' or λ'' to denote the subsequence, it is sufficient to consider four cases depending on the asymptotic behavior of $d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}$: (i) $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in \{-\infty\} \cup (-\infty, 0)$ and $d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in \{-\infty\} \cup (-\infty, 0)$ for all large enough λ ; (ii) $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in (0, \infty) \cup \{\infty\}$ and $d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in (0, \infty) \cup \{\infty\}$ for all large enough λ ; (iii) $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} = 0$ and $d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in (-\infty, 0)$ for all large enough λ ; and (iv) $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} = 0$ and $d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in (0, \infty)$ for all large enough λ . For identical reasons, when $\mu = a$, it is sufficient to consider four cases depending on the asymptotic behavior of d_2^λ : (i) $\lim_{\lambda \rightarrow \infty} d_2^\lambda \in \{-\infty\} \cup (-\infty, 0)$ and $d_2^\lambda \in \{-\infty\} \cup (-\infty, 0)$ for all large enough λ ; (ii) $\lim_{\lambda \rightarrow \infty} d_2^\lambda \in (0, \infty) \cup \{\infty\}$ and $d_2^\lambda \in (0, \infty) \cup \{\infty\}$ for all large enough λ ; (iii) $\lim_{\lambda \rightarrow \infty} d_2^\lambda = 0$ and $d_2^\lambda \in (-\infty, 0)$ for all

large enough λ ; and (iv) $\lim_{\lambda \rightarrow \infty} d_2^\lambda = 0$ and $d_2^\lambda \in (0, \infty)$ for all large enough λ .

To prove Proposition 17, we need the following auxiliary lemmas, whose proofs will appear at the end. These lemmas will also be used later, in the proof of Lemma 6.

Lemma 33. *Under linear staffing (1.14), if $\mu = a$, then*

$$(i) \lim_{\lambda \rightarrow \infty} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} = \exp \left(- \lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \right).$$

$$\text{Furthermore, if } \lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} = 0, \text{ then (ii) } \lim_{\lambda \rightarrow \infty} \frac{1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda}}{d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} = 1.$$

Lemma 34. *Under linear staffing (1.14), if $\mu = a$ and $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in \mathbb{R} \cup \{-\infty\}$, then the following holds: $\lim_{\lambda \rightarrow \infty} \left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^r I^\lambda = 0$ (i) for $r = 1$ if $\lim_{\lambda \rightarrow \infty} d_2^\lambda = 0$, or, (ii) for all $r \in \mathbb{N}$ if $\lim_{\lambda \rightarrow \infty} d_2^\lambda \in (\mathbb{R} \setminus \{0\}) \cup \{-\infty, \infty\}$.*

Lemma 35. *Under linear staffing (1.14), if $\mu = a$ and $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in (\mathbb{R} \setminus \{0\}) \cup \{-\infty, \infty\}$, then $\lim_{\lambda \rightarrow \infty} \frac{\rho^\lambda C^\lambda}{d_2^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} I^\lambda \in (-\infty, \infty)$.*

Lemma 36. *Under linear staffing (1.14), if $\mu = a$ and either (i) $0 < N^\lambda - \frac{\lambda}{a} \in \omega(1)$ or (ii) $0 < \frac{\lambda}{a} - N^\lambda \in \mathcal{O}(\sqrt{\lambda}) \cap \omega(1)$, then $\lim_{\lambda \rightarrow \infty} \frac{\sqrt{\rho^\lambda} I^\lambda}{d_2^\lambda} = 0$.*

Lemma 37. *Under linear staffing (1.14), if $\mu = a$ and $\left| N^\lambda - \frac{\lambda}{a} \right| \in \mathcal{O}(1)$, then the following holds: $\lim_{\lambda \rightarrow \infty} (\rho^\lambda)^r I^\lambda = 0$ (i) for $r \in [0, \frac{1}{2})$ if $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} = 0$, or (ii) for $r \in [0, 1)$ if $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in (\mathbb{R} \setminus \{0\}) \cup \{-\infty, \infty\}$.*

Lemma 38. *Under linear staffing (1.14), if $\mu = a$ and $0 < \frac{\lambda}{a} - N^\lambda \in \omega(\sqrt{\lambda})$, then $\lim_{\lambda \rightarrow \infty} \frac{\rho^\lambda C^\lambda I^\lambda}{d_2^\lambda} \in [-\frac{\mu_1}{a}, 0]$.*

Lemma 39. *Under linear staffing (1.14), if $\mu = a$, $\lim_{\lambda \rightarrow \infty} d_2^\lambda \in (\mathbb{R} \setminus \{0\}) \cup \{-\infty, \infty\}$, and $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} = 0$, then $\lim_{\lambda \rightarrow \infty} \frac{\rho^\lambda C^\lambda}{d_2^\lambda} \left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^r \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} (I^\lambda)^2 = 0$ for all $r \in \mathbb{N}$.*

A.7.2.2 Proof of Proposition 17

From (A.16) in Lemma 22,

$$\begin{aligned}
& \mu_1 \frac{\partial I^\lambda}{\partial \mu_1} \\
&= I^\lambda(1 - I^\lambda) + (I^\lambda)^2 \left(\frac{C^\lambda}{d_2^\lambda} \left[\left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right) \frac{\rho^\lambda}{d_2^\lambda} - (k^\lambda - N^\lambda) \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right] + \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} C^\lambda \right) \\
&\stackrel{(*)}{=} \left[\left(1 + \frac{1}{d_2^\lambda} \frac{\mu_1}{\mu} \right) I^\lambda(1 - I^\lambda) \right] - \left[\frac{\rho^\lambda}{d_2^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} (I^\lambda)^2 \right] - \left[\frac{\rho^\lambda C^\lambda}{d_2^\lambda} \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} (I^\lambda)^2 \right], \tag{A.89}
\end{aligned}$$

where $(*)$ follows by recalling the definition $d_2^\lambda := d_1^\lambda - \rho^\lambda$ and noting that

$$\begin{aligned}
(I^\lambda)^2 \frac{C^\lambda}{d_2^\lambda} \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right) \frac{\rho^\lambda}{d_2^\lambda} &= \frac{1}{d_2^\lambda} \frac{\mu_1}{\mu} I^\lambda(1 - I^\lambda) - \frac{\rho^\lambda}{d_2^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} (I^\lambda)^2, \quad \text{and then} \\
-\frac{C^\lambda}{d_2^\lambda} (k^\lambda - N^\lambda) \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} + \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} C^\lambda &= -\frac{\rho^\lambda C^\lambda}{d_2^\lambda} \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda}.
\end{aligned}$$

Case (I): If $\mu < a$, then $\lim_{\lambda \rightarrow \infty} \frac{1}{d_2^\lambda} = 0$ (recalling the definition $d_2^\lambda := N^\lambda - \rho^\lambda - 1 + \frac{\mu_1}{\mu}$, where $N^\lambda = \frac{\lambda}{a} + o(\lambda)$ under linear staffing (1.14) and $\rho^\lambda := \frac{\lambda}{\mu}$), $\lim_{\lambda \rightarrow \infty} I^\lambda = 0$ (from Proposition 16), and $\lim_{\lambda \rightarrow \infty} \frac{\rho^\lambda}{d_2^\lambda} = (\frac{\mu}{a} - 1)^{-1} \in (-\infty, 0)$. Furthermore, $\frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \in [0, 1]$ for all λ (from Lemma 19 (c)). Using these facts, the first two terms of (A.89) vanish in the limit as $\lambda \rightarrow \infty$. It remains to be shown that the third term follows suit:

$$\begin{aligned}
\lim_{\lambda \rightarrow \infty} \frac{\rho^\lambda C^\lambda}{d_2^\lambda} \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} (I^\lambda)^2 &= \lim_{\lambda \rightarrow \infty} \frac{\frac{\rho^\lambda C^\lambda}{d_2^\lambda} \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda}}{\left(1 + \rho^\lambda \frac{\mu}{\mu_1} \left(\frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} + \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right) \frac{C^\lambda}{d_2^\lambda} \right) \right)^2} \\
&= \lim_{\lambda \rightarrow \infty} \frac{\frac{C^\lambda}{d_2^\lambda} \frac{\rho^\lambda}{d_1^\lambda} (k^\lambda - N^\lambda) \left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda}}{\left(\left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} + \rho^\lambda \frac{\mu}{\mu_1} \left(\frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} + \left(\left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} - 1 \right) \frac{C^\lambda}{d_2^\lambda} \right) \right)^2} = 0,
\end{aligned}$$

because, in addition to the earlier facts, $\lim_{\lambda \rightarrow \infty} \frac{C^\lambda}{d_2^\lambda} = \lim_{\lambda \rightarrow \infty} \frac{C^\lambda/\lambda}{d_2^\lambda/\lambda} = \frac{(a-\mu)^2}{a^2\mu} / \left(\frac{1}{a} - \frac{1}{\mu} \right) = \frac{\mu}{a} - 1 \in (-\infty, 0)$ (using Lemma 30 (a)), $\lim_{\lambda \rightarrow \infty} \frac{d_1^\lambda}{\rho^\lambda} = \frac{\mu}{a} < 1$ (recalling the definition $d_1^\lambda := N^\lambda - 1 + \frac{\mu_1}{\mu}$, where $N^\lambda = \frac{\lambda}{a} + o(\lambda)$ under linear staffing (1.14) and $\rho^\lambda := \frac{\lambda}{\mu}$),

$\left(\frac{d_1^\lambda}{\rho^\lambda}\right)^{k^\lambda - N^\lambda} \leq 1$ for all large enough λ , and $\lim_{\lambda \rightarrow \infty} (k^\lambda - N^\lambda) \left(\frac{d_1^\lambda}{\rho^\lambda}\right)^{k^\lambda - N^\lambda} < \infty$ (because, even if $\lim_{\lambda \rightarrow \infty} k^\lambda - N^\lambda = \infty$, exponential decay in terms of $k^\lambda - N^\lambda$ would dominate its linear growth).

Case (II): If $\mu > a$, then the second and third terms vanish in the limit as $\lambda \rightarrow \infty$, because $\lim_{\lambda \rightarrow \infty} \frac{1}{d_2^\lambda} = 0$ (recalling the definition $d_2^\lambda := N^\lambda - \rho^\lambda - 1 + \frac{\mu_1}{\mu}$ and the linear staffing rule (1.14)), $\lim_{\lambda \rightarrow \infty} \frac{\rho^\lambda}{N^\lambda - \rho^\lambda} = (\frac{\mu}{a} - 1)^{-1} \in (0, \infty)$, $\lim_{\lambda \rightarrow \infty} C^\lambda = 0$ (from Lemma 29 (b)), $I^\lambda \in [0, 1]$ for all λ , $\lim_{\lambda \rightarrow \infty} \frac{1}{d_1^\lambda} = 0$ (recalling the definition $d_1^\lambda := N^\lambda - 1 + \frac{\mu_1}{\mu}$), and $\lim_{\lambda \rightarrow \infty} (k^\lambda - N^\lambda) \left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda} < \infty$ (noting that $\lim_{\lambda \rightarrow \infty} \frac{\rho^\lambda}{d_1^\lambda} = \frac{a}{\mu} < 1 \Rightarrow \frac{\rho^\lambda}{d_1^\lambda} < 1$ for all large enough λ , and, even if $\lim_{\lambda \rightarrow \infty} k^\lambda - N^\lambda = \infty$, exponential decay in terms of $k^\lambda - N^\lambda$ would dominate its linear growth). Thus, (A.89) yields

$$\lim_{\lambda \rightarrow \infty} \frac{\partial I^\lambda}{\partial \mu_1} = \frac{1}{\mu_1} \lim_{\lambda \rightarrow \infty} \left[\left(1 + \frac{1}{d_2^\lambda} \frac{\mu_1}{\mu} \right) I^\lambda \left(1 - I^\lambda \right) \right] = \frac{\frac{a}{\mu_1^2} \left(1 - \frac{a}{\mu} \right)}{\left(1 - \frac{a}{\mu} + \frac{a}{\mu_1} \right)^2},$$

recalling that $\lim_{\lambda \rightarrow \infty} \frac{1}{d_2^\lambda} = 0$ and $\lim_{\lambda \rightarrow \infty} I^\lambda = (1 - \frac{a}{\mu}) / (1 - \frac{a}{\mu} + \frac{a}{\mu_1})$ (from Proposition 16).

Case (III): If $\mu = a$, then, recalling that $N^\lambda = \frac{\lambda}{a} + o(\lambda)$ under linear staffing (1.14), it turns out that $|d_2^\lambda| = |(N^\lambda - \frac{\lambda}{a}) - (1 - \frac{\mu_1}{a})| \in o(\lambda)$, unlike when $\mu \neq a$, leading to the possibility that $\lim_{\lambda \rightarrow \infty} d_2^\lambda$ could be finite. This complicates the analysis. To proceed, we need the following auxiliary claim, whose proof is delayed until the end.

Claim 5. Under linear staffing (1.14), if $\lim_{\lambda \rightarrow \infty} d_2^\lambda = 0$ (implying that $\mu = a$) and $\lim_{\lambda \rightarrow \infty} d_2^\lambda \left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right) \in (\mathbb{R} \setminus \{0\}) \in \{-\infty, \infty\}$, then $\lim_{\lambda \rightarrow \infty} \frac{I^\lambda}{d_2^\lambda} = 0$.

We discuss three cases depending on the asymptotic behavior of $N^\lambda - \frac{\lambda}{a}$.

Case (A): If $0 < N^\lambda - \frac{\lambda}{a} \in \omega(1) \cap o(\lambda)$, then $\lim_{\lambda \rightarrow \infty} d_2^\lambda = \infty$, which implies that $d_2^\lambda > 0$ for all large enough λ . Furthermore, $\frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \geq 0$ for all λ (from Lemma 19 (c)). Therefore,

the second and third terms of (A.89) are non-negative, which implies that

$$\lim_{\lambda \rightarrow \infty} \frac{\partial I^\lambda}{\partial \mu_1} \leq \lim_{\lambda \rightarrow \infty} \frac{1}{\mu_1} \left[\left(1 + \frac{1}{d_2^\lambda} \frac{\mu_1}{a} \right) I^\lambda (1 - I^\lambda) \right] = 0,$$

recalling that $\lim_{\lambda \rightarrow \infty} d_2^\lambda \neq 0$ and $\lim_{\lambda \rightarrow \infty} I^\lambda = 0$ (from Proposition 16). Hence, $\lim_{\lambda \rightarrow \infty} \frac{\partial I^\lambda}{\partial \mu_1} = 0$ by non-negativity (recalling from Lemma 21 (a) that $\lim_{\lambda \rightarrow \infty} \frac{\partial I^\lambda}{\partial \mu_1} \geq 0$).

Case (B): If $0 < \frac{\lambda}{a} - N^\lambda \in \omega(1) \cap o(\lambda)$, then $\lim_{\lambda \rightarrow \infty} d_2^\lambda = -\infty$, which implies that $d_2^\lambda < 0$ for all large enough λ . We investigate the three terms of (A.89) separately and show that each of them converges to 0 as $\lambda \rightarrow \infty$.

- The first term converges to 0 as $\lambda \rightarrow \infty$, recalling that $\lim_{\lambda \rightarrow \infty} d_2^\lambda \neq 0$ and $\lim_{\lambda \rightarrow \infty} I^\lambda = 0$ (from Proposition 16).
- For the second term, we further discuss two cases.

- **Case (B-1):** If $0 < \frac{\lambda}{a} - N^\lambda \in \omega(1) \cap \mathcal{O}(\sqrt{\lambda})$, then, by regrouping the terms, we can write

$$\lim_{\lambda \rightarrow \infty} \frac{\rho^\lambda}{d_2^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} (I^\lambda)^2 = \lim_{\lambda \rightarrow \infty} \left(\frac{\sqrt{\rho^\lambda} I^\lambda}{d_2^\lambda} \right) \left(\sqrt{\rho^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \right) I^\lambda = 0,$$

because $\lim_{\lambda \rightarrow \infty} \frac{\sqrt{\rho^\lambda} I^\lambda}{d_2^\lambda} = 0$ (from Lemma 36), $\lim_{\lambda \rightarrow \infty} \sqrt{\rho^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda}$ is finite (from Lemma 30 (c), noting that $\rho^\lambda = \frac{\lambda}{a}$), and $I^\lambda \in [0, 1]$ for all λ .

- **Case (B-2):** If $0 < \frac{\lambda}{a} - N^\lambda \in \omega(\sqrt{\lambda}) \cap o(\lambda)$, then, by regrouping the terms, we can write

$$\lim_{\lambda \rightarrow \infty} \frac{\rho^\lambda}{d_2^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} (I^\lambda)^2 = \lim_{\lambda \rightarrow \infty} \left(\frac{\rho^\lambda C^\lambda I^\lambda}{d_2^\lambda} \right) \frac{1}{C^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} I^\lambda = 0,$$

because $\lim_{\lambda \rightarrow \infty} \frac{\rho^\lambda C^\lambda I^\lambda}{d_2^\lambda}$ is finite (from Lemma 38), $\lim_{\lambda \rightarrow \infty} C^\lambda = \infty$ (from Lemma 29 (b)); and $\frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \in [0, 1]$ (from Lemma 19 (c)) and $I^\lambda \in [0, 1]$ for all λ .

- For the third term, we further discuss two cases depending on the value of $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}$.
 - If $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} = 0$, then the third term converges to 0 as $\lambda \rightarrow \infty$ by Lemma 39.
 - If $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in \{-\infty\} \cup (-\infty, 0)$, then the third term converges to 0 as $\lambda \rightarrow \infty$ by Lemmas 34 and 35.

Case (C): If $|N^\lambda - \frac{\lambda}{a}| \in \mathcal{O}(1)$, we further discuss three cases depending on the value of $\lim_{\lambda \rightarrow \infty} d_2^\lambda$.

- **Case (C-1):** If $\lim_{\lambda \rightarrow \infty} d_2^\lambda \in (0, \infty) \cup \{\infty\}$, the arguments are identical to those in Case (A).
- **Case (C-2):** If $\lim_{\lambda \rightarrow \infty} d_2^\lambda \in \{-\infty\} \cup (-\infty, 0)$, we investigate the three terms of (A.89) separately and show that each of them converges to 0 as $\lambda \rightarrow \infty$.
 - The first term converges to 0 as $\lambda \rightarrow \infty$, recalling that $\lim_{\lambda \rightarrow \infty} d_2^\lambda \neq 0$ and $\lim_{\lambda \rightarrow \infty} I^\lambda = 0$.
 - The second term satisfies

$$\lim_{\lambda \rightarrow \infty} \frac{\rho^\lambda}{d_2^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} (I^\lambda)^2 = \lim_{\lambda \rightarrow \infty} \left((\rho^\lambda)^{\frac{1}{4}} I^\lambda \right)^2 \left(\sqrt{\rho^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \right) \frac{1}{d_2^\lambda} = 0,$$

by applying Lemma 37 and Lemma 30 (c), and noting that $\lim_{\lambda \rightarrow \infty} d_2^\lambda \neq 0$.

- For the third term, the arguments are identical to those in Case (B).
- **Case (C-3):** If $\lim_{\lambda \rightarrow \infty} d_2^\lambda = 0$, then we further discuss three cases depending on the value of $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}$.
 - If $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in (0, \infty) \cup \{\infty\}$, then

$$\lim_{\lambda \rightarrow \infty} \frac{\partial I^\lambda}{\partial \mu_1} \leq \lim_{\lambda \rightarrow \infty} \frac{1}{\mu_1} \left\{ \left(1 + \frac{1}{d_2} \frac{\mu_1}{\mu} \right) I^\lambda (1 - I^\lambda) \right\} = \lim_{\lambda \rightarrow \infty} \frac{1}{\mu_1} I^\lambda (1 - I^\lambda) + \frac{1}{\mu} (1 - I^\lambda) \frac{I^\lambda}{d_2^\lambda} = 0,$$

where the inequality follows for the same reasons as those in Case (A) and the last equality follows from $\lim_{\lambda \rightarrow \infty} I^\lambda = 0$ and Claim 5. Hence, $\lim_{\lambda \rightarrow \infty} \frac{\partial I^\lambda}{\partial \mu_1} = 0$ by non-negativity (recalling from Lemma 21 (a) that $\lim_{\lambda \rightarrow \infty} \frac{\partial I^\lambda}{\partial \mu_1} \geq 0$).

- If $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in \{-\infty\} \cup (-\infty, 0)$, we investigate the three terms of (A.89) separately and show that each of them converges to 0 as $\lambda \rightarrow \infty$.

★ The first term satisfies

$$\lim_{\lambda \rightarrow \infty} \left(1 + \frac{1}{d_2^\lambda} \frac{\mu_1}{\mu} \right) I^\lambda (1 - I^\lambda) = \lim_{\lambda \rightarrow \infty} I^\lambda (1 - I^\lambda) + \frac{\mu_1}{\mu} (1 - I^\lambda) \frac{I^\lambda}{d_2^\lambda} = 0,$$

recalling that $\lim_{\lambda \rightarrow \infty} I^\lambda = 0$ and by Claim 5.

★ The second term satisfies

$$\lim_{\lambda \rightarrow \infty} \frac{\rho^\lambda}{d_2^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} (I^\lambda)^2 = \lim_{\lambda \rightarrow \infty} \left(\sqrt{\rho^\lambda} I^\lambda \right) \left(\sqrt{\rho^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \right) \frac{I^\lambda}{d_2^\lambda} = 0,$$

by Lemma 37, Lemma 30 (c), and Claim 5.

★ The third term converges to 0 as $\lambda \rightarrow \infty$ by Lemmas 34 and 35.

- If $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} = 0$, then, from (A.89), when $\mu = a$,

$$\begin{aligned} \mu_1 \frac{\partial I^\lambda}{\partial \mu_1} &= \left(1 + \frac{1}{d_2^\lambda} \frac{\mu_1}{\mu} \right) I^\lambda (1 - I^\lambda) - \frac{\rho^\lambda}{d_2^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} (I^\lambda)^2 - \frac{\rho^\lambda C^\lambda}{d_2^\lambda} \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} (I^\lambda)^2 \\ &= I^\lambda (1 - I^\lambda) + \frac{(I^\lambda)^2}{d_2^\lambda} \left\{ \frac{\mu_1}{a} \left(\frac{1}{I^\lambda} - 1 \right) - \rho^\lambda \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} - \rho^\lambda C^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right\} \\ &\stackrel{(*)}{=} I^\lambda (1 - I^\lambda) + \frac{(I^\lambda)^2}{d_2^\lambda} \left\{ \frac{\mu_1}{a} \left[\rho^\lambda \frac{a}{\mu_1} \left(\frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} + \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right) \frac{C^\lambda}{d_2^\lambda} \right) \right] \right. \\ &\quad \left. - \rho^\lambda \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} - \rho^\lambda C^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right\} \\ &= I^\lambda (1 - I^\lambda) + \frac{(I^\lambda)^2}{d_2^\lambda} \frac{\rho^\lambda C^\lambda}{d_2^\lambda} \left[1 - \left(1 + d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \right) \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right], \end{aligned} \tag{A.90}$$

where (*) follows from (A.88).

Note that we can write

$$\left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda} = \left(1 - \frac{d_2^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda} = \exp\left((k^\lambda - N^\lambda) \ln\left(1 - \frac{d_2^\lambda}{d_1^\lambda}\right)\right).$$

Using the properties (i) $-\frac{x}{1-x} \leq \ln(1-x) \leq -x$ for all $x < 1$ and (ii) $\exp(x)$ is an increasing function of x for all $x \in \mathbb{R}$, we then obtain

$$\begin{aligned} \exp\left(-d_2^\lambda \frac{k^\lambda - N^\lambda}{\rho^\lambda}\right) &\leq \left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda} \leq \exp\left(-d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right) \\ \stackrel{(\dagger)}{\Rightarrow} 1 - d_2^\lambda \frac{k^\lambda - N^\lambda}{\rho^\lambda} + \frac{e^{\alpha_1^\lambda}}{2!} \left(-d_2^\lambda \frac{k^\lambda - N^\lambda}{\rho^\lambda}\right)^2 &\leq \left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda} \leq 1 - d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} + \frac{e^{\alpha_2^\lambda}}{2!} \left(-d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right)^2 \\ \Rightarrow 1 - \left(\frac{d_1^\lambda}{\rho^\lambda} - 1\right) \left(d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right) - \frac{d_1^\lambda}{\rho^\lambda} \left(d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right)^2 + \frac{e^{\alpha_1^\lambda}}{2!} \left(\frac{d_1^\lambda}{\rho^\lambda}\right)^2 \left(d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right)^2 \left(1 + d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right) &\leq \left(1 + d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right) \left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda} \leq 1 - \left(d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right)^2 + \frac{e^{\alpha_2^\lambda}}{2!} \left(d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right)^2 \left(1 + d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right) \\ \Rightarrow \left(d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right)^2 \left[1 - \frac{e^{\alpha_1^\lambda}}{2!} \left(1 + d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right)\right] &\leq 1 - \left(1 + d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right) \left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda} \\ &\leq (d_2^\lambda)^2 \frac{k^\lambda - N^\lambda}{d_1^\lambda \rho^\lambda} + \frac{d_1^\lambda}{\rho^\lambda} \left(d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right)^2 \left[1 - \frac{e^{\alpha_1^\lambda}}{2!} \frac{d_1^\lambda}{\rho^\lambda} \left(1 + d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right)\right] \\ \Rightarrow \frac{(I^\lambda)^2 \rho^\lambda C^\lambda}{d_2^\lambda} \left(d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right)^2 \left[1 - \frac{e^{\alpha_2^\lambda}}{2!} \left(1 + d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right)\right] &\leq \frac{(I^\lambda)^2 \rho^\lambda C^\lambda}{d_2^\lambda} \left[1 - \left(1 + d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right) \left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda}\right] \\ &\leq I^\lambda C^\lambda \left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} I^\lambda\right) + \frac{(I^\lambda)^2 \rho^\lambda C^\lambda}{d_2^\lambda} \frac{d_1^\lambda}{\rho^\lambda} \left(d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right)^2 \left[1 - \frac{e^{\alpha_1^\lambda}}{2!} \frac{d_1^\lambda}{\rho^\lambda} \left(1 + d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right)\right], \end{aligned} \tag{A.91}$$

where (\dagger) follows from Taylor's expansion for the function $\exp(x)$ at $x = 0$ up to the first two terms plus the remainder, according to which $|\alpha_1^\lambda| \leq \left|d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \frac{d_1^\lambda}{\rho^\lambda}\right|$ and $|\alpha_2^\lambda| \leq \left|d_2^\lambda \frac{k^\lambda - N^\lambda}{d_2^\lambda}\right|$. Since $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} = 0$ (by assumption), it follows that either $d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} > 0$ for all large enough λ or $d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} < 0$ for all large enough λ . We complete the proof under the former scenario; the proof under the latter is identical, except that the inequalities in the next step are reversed.

Note that $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} = 0$ (by assumption), $\lim_{\lambda \rightarrow \infty} I^\lambda = 0$ (from Propo-

sition 16 when $\mu = a$), $\lim_{\lambda \rightarrow \infty} C^\lambda = 1$ (from Lemma 29 (b), recalling that $|N^\lambda - \frac{\lambda}{a}| \in \mathcal{O}(1)$), $\lim_{\lambda \rightarrow \infty} \frac{k^\lambda - N^\lambda}{d_1^\lambda} = 0$ (from Lemma 33 (i)), $\lim_{\lambda \rightarrow \infty} \frac{d_1^\lambda}{\rho^\lambda} = 1$, $0 \leq \lim_{\lambda \rightarrow \infty} |\alpha_1^\lambda| \leq 0$ and $0 \leq |\alpha_2^\lambda| \leq 0$. Moreover, from (A.88) when $\mu = a$,

$$\begin{aligned} \lim_{\lambda \rightarrow \infty} \frac{(I^\lambda)^2 \rho^\lambda C^\lambda}{d_2^\lambda} \left(d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^2 &= \lim_{\lambda \rightarrow \infty} C^\lambda \left(\sqrt{\rho^\lambda} \frac{k^\lambda - N^\lambda}{d_1^\lambda} I^\lambda \right)^2 \\ &= \lim_{\lambda \rightarrow \infty} C^\lambda \left[\frac{d_1^\lambda}{\sqrt{\rho^\lambda}} \frac{1}{k^\lambda - N^\lambda} + \frac{\mu}{\mu_1} \sqrt{\rho^\lambda} \frac{d_1^\lambda}{k^\lambda - N^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} + \frac{a}{\mu_1} C^\lambda \frac{\sqrt{\rho^\lambda}}{d_2^\lambda} \frac{d_1^\lambda}{k^\lambda - N^\lambda} \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right) \right]^{-2} \end{aligned} \quad (\text{A.92})$$

$$= 0, \quad (\text{A.93})$$

because all three terms within the square bracket in the above display are non-negative, $\lim_{\lambda \rightarrow \infty} C^\lambda = 1$, and in particular, the third term satisfies

$$\frac{a}{\mu_1} \lim_{\lambda \rightarrow \infty} C^\lambda \sqrt{\rho^\lambda} \left(\frac{1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda}}{d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \right) = \infty,$$

recalling that $\lim_{\lambda \rightarrow \infty} C^\lambda = 1$ and $\lim_{\lambda \rightarrow \infty} \frac{1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda}}{d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} = 1$ (from Lemma 33 (ii)).

Based on the above facts, (A.91) implies that

$$\lim_{\lambda \rightarrow \infty} \frac{(I^\lambda)^2 \rho^\lambda C^\lambda}{d_2^\lambda} \left[1 - \left(1 + d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \right) \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right] = 0.$$

Hence, from (A.90), $\lim_{\lambda \rightarrow \infty} \frac{\partial I^\lambda}{\partial \mu_1} = 0$, recalling that $\lim_{\lambda \rightarrow \infty} I^\lambda = 0$.

■

Proof of Claim 5: We first note that $\lim_{\lambda \rightarrow \infty} d_2^\lambda = 0$ implies $\mu = a$ and $\lim_{\lambda \rightarrow \infty} N^\lambda - \frac{\lambda}{a} \in (-\infty, 1)$ (recalling $d_2^\lambda := N^\lambda - \frac{\lambda}{\mu} - (1 - \frac{\mu_1}{\mu})$), in turn, implying $|N^\lambda - \frac{\lambda}{a}| \in \mathcal{O}(1)$. We discuss the following two cases depending on the asymptotic behavior of $d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}$.

Case (I): If $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in (0, \infty) \cup \{\infty\}$. Then, from (A.88), when $\mu = a$,

$$\frac{\rho^\lambda I^\lambda}{d_2^\lambda} = \left[\frac{d_2^\lambda}{\rho^\lambda} + \frac{a}{\mu_1} \frac{d_2^\lambda}{\sqrt{\rho^\lambda}} \sqrt{\rho^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} + \frac{a}{\mu_1} C^\lambda \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right) \right]^{-1}.$$

Note that $\lim_{\lambda \rightarrow \infty} \frac{d_2^\lambda}{\rho^\lambda} = 0$, $\lim_{\lambda \rightarrow \infty} \frac{d_2^\lambda}{\sqrt{\rho^\lambda}} = 0$ (since $\lim_{\lambda \rightarrow \infty} d_2^\lambda = 0$), $\lim_{\lambda \rightarrow \infty} \sqrt{\rho^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \in (0, \infty)$ (from Lemma 30 (c) noting that $|N^\lambda - \frac{\lambda}{a}| \in \mathcal{O}(1)$), and $\lim_{\lambda \rightarrow \infty} C^\lambda = 1$ (from Lemma 29 (b)). Moreover, note that $\lim_{\lambda \rightarrow \infty} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} = e^{-\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \in [0, 1)$ (from Lemma 33 (i)). Thus, it follows that $\lim_{\lambda \rightarrow \infty} \frac{\rho^\lambda I^\lambda}{d_2^\lambda} \in [\frac{\mu_1}{a}, \infty)$. Then, it is clear that $\lim_{\lambda \rightarrow \infty} \frac{I^\lambda}{d_2^\lambda} = \lim_{\lambda \rightarrow \infty} \frac{1}{\rho^\lambda} \frac{\rho^\lambda I^\lambda}{d_2^\lambda} = 0$.

Case (II): If $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in \{-\infty\} \cup (-\infty, 0)$. Then, from (A.88), when $\mu = a$,

$$\frac{I^\lambda}{d_2^\lambda} = \left[d_2^\lambda + \frac{a}{\mu_1} d_2^\lambda \rho^\lambda \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} + \frac{a}{\mu_1} C^\lambda \rho^\lambda \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right) \right]^{-1}.$$

Note that the first two terms in the square bracket in the above display are negative and $C^\lambda \rightarrow 1$ (from Lemma 29 (b)) as $\lambda \rightarrow \infty$. Moreover, note that $\lim_{\lambda \rightarrow \infty} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} = e^{-\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \in (1, \infty]$ (from Lemma 33 (i)), which implies that $\lim_{\lambda \rightarrow \infty} \rho^\lambda \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right) = -\infty$. Then, it is clear that $\lim_{\lambda \rightarrow \infty} \frac{I^\lambda}{d_2^\lambda} = 0$.

Combining both cases, it follows that $\lim_{\lambda \rightarrow \infty} \frac{I^\lambda}{d_2^\lambda} = 0$. ■

We now proceed to prove auxiliary Lemmas 33-39.

Proof of Lemma 33

(i): First, we recall the definition $d_2^\lambda := d_1^\lambda - \rho^\lambda$, where $d_1^\lambda = N^\lambda - 1 + \frac{\mu_1}{\mu} = \frac{\lambda}{a} + o(\lambda) - 1 + \frac{\mu_1}{\mu}$ under linear staffing (1.14). This implies that $\lim_{\lambda \rightarrow \infty} \frac{d_2^\lambda}{d_1^\lambda} = 0$ (recalling that $\rho^\lambda = \frac{\lambda}{a}$ when

$\mu = a$). Next, we can write

$$\begin{aligned} \left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda} &= \left(1 - \frac{d_2^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda} = \left(\left(1 - \frac{d_2^\lambda}{d_1^\lambda}\right)^{-\frac{d_2^\lambda}{d_2^\lambda}}\right)^{-d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \\ &\Rightarrow \lim_{\lambda \rightarrow \infty} \left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda} = \left(\lim_{\lambda \rightarrow \infty} \left(1 - \frac{d_2^\lambda}{d_1^\lambda}\right)^{-\frac{d_2^\lambda}{d_2^\lambda}}\right)^{-\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} = \exp\left(-\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right). \end{aligned}$$

(ii): First, we recall the definition $d_2^\lambda := d_1^\lambda - \rho^\lambda$, where $d_1^\lambda := N^\lambda - 1 + \frac{\mu_1}{\mu} > 0$ (since $N^\lambda \geq 1$, $\mu_1 > 0$, and $\mu > 0$) and $\rho^\lambda := \frac{\lambda}{\mu} > 0$ for all $\lambda > 0$. This implies that $\frac{d_2^\lambda}{d_1^\lambda} < 1$ for all $\lambda > 0$. Next, we can write

$$\left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda} = \left(1 - \frac{d_2^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda} = \exp\left(\left(k^\lambda - N^\lambda\right) \ln\left(1 - \frac{d_2^\lambda}{d_1^\lambda}\right)\right).$$

Using the properties (i) $-\frac{x}{1-x} \leq \ln(1-x) \leq -x$ for all $x < 1$ and (ii) $\exp(x)$ is an increasing function of x for all $x \in \mathbb{R}$, we then obtain

$$\begin{aligned} \exp\left(-d_2^\lambda \frac{k^\lambda - N^\lambda}{\rho^\lambda}\right) &\leq \left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda} \leq \exp\left(-d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right) \\ \stackrel{(*)}{\Rightarrow} 1 - d_2^\lambda \frac{k^\lambda - N^\lambda}{\rho^\lambda} + \frac{e^{\alpha_1^\lambda}}{2!} \left(-d_2^\lambda \frac{k^\lambda - N^\lambda}{\rho^\lambda}\right)^2 &\leq \left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda} \leq 1 - d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} + \frac{e^{\alpha_2^\lambda}}{2!} \left(-d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right)^2 \\ \Rightarrow d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} - \frac{e^{\alpha_2^\lambda}}{2!} \left(-d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right)^2 &\leq 1 - \left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda} \leq d_2^\lambda \frac{k^\lambda - N^\lambda}{\rho^\lambda} - \frac{e^{\alpha_1^\lambda}}{2!} \left(-d_2^\lambda \frac{k^\lambda - N^\lambda}{\rho^\lambda}\right)^2 \\ \Rightarrow d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(1 - \frac{e^{\alpha_2^\lambda}}{2!} \left(d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right)\right) &\leq 1 - \left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda} \leq d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \frac{d_1^\lambda}{\rho^\lambda} \left(1 - \frac{e^{\alpha_1^\lambda}}{2!} \left(d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \frac{d_1^\lambda}{\rho^\lambda}\right)\right), \end{aligned} \tag{A.94}$$

where $(*)$ follows from Taylor's expansion for the function $\exp(x)$ at $x = 0$ up to the first two terms plus the remainder, according to which $|\alpha_1^\lambda| \leq \left|d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \frac{d_1^\lambda}{\rho^\lambda}\right|$ and $|\alpha_2^\lambda| \leq \left|d_2^\lambda \frac{k^\lambda - N^\lambda}{d_2^\lambda}\right|$. Since $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} = 0$ (by assumption), it follows that either $d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} > 0$ for all large enough λ or $d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} < 0$ for all large enough λ . We complete the proof under the former scenario; the proof under the latter is identical, except that the inequalities in the

next step are reversed. Dividing (A.94) throughout by $d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}$, we obtain, for all large enough λ ,

$$1 - \frac{e^{\alpha_2^\lambda}}{2!} \left(d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \right) \leq \frac{1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda}}{d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \leq \frac{d_1^\lambda}{\rho^\lambda} \left(1 - \frac{d_1^\lambda}{\rho^\lambda} \frac{e^{\alpha_1^\lambda}}{2!} \left(d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \right) \right). \quad (\text{A.95})$$

Recalling that $d_1^\lambda = N^\lambda - 1 + \frac{\mu_1}{\mu} = \frac{\lambda}{a} + o(\lambda) - 1 + \frac{\mu_1}{\mu}$ under linear staffing (1.14) and that $\rho^\lambda = \frac{\lambda}{a}$ when $\mu = a$, it follows that $\lim_{\lambda \rightarrow \infty} \frac{d_1^\lambda}{\rho^\lambda} = 1$. In addition, recalling that $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} = 0$ (by assumption), it follows that $\lim_{\lambda \rightarrow \infty} e^{\alpha_1^\lambda} = \lim_{\lambda \rightarrow \infty} e^{\alpha_2^\lambda} = 1$. Using these facts, (A.95) implies that

$$1 \leq \lim_{\lambda \rightarrow \infty} \frac{1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda}}{d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \leq 1 \quad \Rightarrow \quad \lim_{\lambda \rightarrow \infty} \frac{1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda}}{d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} = 1.$$

■

Proof of Lemma 34

By assumption, we know that $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in \mathbb{R} \cup \{-\infty\}$. We discuss three cases depending on the values of $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}$ and $\lim_{\lambda \rightarrow \infty} d_2^\lambda$.

Case (I): If $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in \mathbb{R}$ and $\lim_{\lambda \rightarrow \infty} d_2^\lambda \in (\mathbb{R} \setminus \{0\}) \cup \{-\infty, \infty\}$, then, by multiplying and dividing by $(d_2^\lambda)^r$, we can write

$$\left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^r I^\lambda = \frac{1}{(d_2^\lambda)^r} \left(d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^r I^\lambda = 0 \quad \forall r \in \mathbb{N},$$

recalling that when $\mu = a$, $\lim_{\lambda \rightarrow \infty} I^\lambda = 0$ (from Proposition 16).

For the remaining two cases, we first use (A.88) with $\mu = a$ to expand

$$\left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^r I^\lambda = \left[\left(\frac{d_1^\lambda}{k^\lambda - N^\lambda} \right)^r + \rho^\lambda \frac{a}{\mu_1} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \left(\frac{d_1^\lambda}{k^\lambda - N^\lambda} \right)^r + \rho^\lambda \frac{a}{\mu_1} \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right) \frac{C^\lambda}{d_2^\lambda} \left(\frac{d_1^\lambda}{k^\lambda - N^\lambda} \right)^r \right]^{-1}. \quad (\text{A.96})$$

Next, observe that $\frac{1-C^\lambda}{N^\lambda-\rho^\lambda} \geq 0$ (from Lemma 19 (c)) and $\frac{1}{d_2^\lambda} \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda-N^\lambda}\right) \geq 0$ (recalling the definition $d_2^\lambda := d_1^\lambda - \rho^\lambda$) for all λ , resulting in all three terms within the square bracket of (A.96) being non-negative for all λ . Therefore, when $\lambda \rightarrow \infty$, in order to show that the expression in (A.96) vanishes, we need only show that one of these terms diverges to ∞ . For the remaining two cases, we focus on the third term, which can be multiplied and divided by $(d_2^\lambda)^r$ and expressed as

$$\frac{a}{\mu_1} \rho^\lambda C^\lambda (d_2^\lambda)^{r-1} \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda-N^\lambda}\right) \left(d_2^\lambda \frac{k^\lambda-N^\lambda}{d_1^\lambda}\right)^{-r}. \quad (\text{A.97})$$

Case (II): Suppose $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda-N^\lambda}{d_1^\lambda} \in \mathbb{R}$ and $\lim_{\lambda \rightarrow \infty} d_2^\lambda = 0$. First, when $r = 1$, (A.97) becomes

$$\frac{a}{\mu_1} \rho^\lambda C^\lambda \left(\frac{1 - \left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda-N^\lambda}}{d_2^\lambda \frac{k^\lambda-N^\lambda}{d_1^\lambda}} \right). \quad (\text{A.98})$$

Next, note that $\lim_{\lambda \rightarrow \infty} d_2^\lambda = 0$ implies that $\lim_{\lambda \rightarrow \infty} N^\lambda - \frac{\lambda}{a} = 1 - \frac{\mu_1}{a} \Rightarrow \left|N^\lambda - \frac{\lambda}{a}\right| \in \mathcal{O}(1)$ (recalling that when $\mu = a$, $d_2^\lambda = N^\lambda - \frac{\lambda}{a} - (1 - \frac{\mu_1}{a})$); therefore, $\lim_{\lambda \rightarrow \infty} C^\lambda = 1$ (from Lemma 29 (b)). Furthermore, $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda-N^\lambda}{d_1^\lambda} \in \mathbb{R}$ implies that $\lim_{\lambda \rightarrow \infty} \frac{1 - \left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda-N^\lambda}}{d_2^\lambda \frac{k^\lambda-N^\lambda}{d_1^\lambda}} > 0$ (from Lemma 33 (i)). Therefore, as $\lambda \rightarrow \infty$, the expression in (A.98) diverges to infinity because $\rho^\lambda \rightarrow \infty$, as desired.

Case (III): If $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda-N^\lambda}{d_1^\lambda} = -\infty$, then $d_2^\lambda \frac{k^\lambda-N^\lambda}{d_1^\lambda} < 0$ and $d_2^\lambda < 0$ for all large enough λ . The latter implies that either $0 < \frac{\lambda}{a} - N^\lambda \in \omega(1)$ or $\left|N^\lambda - \frac{\lambda}{a}\right| \in \mathcal{O}(1)$; therefore, $\lim_{\lambda \rightarrow \infty} C^\lambda \geq 1$ (from Lemma 29 (b)). Furthermore, $d_2^\lambda < 0$ for all large enough λ also implies that $\frac{d_2^\lambda}{d_1^\lambda} < 0$ and $\frac{\rho^\lambda}{d_1^\lambda} > 1$ for all large enough λ (recalling the definition $d_2^\lambda := d_1^\lambda - \rho^\lambda$).

Next, we can write

$$\left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda-N^\lambda} = \left(1 - \frac{d_2^\lambda}{d_1^\lambda}\right)^{k^\lambda-N^\lambda} = \exp\left(\left(k^\lambda - N^\lambda\right) \ln\left(1 - \frac{d_2^\lambda}{d_1^\lambda}\right)\right).$$

Using the properties (i) $\ln(1 - x) \geq -\frac{x}{1-x}$ for all $x < 1$ and (ii) $\exp(x)$ is an increasing function of x for all $x \in \mathbb{R}$, we then obtain

$$\begin{aligned} \left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda} &\geq \exp\left(-d_2^\lambda \frac{k^\lambda - N^\lambda}{\rho^\lambda}\right) \stackrel{(*)}{=} 1 + \sum_{i=1}^r \frac{1}{i!} \left(-d_2^\lambda \frac{k^\lambda - N^\lambda}{\rho^\lambda}\right)^i + \frac{e^{\alpha^\lambda}}{(r+1)!} \left(-d_2^\lambda \frac{k^\lambda - N^\lambda}{\rho^\lambda}\right)^{r+1} \\ \Rightarrow \left(\left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda} - 1\right) &\geq \sum_{i=1}^r \frac{1}{i!} \left(-d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right)^i \left(\frac{d_1^\lambda}{\rho^\lambda}\right)^i + \frac{e^{\alpha^\lambda}}{(r+1)!} \left(-d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right)^{r+1} \left(\frac{d_1^\lambda}{\rho^\lambda}\right)^{r+1} \\ &\stackrel{(**)}{\geq} \frac{1}{(r+1)!} \left(-d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right)^{r+1} \left(\frac{d_1^\lambda}{\rho^\lambda}\right)^{r+1}, \end{aligned} \quad (\text{A.99})$$

where (*) follows from Taylor's expansion for the function $\exp(x)$ at $x = 0$ up to the first $r+1$ terms plus the remainder, according to which $\alpha^\lambda \in \left(0, -d_2^\lambda \frac{k^\lambda - N^\lambda}{\rho^\lambda}\right)$; and (**) follows by recalling that $d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} < 0$ for all large enough λ and noting that $e^{\alpha^\lambda} \geq 1$ for all large enough λ .

Recalling that $\frac{\rho^\lambda}{d_1^\lambda} > 1$ and $d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} < 0$ for all large enough λ , it follows that both sides of the inequality (A.99) are strictly positive for all large enough λ . Using (A.99), we write

$$\begin{aligned} \left| \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda}\right) \left(d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right)^{-r} \right| &= \left| \left(\left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda} - 1\right) \left(-d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right)^{-r} \right| \\ &\geq \frac{1}{(r+1)!} \left(-d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right) \left(\frac{d_1^\lambda}{\rho^\lambda}\right)^{r+1}. \end{aligned} \quad (\text{A.100})$$

Recalling that $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} = -\infty$ (by assumption) and noting that $\lim_{\lambda \rightarrow \infty} \frac{d_1^\lambda}{\rho^\lambda} = 1$ (recalling that $d_1^\lambda = N^\lambda - 1 + \frac{\mu_1}{\mu} = \frac{\lambda}{a} + o(\lambda) - 1 + \frac{\mu_1}{\mu}$ under linear staffing (1.14) and that $\rho^\lambda = \frac{\lambda}{a}$ when $\mu = a$), the right-hand side of (A.100) diverges to ∞ in the limit as $\lambda \rightarrow \infty$, implying that

$$\lim_{\lambda \rightarrow \infty} \left| \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda - N^\lambda}\right) \left(d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}\right)^{-r} \right| = \infty. \quad (\text{A.101})$$

Recalling that $\lim_{\lambda \rightarrow \infty} C^\lambda \geq 1$ and using (A.101), the expression in (A.97) diverges to ∞ (i) for $r = 1$ if $\lim_{\lambda \rightarrow \infty} d_2^\lambda = 0$ and (ii) for all $r \in \mathbb{N}$ if $\lim_{\lambda \rightarrow \infty} d_2^\lambda \in (\mathbb{R} \setminus \{0\}) \cup \{-\infty, \infty\}$, as desired. ■

Proof of Lemma 35

From (A.88), when $\mu = a$,

$$\begin{aligned} \frac{\rho^\lambda C^\lambda}{d_2^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} I^\lambda &= \left[\frac{d_2^\lambda}{\rho^\lambda C^\lambda} \left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} + \frac{a}{\mu_1} \frac{d_2^\lambda}{C^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} + \frac{a}{\mu_1} \left(\left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} - 1 \right) \right]^{-1} \\ &= \left[\frac{d_2^\lambda}{\rho^\lambda C^\lambda} \left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} + \frac{a}{\mu_1} \left\{ \left(\frac{d_2^\lambda}{C^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} + 1 \right) \left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} - 1 \right\} \right]^{-1}. \end{aligned} \quad (\text{A.102})$$

We first evaluate the limit of the second term within the square bracket in (A.102). Note that $\frac{d_2^\lambda}{C^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} = \frac{d_2^\lambda}{N^\lambda - \rho^\lambda} \frac{1 - C^\lambda}{C^\lambda} \rightarrow \lim_{\lambda \rightarrow \infty} \frac{1}{C^\lambda} - 1$ and $\left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} \rightarrow e^{\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}}$, as $\lambda \rightarrow \infty$ (from Lemma 33 (i)). Then,

$$\lim_{\lambda \rightarrow \infty} \left(\frac{d_2^\lambda}{C^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} + 1 \right) \left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} - 1 = \left(\lim_{\lambda \rightarrow \infty} \frac{1}{C^\lambda} \right) e^{\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} - 1. \quad (\text{A.103})$$

It suffices to show:

- (i) $\left(\lim_{\lambda \rightarrow \infty} \frac{1}{C^\lambda} \right) e^{\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \neq 1$;
- (ii) $\lim_{\lambda \rightarrow \infty} \frac{d_2^\lambda}{\rho^\lambda C^\lambda} \left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} = 0$, or $\lim_{\lambda \rightarrow \infty} \frac{d_2^\lambda}{\rho^\lambda C^\lambda} \left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda}$ and $\left(\lim_{\lambda \rightarrow \infty} \frac{1}{C^\lambda} \right) e^{\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} - 1$ share the same sign.

Recall that, by assumption, $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in (\mathbb{R} \setminus \{0\}) \cup \{-\infty, \infty\}$. We discuss the following three cases depending on the asymptotic behavior of $N^\lambda - \frac{\lambda}{a}$. These are **Case (I)** $0 < \frac{\lambda}{a} - N^\lambda \in \omega(1) \cap o(\lambda)$, **Case (II)** $0 < \left| \frac{\lambda}{a} - N^\lambda \right| \in \mathcal{O}(1)$, and **Case (III)** $0 < N^\lambda - \frac{\lambda}{a} \in \omega(1) \cap o(\lambda)$.

Case (I): If $0 < \frac{\lambda}{a} - N^\lambda \in \omega(1) \cap o(\lambda)$, then $\lim_{\lambda \rightarrow \infty} C^\lambda \in [1, \infty) \cup \{\infty\}$ and $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in \{-\infty\} \cup (-\infty, 0)$. Then, $\lim_{\lambda \rightarrow \infty} \frac{1}{C^\lambda} \in [0, 1]$ and $e^{\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \in [0, 1)$, implying that $\left(\lim_{\lambda \rightarrow \infty} \frac{1}{C^\lambda} \right) e^{\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \in [0, 1)$. Moreover, $\lim_{\lambda \rightarrow \infty} \frac{d_2^\lambda}{\rho^\lambda C^\lambda} \left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} < 0$. There-

fore, both (i) and (ii) hold.

Case (II): If $0 < \left| \frac{\lambda}{a} - N^\lambda \right| \in \mathcal{O}(1)$, then $\lim_{\lambda \rightarrow \infty} C^\lambda = 1$ and $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in (\mathbb{R} \setminus \{0\}) \cup \{-\infty, \infty\}$. Then, $\lim_{\lambda \rightarrow \infty} \frac{1}{C^\lambda} = 1$ and $e^{\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \in [0, 1) \cup (1, \infty) \cup \{\infty\}$, implying that $\left(\lim_{\lambda \rightarrow \infty} \frac{1}{C^\lambda} \right) e^{\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \in [0, 1) \cup (1, \infty) \cup \{\infty\}$. Hence, (i) holds.

- If $\lim_{\lambda \rightarrow \infty} d_2^\lambda \in \{-\infty\} \cup (-\infty, 0]$, then $\lim_{\lambda \rightarrow \infty} \frac{d_2^\lambda}{\rho^\lambda C^\lambda} \left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} \leq 0$, and $\left(\lim_{\lambda \rightarrow \infty} \frac{1}{C^\lambda} \right) e^{\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \in [0, 1)$. Therefore, (ii) holds.
 - If $\lim_{\lambda \rightarrow \infty} d_2^\lambda \in (0, \infty) \cup \{\infty\}$, then $\lim_{\lambda \rightarrow \infty} \frac{d_2^\lambda}{\rho^\lambda C^\lambda} \left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} > 0$, and $\left(\lim_{\lambda \rightarrow \infty} \frac{1}{C^\lambda} \right) e^{\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \in (1, \infty) \cup \{\infty\}$. Therefore, (ii) holds.
- Case (III):** If $0 < N^\lambda - \frac{\lambda}{a} \in \omega(1) \cap o(\lambda)$, then $\lim_{\lambda \rightarrow \infty} C^\lambda \in [0, 1]$ and $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in (0, \infty) \cup \{\infty\}$. Then, $\lim_{\lambda \rightarrow \infty} \frac{1}{C^\lambda} \in [1, \infty) \cup \{\infty\}$ and $e^{\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \in (1, \infty) \cup \{\infty\}$, implying that $\left(\lim_{\lambda \rightarrow \infty} \frac{1}{C^\lambda} \right) e^{\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \in (1, \infty) \cup \{\infty\}$. Moreover, $\lim_{\lambda \rightarrow \infty} \frac{d_2^\lambda}{\rho^\lambda C^\lambda} \left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} > 0$. Therefore, both (i) and (ii) hold. ■

Proof of Lemma 36

From (A.88),

$$\frac{\sqrt{\rho^\lambda} I^\lambda}{d_2^\lambda} = \left[\frac{d_2^\lambda}{\sqrt{\rho^\lambda}} + \frac{\mu}{\mu_1} \sqrt{\rho^\lambda} d_2^\lambda \left(\frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \right) + \frac{\mu}{\mu_1} C^\lambda \sqrt{\rho^\lambda} \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right) \right]^{-1}. \quad (\text{A.104})$$

(i): If $0 < N^\lambda - \frac{\lambda}{a} \in \omega(1)$, then $d_2^\lambda > 0$ for all large enough λ . We rewrite (A.104) as

$$\frac{\sqrt{\rho^\lambda} I^\lambda}{d_2^\lambda} = \frac{1}{\sqrt{\rho^\lambda}} \left[\frac{d_2^\lambda}{\rho^\lambda} + \frac{a}{\mu_1} (1 - C^\lambda) \left(\frac{d_2^\lambda}{N^\lambda - \rho^\lambda} \right) + \frac{a}{\mu_1} C^\lambda \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right) \right]^{-1}.$$

Note that $\lim_{\lambda \rightarrow \infty} \frac{d_2^\lambda}{\rho^\lambda} = 0$, $\lim_{\lambda \rightarrow \infty} \frac{d_2^\lambda}{N^\lambda - \rho^\lambda} = 1$, $\lim_{\lambda \rightarrow \infty} C^\lambda \in [0, 1]$ (from Lemma 29 (b)),

and $\lim_{\lambda \rightarrow \infty} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} = e^{-\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \in [0, 1]$ (from Lemma 33 (i)). Thus, the terms within the square bracket in the above display are all finite, which implies that $\lim_{\lambda \rightarrow \infty} \frac{\sqrt{\rho^\lambda} I^\lambda}{d_2^\lambda} = 0$.

(ii): If $0 < \frac{\lambda}{a} - N^\lambda \in \mathcal{O}(\sqrt{\lambda}) \cap \omega(1)$, then $d_2^\lambda < 0$ for all large enough λ . Since $d_2^\lambda = d_1^\lambda - \rho^\lambda$, all three terms in (A.104) are non-positive for all large enough λ . Furthermore, note that $\lim_{\lambda \rightarrow \infty} \sqrt{\rho^\lambda} \left(\frac{1-C^\lambda}{N^\lambda - \rho^\lambda} \right) \in (0, \infty)$ (from Lemma 30 (c)), implying that the second term diverges to $-\infty$ as $\lambda \rightarrow \infty$ (since $|d_2^\lambda| \in \omega(1)$). Therefore, $\lim_{\lambda \rightarrow \infty} \frac{\sqrt{\rho^\lambda} I^\lambda}{d_2^\lambda} = 0$. \blacksquare

Proof of Lemma 37

From (A.88), when $\mu = a$,

$$(\rho^\lambda)^r I^\lambda = \left[\frac{1}{(\rho^\lambda)^r} + \frac{a}{\mu_1} (\rho^\lambda)^{1-r} \frac{1-C^\lambda}{N^\lambda - \rho^\lambda} + \frac{a}{\mu_1} (\rho^\lambda)^{1-r} \frac{C^\lambda}{d_2^\lambda} \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right) \right]^{-1}. \quad (\text{A.105})$$

Note that $\lim_{\lambda \rightarrow \infty} C^\lambda = 1$ (from Lemma 29 (b)) and $|d_2^\lambda| \in \mathcal{O}(1)$, implying that $\lim_{\lambda \rightarrow \infty} \frac{C^\lambda}{d_2^\lambda} \neq 0$. Note that $\frac{1-C^\lambda}{N^\lambda - \rho^\lambda} \in [0, 1]$ for all λ (from Lemma 19 (c)) and, since $d_2^\lambda = d_1^\lambda - \rho^\lambda$, $\frac{1}{d_2^\lambda} \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right) \geq 0$ for all λ . As a result, all three terms within (A.105) are non-negative for all λ .

(i): If $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} = 0$, then, note that

$$\lim_{\lambda \rightarrow \infty} (\rho^\lambda)^{1-r} \frac{1-C^\lambda}{N^\lambda - \rho^\lambda} = \lim_{\lambda \rightarrow \infty} \left(\sqrt{\rho^\lambda} \frac{1-C^\lambda}{N^\lambda - \rho^\lambda} \right) (\rho^\lambda)^{\frac{1}{2}-r} = \infty, \quad \forall r \in \left[0, \frac{1}{2} \right),$$

since $\sqrt{\rho^\lambda} \frac{1-C^\lambda}{N^\lambda - \rho^\lambda} \in (0, \infty)$ (from Lemma 30 (c)). Thus, the second term of (A.105) diverges to ∞ as $\lambda \rightarrow \infty$, which implies that $\lim_{\lambda \rightarrow \infty} (\rho^\lambda)^r I^\lambda = 0$ for all $r \in [0, \frac{1}{2})$.

(ii): If $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in (\mathbb{R} \setminus \{0\}) \cup \{-\infty, \infty\}$, then, note that

$$\lim_{\lambda \rightarrow \infty} (\rho^\lambda)^{1-r} \frac{C^\lambda}{d_2^\lambda} \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right) = \infty, \quad \forall r \in [0, 1),$$

recalling that $\lim_{\lambda \rightarrow \infty} \frac{C^\lambda}{d_2^\lambda} \neq 0$ and $\lim_{\lambda \rightarrow \infty} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} = e^{-\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \neq 1$ (where the equality follows from Lemma 33 (i)). Thus, the third term of (A.105) diverges to ∞ as $\lambda \rightarrow \infty$, which implies that $\lim_{\lambda \rightarrow \infty} (\rho^\lambda)^r I^\lambda = 0$ for all $r \in [0, 1)$. \blacksquare

Proof of Lemma 38

From (A.88), when $\mu = a$,

$$\begin{aligned} \frac{\rho^\lambda C^\lambda I^\lambda}{d_2^\lambda} &= \left[\frac{d_2^\lambda}{\rho^\lambda C^\lambda} \frac{1}{N^\lambda - \rho^\lambda} + \frac{a}{\mu_1} \frac{d_2^\lambda}{N^\lambda - \rho^\lambda} \left(\frac{1}{C^\lambda} - 1 \right) + \frac{a}{\mu_1} \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right) \right]^{-1} \\ &= \left[\frac{d_2^\lambda}{\rho^\lambda C^\lambda} \frac{1}{N^\lambda - \rho^\lambda} \left(\frac{1}{C^\lambda} - 1 \right) + 1 - \frac{a}{\mu_1} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right]^{-1}. \end{aligned}$$

Note that $\frac{d_2^\lambda}{\rho^\lambda} \rightarrow 0$, $C^\lambda \rightarrow \infty$ (from Lemma 29 (b) when $0 < \frac{\lambda}{a} - N^\lambda \in \omega(\sqrt{\lambda})$ (by assumption)), $\frac{d_2^\lambda}{N^\lambda - \rho^\lambda} \rightarrow 1$ (since $0 < \frac{\lambda}{a} - N^\lambda \in \omega(\sqrt{\lambda})$ (by assumption)), and $\left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \rightarrow e^{-\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \in [1, \infty]$, as $\lambda \rightarrow \infty$ (from Lemma 33 (i)). Hence, $\lim_{\lambda \rightarrow \infty} \frac{\rho^\lambda C^\lambda I^\lambda}{d_2^\lambda} \in [-\frac{\mu_1}{a}, 0]$. \blacksquare

Proof of Lemma 39

By multiplying and dividing by $(d_2^\lambda)^{r-1}$ and regrouping the terms, we can write

$$\frac{\rho^\lambda C^\lambda}{d_2^\lambda} \left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^r \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} (I^\lambda)^2 = \frac{\rho^\lambda C^\lambda}{d_2^\lambda} \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} (I^\lambda)^2 \left(\frac{d_2^\lambda}{d_1^\lambda} \frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^{r-1} \frac{1}{(d_2^\lambda)^{r-1}}.$$

Note that $\lim_{\lambda \rightarrow \infty} \left(d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^{r-1} \in \{0, 1\}$ (because $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} = 0$ by assumption and $r \in \mathbb{N}$), and $\lim_{\lambda \rightarrow \infty} \frac{1}{(d_2^\lambda)^{r-1}} \in (-\infty, \infty)$ (because $\lim_{\lambda \rightarrow \infty} d_2^\lambda \neq 0$ and $r \in \mathbb{N}$). Therefore, to complete the proof, it suffices to show $\lim_{\lambda \rightarrow \infty} \frac{\rho^\lambda C^\lambda}{d_2^\lambda} \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} (I^\lambda)^2 = 0$.

Using (A.88) and after algebra, when $\mu = a$,

$$\begin{aligned} & \left[\frac{\rho^\lambda C^\lambda}{d_2^\lambda} \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} (I^\lambda)^2 \right]^{-1} \\ &= \frac{(d_2^\lambda)^2}{\rho^\lambda C^\lambda} \left(\frac{\left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda}}{d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \right) \left(1 + \frac{a}{\mu_1} \rho^\lambda \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \right)^2 + \left(\frac{a}{\mu_1} \right)^2 \rho^\lambda C^\lambda \left(\frac{\left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda}}{d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \right) \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right)^2 \\ &+ 2 \frac{a}{\mu_1} d_2^\lambda \left(\frac{\left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda}}{d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \right) \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right) \left(1 + \frac{a}{\mu_1} \rho^\lambda \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \right). \end{aligned} \quad (\text{A.106})$$

Note that each of the three terms in the above display either has the same sign as $d_2^\lambda = d_1^\lambda - \rho^\lambda$ or is 0 for all λ . Furthermore, when examining the third term, we note that

$$\begin{aligned} & \lim_{\lambda \rightarrow \infty} \left| d_2^\lambda \left(\frac{\left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda}}{d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \right) \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right) \left(1 + \frac{a}{\mu_1} \rho^\lambda \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \right) \right| \\ &= \lim_{\lambda \rightarrow \infty} \left| d_2^\lambda \left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} \left(\frac{1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda}}{d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \right) \left(1 + \frac{a}{\mu_1} \rho^\lambda \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \right) \right| \stackrel{(i)}{=} \lim_{\lambda \rightarrow \infty} \left| d_2^\lambda \left(1 + \frac{1}{\mu_1} \lambda \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \right) \right| \stackrel{(ii)}{=} \infty, \end{aligned}$$

where (i) follows by noting that $\lim_{\lambda \rightarrow \infty} \left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} = \exp \left(\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \right) = 1$ (because $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} = 0$ by assumption), from Lemma 33, and by recalling that $\rho^\lambda = \frac{\lambda}{a}$; and (ii) follows because $\lim_{\lambda \rightarrow \infty} d_2^\lambda \neq 0$ (by assumption) and from Lemma 30 (d).

Therefore, from (A.106), it follows that $[\lim_{\lambda \rightarrow \infty} \left| \frac{\rho^\lambda C^\lambda}{d_2^\lambda} \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} (I^\lambda)^2 \right|]^{-1} = \infty$, or, equivalently, $\lim_{\lambda \rightarrow \infty} \frac{\rho^\lambda C^\lambda}{d_2^\lambda} \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} (I^\lambda)^2 = 0$. ■

A.7.3 Preliminaries C: Properties of the Limiting FOC

Recall the limiting FOC (1.16) is

$$c'(\mu) = p \left(1 - \left[1 - \frac{a^2}{\mu^2} \right]^+ \right) + v \frac{a}{\mu^2} \left[1 - \frac{a}{\mu} \right]^+. \quad (\text{A.107})$$

Let

$$h(\mu; a, p, v) := \frac{a^2}{\mu^2} \left(p + \frac{v}{a} - \frac{v}{\mu} \right), \quad \mu > 0. \quad (\text{A.108})$$

Then, (A.107) can be written as

$$c'(\mu) = \begin{cases} p, & \mu < a, \\ h(\mu; a, p, v), & \mu \geq a. \end{cases} \quad (\text{A.109})$$

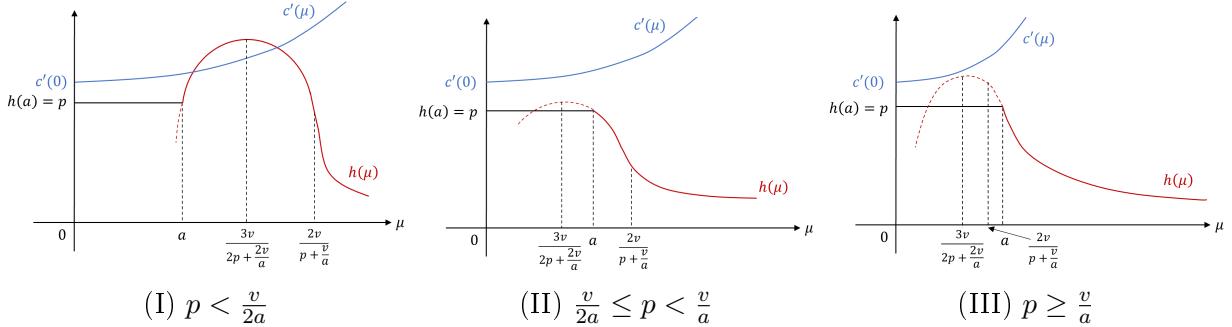


Figure A.4: Illustration of the left-hand side (solid blue curve) and the right-hand side (solid black and solid red curves) of the limiting FOC (A.109), suppose $p \leq c'(0)$.

The next lemma provides properties of $h(\mu; a, p, v)$. We might suppress the dependence of $h(\mu; a, p, v)$ on a, p and v henceforth, when the context is clear.

Lemma 40. *The function $h(\mu; a, p, v)$, defined in (A.108), satisfies the following properties:*

- (a) *$h(\mu)$ is strictly increasing in μ for $\mu \in \left(0, \frac{3v}{2p+\frac{2v}{a}}\right)$, and strictly decreasing in μ for $\mu \in \left(\frac{3v}{2p+\frac{2v}{a}}, \infty\right)$.*

- (b) $h(\mu)$ is strictly concave in μ for $\mu \in \left(0, \frac{2v}{p+\frac{v}{a}}\right)$, and strictly convex in μ for $\mu \in \left(\frac{2v}{p+\frac{v}{a}}, \infty\right)$.
- (c) $h(\mu; a, p, v)$ is strictly decreasing in $a > 0$ when $\mu \in \left(0, \frac{2v}{2p+\frac{v}{a}}\right)$, and strictly increasing in $a > 0$ when $\mu \in \left(\frac{2v}{2p+\frac{v}{a}}, \infty\right)$.
- (d) $h(\mu; a, p, v)$ is strictly increasing in $p \geq 0$ for all $\mu \in (0, \infty)$.
- (e) $h(\mu; a, p, v)$ is strictly decreasing in $v > 0$ when $\mu \in (0, a)$, and strictly increasing in $v > 0$ when $\mu \in (a, \infty)$.

A.7.3.1 Proof of Lemma 40

(a): Differentiating $h(\mu)$ yields

$$h'(\mu) = -\frac{2a^2}{\mu^3} \left(p + \frac{v}{a} - \frac{v}{\mu} \right) + \frac{a^2}{\mu^2} \frac{v}{\mu^2} = \frac{a^2}{\mu^3} \left(-2p - \frac{2v}{a} + \frac{3v}{\mu} \right),$$

which is strictly positive when $\mu < \frac{3v}{2p+\frac{2v}{a}}$, and strictly negative when $\mu > \frac{3v}{2p+\frac{2v}{a}}$. This implies that $h(\mu)$ is strictly increasing in μ for $\mu \in \left(0, \frac{3v}{2p+\frac{2v}{a}}\right)$ and strictly decreasing in μ for $\mu \in \left(\frac{3v}{2p+\frac{2v}{a}}, \infty\right)$.

(b): Differentiating $h(\mu)$ twice yields

$$h''(\mu) = -\frac{3a^2}{\mu^4} \left(-2p - \frac{2v}{a} + \frac{3v}{\mu} \right) + \frac{a^2}{\mu^3} \left(-\frac{3v}{\mu^2} \right) = \frac{6a^2}{\mu^4} \left(p + \frac{v}{a} - \frac{2v}{\mu} \right),$$

which is strictly negative when $\mu < \frac{2v}{p+\frac{v}{a}}$, and strictly positive when $\mu > \frac{2v}{p+\frac{v}{a}}$. This implies that $h(\mu)$ is strictly concave in μ for $\mu \in \left(0, \frac{2v}{p+\frac{v}{a}}\right)$ and strictly convex in μ for $\mu \in \left(\frac{2v}{p+\frac{v}{a}}, \infty\right)$.

(c): Differentiating $h(\mu; a, p, v)$ with respect to a yields

$$\frac{\partial h}{\partial a} = \frac{2a}{\mu^2} \left(p - \frac{v}{\mu} \right) + \frac{v}{\mu^2} = \frac{1}{\mu^2} \left(2ap + v - \frac{2av}{\mu} \right),$$

which is strictly negative when $\mu \in \left(0, \frac{2v}{2p+\frac{v}{a}}\right)$, and strictly positive when $\mu \in \left(\frac{2v}{2p+\frac{v}{a}}, \infty\right)$. This implies that $h(\mu; a, p, v)$ is strictly decreasing in $a > 0$ when $\mu \in \left(0, \frac{2v}{2p+\frac{v}{a}}\right)$, strictly increasing in $a > 0$ when $\mu \in \left[\frac{2v}{2p+\frac{v}{a}}, \infty\right)$.

(d): Differentiating $h(\mu; a, p, v)$ with respect to p yields

$$\frac{\partial h}{\partial p} = \frac{a^2}{\mu^2} > 0,$$

for all $\mu \in (0, \infty)$, which implies that $h(\mu; a, p, v)$ is strictly increasing in $p \geq 0$ for all $\mu \in (0, \infty)$.

(e): Differentiating $h(\mu; a, p, v)$ with respect to v yields

$$\frac{\partial h}{\partial v} = \frac{a^2}{\mu^2} \left(\frac{1}{a} - \frac{1}{\mu} \right),$$

which is strictly negative when $\mu \in (0, a)$, and strictly positive when $\mu \in (a, \infty)$. This implies that $h(\mu; a, p, v)$ is strictly decreasing in $v > 0$ when $\mu \in (0, a)$, and strictly increasing in $v > 0$ when $\mu \in (a, \infty)$.

■

A.7.4 Preliminaries D: Auxiliary Definitions

In this section, we introduce auxiliary results that will be heavily used in the proofs of all the results from Section 1.5.2-1.5.3, except that of Proposition 6. Noting that they all rely on Assumption 1, throughout this section, we implicitly operate under this assumption whenever it is necessary.

We begin with $\mu^\dagger(a, p)$, which is the unique value of $\mu \in \left[a, \frac{3}{2}a\right)$ that simultaneously satisfies $c'(\mu) = h(\mu)$ (the limiting FOC (A.109) for underloaded or critically loaded equilibria) and $c''(\mu) = h'(\mu)$, for a given $a > 0$ and $p \in [0, c'(a)]$. In other words, $c'(\mu)$ and $h(\mu)$ are tangent at $\mu^\dagger(a, p)$, as illustrated in Figure A.5 (I)(II). $v^\dagger(a, p)$ is the unique value of v that

induces $\mu^\dagger(a, p)$.

Definition 15.

(a) For any $a > 0$ and $p \in [0, c'(a)]$, $\mu^\dagger(a, p) \in \left[a, \frac{3}{2}a\right]$ is the unique solution for $\mu \in [a, \infty)$ that solves

$$\frac{\mu^2}{a^3} \left[(3a - 2\mu) c'(\mu) - (\mu - a)\mu c''(\mu) \right] = p. \quad (\text{A.110})$$

(b) For any $a > 0$ and $p \in [0, c'(a)]$,

$$v^\dagger(a, p) := \frac{(\mu^\dagger(a, p))^3}{a^2} \left(2c'(\mu^\dagger(a, p)) + \mu^\dagger(a, p)c''(\mu^\dagger(a, p)) \right). \quad (\text{A.111})$$

Remark 16. For any $a > 0$, when $p = c'(a)$, $\mu = a$ solves (A.110); therefore, by uniqueness, $\mu^\dagger(a, c'(a)) = a$. Then, (A.111) yields $v^\dagger(a, c'(a)) = 2ac'(a) + a^2c''(a)$.

Remark 17. For any $a > 0$ and $p \in [0, c'(a)]$, $v^\dagger(a, p) > 2ap$.

Lemma 41 (Validating Definition 15). For any $a > 0$, the two equations $h(\mu; a, p, v) = c'(\mu)$ and $h'(\mu; a, p, v) = c''(\mu)$ are simultaneously satisfied (i.e., $h(\mu; a, p, v)$ and $c'(\mu)$ are tangent) for some $p \geq 0$, $v > 0$, and $\mu \geq a$ if and only if $p \leq c'(a)$, $v = v^\dagger(a, p)$, and $\mu = \mu^\dagger(a, p) \in \left[a, \frac{3}{2}a\right]$.

Lemma 42. For any $a > 0$ and $p \in [0, c'(a)]$, $a \leq \mu^\dagger(a, p) < \frac{3v^\dagger(a, p)}{2p + \frac{2v^\dagger(a, p)}{a}} \leq \frac{3}{2}a$.

Lemma 43. For any $a > 0$ and $p \in [0, c'(a)]$, $c'(\mu) \geq h(\mu; a, p, v^\dagger(a, p))$ for all $\mu \in [a, \infty)$, with equality holding only at $\mu = \mu^\dagger(a, p)$.

Next, we revisit the piece-rate payment thresholds $p^\dagger(v)$ and $p^\ddagger(v)$, introduced in Theorem 4 and Proposition 3 (and used in Proposition 7), respectively.

Definition 16.

(a) For any $v > 0$, $p^\dagger(v)$ is the unique solution for $p \in (c'(0), \infty)$ that solves

$$p(c')^{-1}(p) + \frac{1}{2}((c')^{-1}(p))^2 c''((c')^{-1}(p)) = \frac{v}{2}. \quad (\text{A.112})$$

This is identical to the definition in Theorem 4.

(b) For any $v > 0$, $p^\ddagger(v)$ is the unique solution for $p \in (p^\dagger(v), \infty)$ that solves

$$p(c')^{-1}(p) = \frac{v}{2}. \quad (\text{A.113})$$

This is equivalent to the definition in Proposition 3.

Remark 18. When $v = 0$, (A.112) is satisfied only when $p = c'(0)$, due to the strict convexity of c . As a result, $\lim_{v \downarrow 0} p^\dagger(v) = c'(0)$.

Remark 19. $p^\dagger\left(v^\dagger(a, c'(a))\right) = c'(a)$, or, equivalently, $v^\dagger((c')^{-1}(p^\dagger(v)), p^\dagger(v)) = v$.

Lemma 44 (Validating Definition 16). For any $v > 0$,

(a) there exists a unique solution for $p \in (c'(0), \infty)$ that solves (A.112).

(b) there exists a unique solution for $p \in (p^\dagger(v), \infty)$ that solves (A.113).

The following result provides some useful monotonicity properties of the quantities defined in Definitions 15 and 16.

Lemma 45.

(a) $\mu^\dagger(a, p)$ is strictly increasing in $a > 0$ and strictly decreasing in $p \in [0, c'(a)]$.

(b) $v^\dagger(a, p)$ is strictly increasing in $a > 0$ and strictly decreasing in $p \in [0, c'(a)]$.

(c) $p^\dagger(v)$ is strictly increasing in $v > 0$.

(d) $p^\ddagger(v)$ is strictly increasing in $v > 0$.

Definition 17. For any $v > 0$ and $p \in [0, p^\dagger(v)]$, $\bar{a}(p, v)$ is the unique $a > 0$ such that $v^\dagger(a, p) = v$.

Remark 20. For any $v > 0$ and $p \in (0, p^\dagger(v)]$, it follows, from Definition 17 and Remark 17, that $v > 2\bar{a}(p, v)$, or, equivalently, $\bar{a}(p, v) < \frac{v}{2p}$.

Lemma 46. Fix $p \geq 0$. Let $A(p) := \{a > 0 : c'(a) \geq p\}$ and $V(p) := \{v > 0 : p^\dagger(v) \geq p\}$. Then, $v^\dagger(a, p) : A(p) \rightarrow V(p)$, given by (A.111), is an invertible function.

Corollary 5 (Validating Definition 17). For any $v > 0$ and $p \in [0, p^\dagger(v)]$, there exists unique $a > 0$ such that $v^\dagger(a, p) = v$.

Corollary 6. For any $v > 0$ and $p \in [0, p^\dagger(v)]$, $v < v^\dagger(a, p)$ if $a > \bar{a}(p, v)$; $v = v^\dagger(a, p)$ if $a = \bar{a}(p, v)$; and $v > v^\dagger(a, p)$ if $a < \bar{a}(p, v)$.

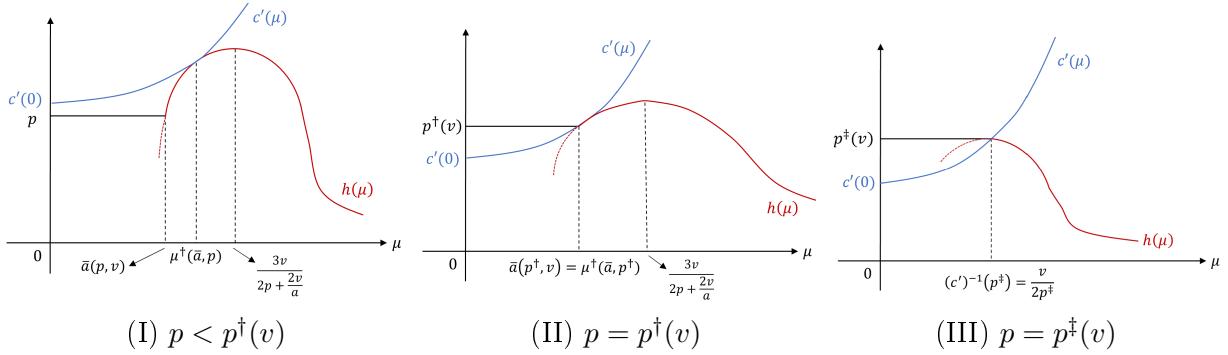


Figure A.5: Illustration of $\mu^\dagger(a, p)$, $p^\dagger(v)$, $p^\dagger(v)$ and $\bar{a}(p, v)$. Note that $v = v^\dagger(\bar{a}(p, v), p)$ in (I) and (II).

A.7.4.1 Proof of Remark 17

Recalling that c is strictly convex and $\mu^\dagger(a, p) \geq a$, from (A.111),

$$\begin{aligned} v^\dagger(a, p) &= \frac{2(\mu^\dagger(a, p))^3}{a^2} c'(\mu^\dagger(a, p)) + \frac{(\mu^\dagger(a, p))^4}{a^2} c''(\mu^\dagger(a, p)) \\ &\geq 2ac'(a) + a^2 c''(\mu^\dagger(a, p)) > 2ac'(a) \geq 2ap, \end{aligned}$$

where the last inequality follows because $p \leq c'(a)$. ■

A.7.4.2 Proof of Lemma 41

Recall, from (A.108), that $h(\mu; a, p, v) = \frac{a^2}{\mu^2} \left(p + \frac{v}{a} - \frac{v}{\mu} \right)$ for all $\mu > 0$. Therefore, the two equations $h(\mu; a, p, v) = c'(\mu)$ and $h'(\mu; a, p, v) = c''(\mu)$ are equivalent to

$$\frac{a^2}{\mu^2} \left(p + \frac{v}{a} - \frac{v}{\mu} \right) = c'(\mu) \quad \text{and} \quad \frac{a^2}{\mu^2} \left(\frac{3v}{\mu} - \frac{2v}{a} - 2p \right) = \mu c''(\mu), \quad (\text{A.114})$$

respectively. Combining these two equations by eliminating v , we obtain

$$\frac{\mu^2}{a^3} [(3a - 2\mu) c'(\mu) - (\mu - a) \mu c''(\mu)] = p,$$

which establishes (A.110). In order to study the properties of solutions to (A.110), we differentiate its left-hand side with respect to μ :

$$\begin{aligned} & \frac{d}{d\mu} \left(\frac{\mu^2}{a^3} [(3a - 2\mu) c'(\mu) - (\mu - a) \mu c''(\mu)] \right) \\ &= \frac{2\mu}{a^3} [(3a - 2\mu) c'(\mu) - (\mu - a) \mu c''(\mu)] \\ & \quad + \frac{\mu^2}{a^3} [-2c'(\mu) + (3a - 2\mu)c''(\mu) - \mu c''(\mu) - (\mu - a)(c''(\mu) + \mu c'''(\mu))] \\ &= -\frac{\mu(\mu - a)}{a^3} [6c'(\mu) + 6\mu c''(\mu) + \mu^2 c'''(\mu)], \end{aligned}$$

which, under Assumption 1, is strictly positive for $\mu \in (0, a)$, zero at $\mu = a$, and strictly negative for $\mu \in (a, \infty)$. This implies that the left-hand side of (A.110) attains a global maximum value of $c'(a)$ at $\mu = a$; therefore, (A.110) is true for some $p \geq 0$ only if $p \leq c'(a)$. Next, observe that the left-hand side of (A.110) is strictly decreasing in μ for $\mu \in [a, \infty)$, becoming negative at some $\mu \in (a, \frac{3}{2}a)$; therefore, given any $p \in [0, c'(a)]$, (A.110) admits a unique solution for $\mu \in [a, \infty)$, denoted by $\mu^\dagger(a, p) \in [a, \frac{3}{2}a]$. Finally, combining the two equations in (A.114) by eliminating p (instead of v), we obtain

$$\frac{\mu^3}{a^2} (2c'(\mu) + \mu c''(\mu)) = v,$$

which, after plugging in $\mu = \mu^\dagger(a, p)$, establishes (A.111); consequently, given $p \in [0, c'(a)]$, $v^\dagger(a, p)$ inherits its uniqueness from $\mu^\dagger(a, p)$. \blacksquare

A.7.4.3 Proof of Lemma 42

From Lemma 40 (a), it follows that $h'(\mu; a, p, v) > 0$ if and only if $\mu < \frac{3v}{2p + \frac{2v}{a}}$. By definition and from Lemma 41, given $a > 0$ and $p \in [0, c'(a)]$, when $v = v^\dagger(a, p)$, $\mu^\dagger(v, p) \in [a, \frac{3}{2}a]$ satisfies $h'(\mu^\dagger(a, p); a, p, v^\dagger) = c''(\mu^\dagger) > 0$ (due to the strict convexity of c); therefore, $\mu^\dagger(a, p) < \frac{3v^\dagger}{2p + \frac{2v^\dagger}{a}}$, which is naturally no greater than $\frac{3}{2}a$ for any $p \geq 0$. \blacksquare

A.7.4.4 Proof of Lemma 43

Given $a > 0$ and $p \in [0, c'(a)]$, when $v = v^\dagger$, it follows from Lemma 42 that $\mu^\dagger \in \left[a, \frac{3v^\dagger}{2p + \frac{2v^\dagger}{a}}\right]$ satisfies $c'(\mu^\dagger) = h(\mu^\dagger)$ and $c''(\mu^\dagger) = h'(\mu^\dagger)$. Observe that:

- Under Assumption 1, $c'(\mu)$ is convex in μ for all $\mu \in (0, \infty)$:

$$c''(\mu_1) \leq \frac{c'(\mu_2) - c'(\mu_1)}{\mu_2 - \mu_1} \leq c''(\mu_2) \quad \forall 0 < \mu_1 < \mu_2 < \infty. \quad (\text{A.115})$$

- From Lemmas 40 (b) and 42, $h(\mu)$ is strictly concave in μ for all $\mu \in \left[a, \frac{3v^\dagger}{2p + \frac{2v^\dagger}{a}}\right]$:

$$h'(\mu_2) < \frac{h(\mu_2) - h(\mu_1)}{\mu_2 - \mu_1} < h'(\mu_1) \quad \forall a \leq \mu_1 < \mu_2 \leq \frac{3v^\dagger}{2p + \frac{2v^\dagger}{a}}. \quad (\text{A.116})$$

First, for any $\mu \in [a, \mu^\dagger]$, we set $\mu_1 = \mu$, $\mu_2 = \mu^\dagger$, and use the fact that $c''(\mu^\dagger) = h'(\mu^\dagger)$ to combine (A.115) and (A.116) and obtain

$$\begin{aligned} \frac{c'(\mu^\dagger) - c'(\mu)}{\mu^\dagger - \mu} &< \frac{h(\mu^\dagger) - h(\mu)}{\mu^\dagger - \mu} \\ \xrightarrow{(*)} \quad c'(\mu) &> h(\mu) \quad \forall \mu \in [a, \mu^\dagger], \end{aligned} \quad (\text{A.117})$$

where $(*)$ follows due to the fact that $c'(\mu^\dagger) = h(\mu^\dagger)$. Similarly, second, for any $\mu \in \left(\mu^\dagger, \frac{3v^\dagger}{2p + \frac{2v^\dagger}{a}}\right]$, we set $\mu_1 = \mu^\dagger$, $\mu_2 = \mu$, and use the fact that $c''(\mu^\dagger) = h'(\mu^\dagger)$ to combine (A.115) and (A.116) and obtain

$$\begin{aligned} \frac{h(\mu) - h(\mu^\dagger)}{\mu - \mu^\dagger} &< \frac{c'(\mu) - c'(\mu^\dagger)}{\mu - \mu^\dagger} \\ \stackrel{(*)}{\implies} c'(\mu) > h(\mu) \quad \forall \mu \in \left(\mu^\dagger, \frac{3v^\dagger}{2p + \frac{2v^\dagger}{a}}\right], \end{aligned} \quad (\text{A.118})$$

where $(*)$ follows due to the fact that $c'(\mu^\dagger) = h(\mu^\dagger)$. Finally, since $c'(\mu)$ is strictly increasing in μ , $h(\mu)$ is decreasing in μ for all $\mu \in \left[\frac{3v^\dagger}{2p + \frac{2v^\dagger}{a}}, \infty\right)$ (Lemma 40 (a)), and $c'\left(\frac{3v^\dagger}{2p + \frac{2v^\dagger}{a}}\right) > h\left(\frac{3v^\dagger}{2p + \frac{2v^\dagger}{a}}\right)$ (special case of (A.118)), it follows that

$$c'(\mu) > h(\mu) \quad \forall \mu \in \left(\frac{3v^\dagger}{2p + \frac{2v^\dagger}{a}}, \infty\right). \quad (\text{A.119})$$

The proof is complete when combining (A.117)-(A.119) and recalling that $c'(\mu^\dagger) = h(\mu^\dagger)$.

■

A.7.4.5 Proof of Remark 19

From Remark 16, $v^\dagger(a, c'(a)) = 2ac'(a) + a^2c''(a) > 0$. Thus, from Definition 16 (a), $p^\dagger(v^\dagger(a, c'(a)))$ is the unique solution in p to

$$p(c')^{-1}(p) + \frac{1}{2}((c')^{-1}(p))^2 c''((c')^{-1}(p)) = ac'(a) + \frac{1}{2}a^2c''(a).$$

Note that $p = c'(a)$ solves the above equation. By uniqueness, it follows that $p^\dagger(v^\dagger(a, c'(a))) = c'(a)$. Similarly, when $p = p^\dagger(v) > c'(0)$ and $a = (c')^{-1}(p^\dagger(v))$, it follows from Remark 16 that $v^\dagger((c')^{-1}(p^\dagger(v)), p^\dagger(v)) = 2p^\dagger(v)(c')^{-1}(p^\dagger(v)) + ((c')^{-1}(p^\dagger(v)))^2 c''((c')^{-1}(p^\dagger(v))) \stackrel{(*)}{=} 2 \cdot \frac{v}{2} = v$, where $(*)$ follows from Definition 16 (a).

A.7.4.6 Proof of Lemma 44

(a): Denote the left-hand side of (A.112) by $LHS^\dagger(p)$. We first note that $LHS^\dagger(p)$ is a strictly increasing function of p (since c' is strictly increasing due to strict convexity of c , and c'' is also increasing by Assumption 1), and the right-hand side of (A.112), $\frac{v}{2}$, does not depend on p . Moreover, note that when $LHS^\dagger(c'(0)) = 0 < \frac{v}{2}$ and $\lim_{p \rightarrow \infty} LHS^\dagger(p) = \infty > \frac{v}{2}$. Hence, it follows that (A.112) admits a unique solution in $(c'(0), \infty)$.

(b): Denote the left-hand side of (A.113) by $LHS^\ddagger(p)$. We first note that $LHS^\ddagger(p)$ is a strictly increasing function of p (since c' is strictly increasing due to strict convexity of c), and the right-hand side of (A.113), $\frac{v}{2}$, does not depend on p . Moreover, note that $LHS^\dagger(p) > LHS^\ddagger(p)$ for all $p > 0$, while the right-hand sides of (A.112) and (A.113) are both $\frac{v}{2}$. This implies that $LHS^\dagger(p^\ddagger(v)) > LHS^\ddagger(p^\ddagger(v)) = \frac{v}{2} = LHS^\dagger(p^\dagger(v))$, and hence $p^\ddagger(v) > p^\dagger(v) > c'(0)$ (recalling that $LHS^\dagger(p)$ is a strictly increasing function of p). Additionally, note that $\lim_{p \rightarrow \infty} LHS^\ddagger(p) = \infty > \frac{v}{2}$. Hence, it follows that (A.113) admits a unique solution in $(p^\dagger(v), \infty)$. ■

A.7.4.7 Proof of Lemma 45

(a): By definition, $\mu^\dagger(a, p)$ is given by (A.110), which can be rewritten as

$$\mu^2 [(3a - 2\mu)c'(\mu) - (\mu - a)\mu c''(\mu)] = a^3 p. \quad (\text{A.120})$$

Differentiating both sides of (A.120) with respect to a yields

$$\begin{aligned} 2\mu \frac{\partial \mu}{\partial a} [(3a - 2\mu)c'(\mu) - (\mu - a)\mu c''(\mu)] + \mu^2 \left[\left(3 - 2\frac{\partial \mu}{\partial a} \right) c'(\mu) + (3a - 2\mu)c''(\mu) \frac{\partial \mu}{\partial a} \right. \\ \left. - \left(\frac{\partial \mu}{\partial a} - 1 \right) \mu c''(\mu) - (\mu - a) \left(\frac{\partial \mu}{\partial a} c''(\mu) + \mu c'''(\mu) \frac{\partial \mu}{\partial a} \right) \right] = 3a^2 p. \end{aligned}$$

After algebra,

$$[-6\mu(\mu-a)c'(\mu) - 6\mu^2(\mu-a)c''(\mu) - \mu^3(\mu-a)c'''(\mu)] \frac{\partial\mu}{\partial a} = 3a^2p - 3\mu^2c'(\mu) - \mu^3c''(\mu). \quad (\text{A.121})$$

Thus, $\mu^\dagger(a, p)$ satisfies (A.121). Note that the right-hand side of (A.121) at $\mu = \mu^\dagger(a, p)$ satisfies

$$3a^2p - 3\mu^2c'(\mu) - \mu^3c''(\mu) < 3a^2p - 3\mu^2c'(\mu) \leq 0,$$

noting that $\mu^\dagger(a, p) \geq a$ and $c'(\mu^\dagger(a, p)) \geq c'(a) \geq p$ (where $c'(a) \geq p$ is by definition of $\mu^\dagger(a, p)$ in Definition 15 (a)). Thus, the left-hand side of (A.121) is strictly negative. Moreover, the term in the square bracket on the left-hand side of (A.121) is also negative, since $\mu^\dagger(a, p) \geq a$, $c' > 0$, $c'' > 0$ and $c''' \geq 0$ (Assumption 1.9). Thus, $\frac{\partial\mu^\dagger(a,p)}{\partial a} > 0$, i.e., $\mu^\dagger(a, p)$ is strictly increasing in $a > 0$.

Similarly, differentiating both sides of (A.120) with respect to p yields

$$\begin{aligned} & 2\mu \frac{\partial\mu}{\partial p} [(3a - 2\mu)c'(\mu) - (\mu - a)\mu c''(\mu)] \\ & + \mu^2 \left[-2\frac{\partial\mu}{\partial p}c'(\mu) + (3a - 2\mu)c''(\mu)\frac{\partial\mu}{\partial p} - \frac{\partial\mu}{\partial p}\mu c''(\mu) - (\mu - a) \left(\frac{\partial\mu}{\partial p}c''(\mu) + \mu c'''(\mu)\frac{\partial\mu}{\partial p} \right) \right] = a^3. \end{aligned}$$

After algebra,

$$[-6\mu^2(\mu-a)c'(\mu) - 6\mu^2(\mu-a)c''(\mu) - \mu^3(\mu-a)c'''(\mu)] \frac{\partial\mu}{\partial p} = a^3.$$

Thus, $\mu^\dagger(a, p)$ satisfies the above equation. Note that the right-hand side of the above equation is strictly positive, which implies that the left-hand side of the above equation is also strictly positive. Moreover, note that the term in the square bracket on the left-hand side of the above equation is negative, because $\mu^\dagger(a, p) \geq a$, $c' > 0$, $c'' > 0$ and $c''' \geq 0$ (Assumption 1.9). Thus, $\frac{\partial\mu^\dagger(a,p)}{\partial p} < 0$, i.e., $\mu^\dagger(a, p)$ is strictly decreasing in $p \in [0, c'(a)]$.

(b): Differentiating both sides of (A.111) yields

$$\begin{aligned}\frac{\partial v^\dagger(a, p)}{\partial a} &= \frac{3(\mu^\dagger)^2 \frac{\partial \mu^\dagger}{\partial a} a^2 - (\mu^\dagger)^3 \cdot 2a}{a^4} \left[2c'(\mu^\dagger) + \mu^\dagger c''(\mu^\dagger) \right] \\ &\quad + \frac{(\mu^\dagger)^3}{a^2} \left[2c''(\mu^\dagger) \frac{\partial \mu^\dagger}{\partial a} + \frac{\partial \mu^\dagger}{\partial a} c''(\mu^\dagger) + \mu^\dagger c'''(\mu^\dagger) \frac{\partial \mu^\dagger}{\partial a} \right] \\ &= \left[6 \frac{(\mu^\dagger)^2}{a^2} c'(\mu^\dagger) + 6 \frac{(\mu^\dagger)^3}{a^2} c''(\mu^\dagger) + \frac{(\mu^\dagger)^4}{a^2} c'''(\mu^\dagger) \right] \frac{\partial \mu^\dagger}{\partial a} - 4 \frac{(\mu^\dagger)^4}{a^3} c'(\mu^\dagger) - 2 \frac{(\mu^\dagger)^4}{a^3} c''(\mu^\dagger).\end{aligned}$$

Substituting for $\frac{\partial \mu^\dagger}{\partial a}$ from (A.121) into the above display, and after algebra, implies

$$\begin{aligned}\frac{\partial v^\dagger(a, p)}{\partial a} &= \frac{\mu^\dagger}{a^2(\mu^\dagger - a)} \left(3(\mu^\dagger)^2 c'(\mu^\dagger) + (\mu^\dagger)^3 c''(\mu^\dagger) - 3a^2 p \right) - 4 \frac{(\mu^\dagger)^3}{a^3} c'(\mu^\dagger) - 2 \frac{(\mu^\dagger)^4}{a^3} c''(\mu^\dagger) \\ &= \left[\frac{3(\mu^\dagger)^3}{a^2(\mu^\dagger - a)} - 4 \frac{(\mu^\dagger)^3}{a^3} \right] c'(\mu^\dagger) + \left[\frac{(\mu^\dagger)^4}{a^2(\mu^\dagger - a)} - 2 \frac{(\mu^\dagger)^4}{a^3} \right] c''(\mu^\dagger) - \frac{\mu^\dagger}{a^2(\mu^\dagger - a)} 3a^2 p.\end{aligned}\tag{A.122}$$

Additionally, substituting for p in the last term of (A.122) using (A.120) yields

$$\begin{aligned}\frac{\mu^\dagger}{a^2(\mu^\dagger - a)} 3a^2 p &= \frac{3(\mu^\dagger)^3}{a^3(\mu^\dagger - a)} \left[(3a - 2\mu^\dagger) c'(\mu^\dagger) - (\mu^\dagger - a) \mu^\dagger c''(\mu^\dagger) \right] \\ &= \frac{3(\mu^\dagger)^3 (3a - 2\mu^\dagger)}{a^3(\mu^\dagger - a)} c'(\mu^\dagger) - 3 \frac{(\mu^\dagger)^4}{a^3} c''(\mu^\dagger).\end{aligned}$$

Substitution into (A.122) yields

$$\begin{aligned}\frac{\partial v^\dagger(a, p)}{\partial a} &= \left[\frac{3(\mu^\dagger)^3}{a^2(\mu^\dagger - a)} - 4 \frac{(\mu^\dagger)^3}{a^3} - \frac{3(\mu^\dagger)^3 (3a - 2\mu^\dagger)}{a^3(\mu^\dagger - a)} \right] c'(\mu^\dagger) + \left[\frac{(\mu^\dagger)^4}{a^2(\mu^\dagger - a)} - 2 \frac{(\mu^\dagger)^4}{a^3} + 3 \frac{(\mu^\dagger)^4}{a^3} \right] c''(\mu^\dagger) \\ &= 2 \frac{(\mu^\dagger)^3}{a^3} c'(\mu^\dagger) + \left[\frac{(\mu^\dagger)^4}{a^2(\mu^\dagger - a)} + \frac{(\mu^\dagger)^4}{a^3} \right] c''(\mu^\dagger) > 0,\end{aligned}$$

noting that $c' > 0$ and $c'' > 0$. Hence, $v^\dagger(a, p)$ is strictly increasing in $a > 0$.

Next, we investigate the monotonicity property of $v^\dagger(a, p)$ in terms of p . From (a), $\mu^\dagger(a, p)$ is strictly decreasing in $p \in [0, c'(a)]$; that is, $\mu^\dagger(a, p_1) > \mu^\dagger(a, p_2)$ for any $0 \leq p_1 < p_2 \leq c'(a)$. Recalling that c' is a strictly increasing function and c'' is also an increasing function (Assumption 1), it follows that $c'(\mu^\dagger(a, p_1)) > c'(\mu^\dagger(a, p_2))$ and $c''(\mu^\dagger(a, p_1)) \geq$

$c''(\mu^\dagger(a, p_2))$. Therefore, $v^\dagger(a, p_1) > v^\dagger(a, p_2)$ for $0 \leq p_1 < p_2 \leq c'(a)$; that is, $v^\dagger(a, p)$ is strictly decreasing in $p \in [0, c'(a)]$.

(c): As v increases, it is clear that the right-hand side of (A.112) increases. To equate the left-hand side of (A.112) with the right-hand side of (A.112), p should increase (because the left-hand side of (A.112) is strictly increasing in p). Hence, $p^\dagger(v)$ is strictly increasing in $v > 0$.

(d): Similar to (c), we can conclude that $p^\dagger(v)$ is strictly increasing in $v > 0$. ■

A.7.4.8 Proof of Lemma 46

Fix $p \geq 0$. First, we show that for any $a \in A(p)$, $v^\dagger(a, p) \in V(p)$. Recall that $v^\dagger(a, p)$ is strictly decreasing in $p \in [0, c'(a)]$ (Lemma 45 (b)) and $p^\dagger(v)$ is strictly increasing in $v > 0$ (Lemma 45 (c)). Thus, $a \in A(p) \Rightarrow p \leq c'(a) \Rightarrow v^\dagger(a, p) \geq v^\dagger(a, c'(a)) \Rightarrow p^\dagger(v^\dagger(a, p)) \geq p^\dagger(v^\dagger(a, c'(a)))$. Moreover, from Remark 19, we know that $p^\dagger(v^\dagger(a, c'(a))) = c'(a)$; therefore, $p^\dagger(v^\dagger(a, p)) \geq p^\dagger(v^\dagger(a, c'(a))) = c'(a) \geq p \Rightarrow v^\dagger(a, p) \in V(p)$.

Next, since $v^\dagger(a, p) : A(p) \rightarrow V(p)$ is a continuous and strictly increasing function of $a > 0$ (Lemma 45 (b)), it follows that it is a one-to-one function. What remains to be shown is that it is also an onto function. For this, we rely on the following claim, stated below without proof:

Claim 6. *A function $f : [a, \infty) \rightarrow [b, \infty)$ that is continuous and strictly increasing is an onto function if and only if $f(a) = f(b)$ and $f(x)$ grows unboundedly with x .*

In order to apply this claim to $v^\dagger(a, p) : A(p) \rightarrow V(p)$, we need to understand the structure of the sets $A(p)$ and $V(p)$, which depends upon the relationship between p and $c'(0)$. If $0 \leq p \leq c'(0)$, then $A(p) = V(p) = (0, \infty)$; otherwise, $A(p) = [(c')^{-1}(p), \infty)$ and $V(p) = [(p^\dagger)^{-1}(p), \infty)$. Therefore, to complete the proof, we must show the following:

- (i) $\lim_{a \uparrow \infty} v^\dagger(a, p) = \infty$ for all $p \geq 0$.

(ii) If $0 \leq p \leq c'(0)$, then $\lim_{a \downarrow 0} v^\dagger(a, p) = 0$.

(iii) If $p > c'(0)$, then $v^\dagger((c')^{-1}(p), p) = (p^\dagger)^{-1}(p)$, or, equivalently, $p^\dagger(v^\dagger((c')^{-1}(p), p)) = p$.

Proof of (i): For any $a > 0$ and $p \geq 0$, $c'(a) \geq p$ for all large enough a due to the strict convexity of c . Then, from Remark 17, $v^\dagger(a, p) > 2ap$ for all large enough a ; therefore, $\lim_{a \uparrow \infty} v^\dagger(a, p) = \infty$.

Proof of (ii): For any $a > 0$ and $p \leq c'(0)$, $c'(a) \geq p$ for all $a > 0$. In order to evaluate $\lim_{a \downarrow 0} v^\dagger(a, p)$ using Definition 15 (b), we must first evaluate $\lim_{a \downarrow 0} \mu^\dagger(a, p)$ using Definition 15 (a), according to which $\mu^\dagger(a, p) \in \left[a, \frac{3a}{2}\right)$; therefore, $\lim_{a \downarrow 0} \mu^\dagger(a, p) = 0$. Moreover, (A.110) of Definition 15 (a) can equivalently be written as

$$\left(\frac{\mu}{a}\right)^2 \left[\left(\frac{3}{2} - \frac{\mu}{a}\right) 2c'(\mu) - \left(\frac{\mu}{a} - 1\right) \mu c''(\mu) \right] = p. \quad (\text{A.123})$$

Letting $a \rightarrow 0$ in the above equation implies that, for any subsequence a' for which $\lim_{a' \downarrow 0} \frac{\mu^\dagger(a, p)}{a}$ exists, this limit must lie in the interval $\left[1, \frac{3}{2}\right]$. We simply use a rather than a' to denote the subsequence. Then, it would follow from Definition 15 (b) that

$$\lim_{a \downarrow 0} v^\dagger(a, p) = \left(\lim_{a \downarrow 0} \frac{\mu^\dagger(a, p)}{a} \right)^2 \lim_{a \downarrow 0} \mu^\dagger(a, p) \left(2c' \left(\lim_{a \downarrow 0} \mu^\dagger(a, p) \right) + \lim_{a \downarrow 0} \mu^\dagger(a, p) \cdot c'' \left(\lim_{a \downarrow 0} \mu^\dagger(a, p) \right) \right) = 0.$$

Proof of (iii): This follows immediately from Remark 19 by substituting $a = (c')^{-1}(p)$. ■

A.7.4.9 Proof of Corollary 5

It immediately follows from Lemma 46 that for any $v > 0$ and $p \in [0, p^\dagger(v)]$, there exists a unique a such that $v^\dagger(a, p) = v$ (by the invertible function $v^\dagger(a, p)$). In particular, when $a = (c')^{-1}(p)$ and $p = p^\dagger(v)$, Definitions 15(b) and 16(a) imply that $v^\dagger(a, p) = v^\dagger((c')^{-1}(p^\dagger(v)), p^\dagger(v)) = v$. Equivalently, $p^\dagger(v^\dagger(a, c'(a))) = c'(a)$. ■

A.7.4.10 Proof of Corollary 6

By Definition 17, $v^\dagger(a, p) = v$ if $a = \bar{a}(p, v)$. Recall from Lemma 45 (b) that $v^\dagger(a, p)$ is strictly increasing in $a > 0$. Then, if $a > \bar{a}(p, v)$, then $v^\dagger(a, p) > v^\dagger(\bar{a}(p, v), p) = v$. Similarly, if $a < \bar{a}(p, v)$, then $v^\dagger(a, p) < v^\dagger(\bar{a}(p, v), p) = v$. ■

A.7.5 Proof of Lemma 5

(a): Setting $\mu_1 = \mu$ in Propositions 16 and 17 yields

$$\lim_{\lambda \rightarrow \infty} I^\lambda(\mu, \mu) = \left[1 - \frac{a}{\mu}\right]^+ \text{ and } \lim_{\lambda \rightarrow \infty} \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} = \frac{a[\mu - a]^+}{\mu^3}.$$

(b): Suppose $N^\lambda = f(\lambda) + o(f(\lambda))$ for $f(\lambda) \in o(\lambda) \cap \omega(1)$ or $f(\lambda) \in \omega(\lambda)$.

- When $f(\lambda) \in o(\lambda) \cap \omega(1)$, let $N_0^\lambda = \frac{\lambda}{a} + N^\lambda$. Then, it is clear that $N^\lambda \leq N_0^\lambda$ for all $a > 0$ and $\lambda > 0$. From Lemma 18 (b), it follows that

$$\begin{aligned} ErlC\left(N^\lambda, \frac{\lambda}{\mu}\right) &\geq ErlC\left(N_0^\lambda, \frac{\lambda}{\mu}\right), \quad \forall a > 0, \forall \lambda > 0, \\ \Rightarrow \lim_{\lambda \rightarrow \infty} ErlC\left(N^\lambda, \frac{\lambda}{\mu}\right) &\geq \lim_{\lambda \rightarrow \infty} ErlC\left(N_0^\lambda, \frac{\lambda}{\mu}\right), \quad \forall a > 0. \\ \Rightarrow \lim_{\lambda \rightarrow \infty} ErlC\left(N^\lambda, \frac{\lambda}{\mu}\right) &\geq \sup_{a>0} \lim_{\lambda \rightarrow \infty} ErlC\left(N_0^\lambda, \frac{\lambda}{\mu}\right) = \infty, \end{aligned}$$

where the last equality follows from Lemma 29 (b) when $\mu < a$.

- When $f(\lambda) \in \omega(\lambda)$, let $N_0^\lambda = \frac{\lambda}{a}$ for any $a > 0$, so that $N_0^\lambda \in o(N^\lambda)$. Then, there exists $\Lambda(a) > 0$ such that $N^\lambda \geq N_0^\lambda$ for all $\lambda \geq \Lambda(a)$. From Lemma 18 (b), it follows that

$$\begin{aligned} ErlC\left(N^\lambda, \frac{\lambda}{\mu}\right) &\leq ErlC\left(N_0^\lambda, \frac{\lambda}{\mu}\right), \quad \forall a > 0, \forall \lambda > 0, \\ \Rightarrow \lim_{\lambda \rightarrow \infty} ErlC\left(N^\lambda, \frac{\lambda}{\mu}\right) &\leq \lim_{\lambda \rightarrow \infty} ErlC\left(N_0^\lambda, \frac{\lambda}{\mu}\right), \quad \forall a > 0. \\ \Rightarrow \lim_{\lambda \rightarrow \infty} ErlC\left(N^\lambda, \frac{\lambda}{\mu}\right) &\leq \inf_{a>0} \lim_{\lambda \rightarrow \infty} ErlC\left(N_0^\lambda, \frac{\lambda}{\mu}\right) = 0, \end{aligned}$$

where the last equality follows from Lemma 29 (b) when $\mu > a$.

Therefore, we conclude that

$$\lim_{\lambda \rightarrow \infty} ErlC \left(N^\lambda, \frac{\lambda}{\mu} \right) = \begin{cases} \infty, & f(\lambda) \in o(\lambda) \cap \omega(1), \\ 0, & f(\lambda) \in \omega(\lambda). \end{cases} \quad (\text{A.124})$$

Using (A.124) to evaluate the limiting value of $I^\lambda(\mu, \mu)$ from (A.20) in Corollary 2,

$$\lim_{\lambda \rightarrow \infty} I^\lambda(\mu, \mu) = \begin{cases} 0, & f(\lambda) \in o(\lambda) \cap \omega(1), \\ 1, & f(\lambda) \in \omega(\lambda). \end{cases}$$

Next, from Corollary 3 (a),

$$\lim_{\lambda \rightarrow \infty} \mu \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} \leq \lim_{\lambda \rightarrow \infty} I^\lambda(\mu, \mu) \left(1 - I^\lambda(\mu, \mu) \right) + \frac{2}{\sqrt{N^\lambda}} = 0,$$

because $\lim_{\lambda \rightarrow \infty} N^\lambda = \infty$, and $\lim_{\lambda \rightarrow \infty} I^\lambda(\mu, \mu) \left(1 - I^\lambda(\mu, \mu) \right) = 0$ regardless of whether $I^\lambda(\mu, \mu) = 0$ (when $f(\lambda) \in o(\lambda) \cap \omega(1)$) or 1 (when $f(\lambda) \in \omega(\lambda)$). Hence, $\lim_{\lambda \rightarrow \infty} \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} = 0$ by non-negativity (recalling from Lemma 21 (a) that $\lim_{\lambda \rightarrow \infty} \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \geq 0$ for all $\mu_1 > 0$ and $\mu > 0$). ■

A.7.6 Proof of Lemma 6

Recall that the tagged server's utility function in the finite system is given by (1.4). Using the notation established in Section 1.5.1, (1.4) becomes

$$U^\lambda(\mu_1, \mu) = p\mu_1 + (v - p\mu_1)I^\lambda(\mu_1, \mu) - c(\mu_1).$$

Differentiating with respect to μ_1 yields

$$\frac{\partial U^\lambda(\mu_1, \mu)}{\partial \mu_1} = p(1 - I^\lambda(\mu_1, \mu)) + (v - p\mu_1) \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} - c'(\mu_1).$$

Differentiating once more with respect to μ_1 yields

$$\begin{aligned} \frac{\partial^2 U^\lambda(\mu_1, \mu)}{\partial \mu_1^2} &= -2p \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} + (v - p\mu_1) \frac{\partial^2 I^\lambda}{\partial \mu_1^2} - c''(\mu_1) \\ &= v \frac{\partial^2 I^\lambda(\mu_1, \mu)}{\partial \mu_1^2} - p\mu_1 \left(\frac{\partial^2 I^\lambda(\mu_1, \mu)}{\partial \mu_1^2} + \frac{2}{\mu_1} \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \right) - c''(\mu_1). \end{aligned} \quad (\text{A.125})$$

For the remainder of this proof, we inherit the shorthand notation and the setup outlined around the proof of Proposition 16 in the preliminary Section A.7.2.

To begin, we recall, from Propositions 16 and 17, that

$$\lim_{\lambda \rightarrow \infty} I^\lambda = \begin{cases} 0, & \mu \leq a, \\ \frac{1-\frac{a}{\mu}}{1-\frac{a}{\mu}+\frac{a}{\mu_1}}, & \mu > a, \end{cases} \quad (\text{A.126})$$

and

$$\lim_{\lambda \rightarrow \infty} \frac{\partial I^\lambda}{\partial \mu_1} = \begin{cases} 0, & \mu \leq a, \\ \frac{\frac{a}{\mu_1^2}(1-\frac{a}{\mu})}{\left(1-\frac{a}{\mu}+\frac{a}{\mu_1}\right)^2}, & \mu > a. \end{cases} \quad (\text{A.127})$$

Next, we recall, from (A.17) in Lemma 22, that

$$\begin{aligned} \frac{\partial^2 I^\lambda}{\partial \mu_1^2} &= \frac{(I^\lambda)^2}{\mu_1^2} \left\{ -2(1 - I^\lambda) + \frac{2\rho^\lambda C^\lambda}{(d_2^\lambda)^2} \left[1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right] \left[(1 - 2I^\lambda) - \frac{\frac{\mu_1}{\mu} + \rho^\lambda \left(\frac{1-C^\lambda}{N^\lambda - \rho^\lambda} \right)}{d_2^\lambda} I^\lambda \right] \right. \\ &\quad \left. - \frac{\rho^\lambda C^\lambda}{d_2^\lambda} \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \left[2(1 - 2I^\lambda) + \left(\frac{2}{d_2^\lambda} - \frac{1}{d_1^\lambda} \right) \frac{\mu_1}{\mu} - 4 \frac{\frac{\mu_1}{\mu} + \rho^\lambda \left(\frac{1-C^\lambda}{N^\lambda - \rho^\lambda} \right)}{d_2^\lambda} I^\lambda \right] \right\} \end{aligned}$$

$$\begin{aligned}
& - \rho^\lambda C^\lambda \left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^2 \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \left[\frac{1}{d_2^\lambda} \frac{\mu_1}{\mu} - 2 \frac{\frac{\mu_1}{\mu} + \rho^\lambda \left(\frac{1-C^\lambda}{N^\lambda - \rho^\lambda} \right)}{d_2^\lambda} I^\lambda - \frac{2\rho^\lambda C^\lambda}{(d_2^\lambda)^2} I^\lambda \right] \Bigg) \\
& =: \left(\frac{I^\lambda}{\mu_1} \right)^2 \left\{ -2(1 - I^\lambda) + t_1^\lambda - t_2^\lambda - t_3^\lambda \right\}, \tag{A.128}
\end{aligned}$$

where

$$\begin{aligned}
t_1^\lambda &:= \frac{2\rho^\lambda C^\lambda}{(d_2^\lambda)^2} \left[1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right] \left[(1 - 2I^\lambda) - \frac{\frac{\mu_1}{\mu} + \rho^\lambda \left(\frac{1-C^\lambda}{N^\lambda - \rho^\lambda} \right)}{d_2^\lambda} I^\lambda \right], \\
t_2^\lambda &:= \frac{\rho^\lambda C^\lambda}{d_2^\lambda} \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \left[2(1 - 2I^\lambda) + \left(\frac{2}{d_2^\lambda} - \frac{1}{d_1^\lambda} \right) \frac{\mu_1}{\mu} - 4 \frac{\frac{\mu_1}{\mu} + \rho^\lambda \left(\frac{1-C^\lambda}{N^\lambda - \rho^\lambda} \right)}{d_2^\lambda} I^\lambda \right], \text{ and} \\
t_3^\lambda &:= \rho^\lambda C^\lambda \left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^2 \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \left[\frac{1}{d_2^\lambda} \frac{\mu_1}{\mu} - 2 \frac{\frac{\mu_1}{\mu} + \rho^\lambda \left(\frac{1-C^\lambda}{N^\lambda - \rho^\lambda} \right)}{d_2^\lambda} I^\lambda - \frac{2\rho^\lambda C^\lambda}{(d_2^\lambda)^2} I^\lambda \right].
\end{aligned}$$

From (A.125), and recalling that $c''(\mu_1) > 0$ for all $\mu_1 > 0$ (due to the strict convexity of c), it follows that, in order to show that $\lim_{\lambda \rightarrow \infty} \frac{\partial^2 U^\lambda(\mu_1, \mu)}{\partial \mu_1^2}$ is strictly less than 0, it suffices to show that

- $\lim_{\lambda \rightarrow \infty} \frac{\partial^2 I^\lambda}{\partial \mu_1^2} \leq 0$, and
- $\lim_{\lambda \rightarrow \infty} \left(\frac{\partial^2 I^\lambda}{\partial \mu_1^2} + \frac{2}{\mu_1} \frac{\partial I^\lambda}{\partial \mu_1} \right) \geq 0$.

To do this, in what follows, we evaluate $\lim_{\lambda \rightarrow \infty} \frac{\partial^2 I^\lambda}{\partial \mu_1^2}$ under three cases: (I) $\mu < a$, (II) $\mu > a$, and (III) $\mu = a$.

Case (I): If $\mu < a$, then $\frac{1-C^\lambda}{N^\lambda - \rho^\lambda} \in [0, 1]$ for all λ (from Lemma 30 (c)), $\left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} \leq 1$ for all large enough λ ; and, as $\lambda \rightarrow \infty$, $\frac{\rho^\lambda}{d_1^\lambda} \rightarrow \frac{a}{\mu} \in (0, \infty)$, $\frac{\rho^\lambda}{d_2^\lambda} \rightarrow (\frac{\mu}{a} - 1)^{-1} \in (-\infty, 0)$, $d_1^\lambda \rightarrow \infty$, $d_2^\lambda \rightarrow -\infty$, $\frac{C^\lambda}{d_2^\lambda} = \frac{C^\lambda}{\lambda} / \frac{d_2^\lambda}{\lambda} \rightarrow \frac{\mu}{a} - 1 \in (-\infty, 0)$ (from Lemma 30 (a)), $\frac{C^\lambda}{d_1^\lambda} = \frac{C^\lambda}{d_2^\lambda} / \frac{d_1^\lambda}{d_2^\lambda} \rightarrow \frac{\mu}{a} \left(1 - \frac{a}{\mu} \right)^2$, $I^\lambda \rightarrow 0$ (from Proposition 16), $\lim_{\lambda \rightarrow \infty} \left(k^\lambda - N^\lambda \right)^r \left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} < \infty$ for all $r \in \mathbb{N}$ (because even if $\limsup_{\lambda \rightarrow \infty} k^\lambda - N^\lambda = \infty$, exponential decay in terms of $k^\lambda - N^\lambda$

dominates its polynomial growth), and

$$\begin{aligned}
& \lim_{\lambda \rightarrow \infty} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} (I^\lambda)^2 \\
&= \lim_{\lambda \rightarrow \infty} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \left(1 + \rho^\lambda \frac{\mu}{\mu_1} \left(\frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} + \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right) \frac{C^\lambda}{d_2^\lambda} \right) \right)^{-2} \\
&= \lim_{\lambda \rightarrow \infty} \left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} \left(\left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} + \rho^\lambda \frac{\mu}{\mu_1} \left(\frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} + \left(\left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} - 1 \right) \frac{C^\lambda}{d_2^\lambda} \right) \right)^{-2} \\
&= 0,
\end{aligned}$$

recalling that $\left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} \leq 1$ for all large enough λ , $\frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \in [0, 1]$ for all λ , and $\lim_{\lambda \rightarrow \infty} \frac{C^\lambda}{d_2^\lambda} = \frac{\mu}{a} - 1$. Based on these facts, it can be shown that

$$\begin{aligned}
\lim_{\lambda \rightarrow \infty} (I^\lambda)^2 t_1^\lambda &= 2 \frac{\rho^\lambda}{d_2^\lambda} \frac{C^\lambda}{d_2^\lambda} \left[(I^\lambda)^2 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} (I^\lambda)^2 \right] \left[(1 - 2I^\lambda) - \frac{\frac{\mu_1}{\mu} + \rho^\lambda \left(\frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \right)}{d_2^\lambda} I^\lambda \right] = 0, \\
\lim_{\lambda \rightarrow \infty} (I^\lambda)^2 t_2^\lambda &= \frac{\rho^\lambda}{d_2^\lambda} \frac{C^\lambda}{d_1^\lambda} (k^\lambda - N^\lambda) \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} (I^\lambda)^2 \left[2(1 - 2I^\lambda) + \left(\frac{2}{d_2^\lambda} - \frac{1}{d_1^\lambda} \right) \frac{\mu_1}{\mu} - 4 \frac{\frac{\mu_1}{\mu} + \rho^\lambda \left(\frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \right)}{d_2^\lambda} I^\lambda \right] = 0, \\
\lim_{\lambda \rightarrow \infty} (I^\lambda)^2 t_3^\lambda &= \frac{\rho^\lambda}{d_1^\lambda} \frac{C^\lambda}{d_1^\lambda} (k^\lambda - N^\lambda)^2 \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} (I^\lambda)^2 \left[\frac{1}{d_2^\lambda} \frac{\mu_1}{\mu} - 2 \frac{\frac{\mu_1}{\mu} + \rho^\lambda \left(\frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \right)}{d_2^\lambda} I^\lambda - 2 \frac{\rho^\lambda}{d_2^\lambda} \frac{C^\lambda}{d_2^\lambda} I^\lambda \right] = 0,
\end{aligned}$$

which, from (A.128), implies that

$$\lim_{\lambda \rightarrow \infty} \frac{\partial^2 I^\lambda}{\partial \mu_1^2} = -\frac{2}{\mu_1^2} \left(\lim_{\lambda \rightarrow \infty} I^\lambda \right)^2 \left(1 - \lim_{\lambda \rightarrow \infty} I^\lambda \right) = 0. \quad (\text{A.129})$$

Together with (A.127), it follows that

$$\lim_{\lambda \rightarrow \infty} \frac{\partial^2 I^\lambda}{\partial \mu_1^2} + \frac{2}{\mu_1} \frac{\partial I^\lambda}{\partial \mu_1} = 0. \quad (\text{A.130})$$

Hence, from (A.125),

$$\lim_{\lambda \rightarrow \infty} \frac{\partial^2 U^\lambda}{\partial \mu_1^2} = \lim_{\lambda \rightarrow \infty} v \frac{\partial^2 I^\lambda}{\partial \mu_1^2} - p \mu_1 \left(\frac{\partial^2 I^\lambda}{\partial \mu_1^2} + \frac{2}{\mu_1} \frac{\partial I^\lambda}{\partial \mu_1} \right) - c''(\mu_1) < 0,$$

recalling that $\lim_{\lambda \rightarrow \infty} \frac{\partial^2 I^\lambda}{\partial \mu_1^2} = 0$ (from (A.129)), $\lim_{\lambda \rightarrow \infty} \frac{\partial^2 I^\lambda}{\partial \mu_1^2} + \frac{2}{\mu_1} \frac{\partial I^\lambda}{\partial \mu_1} = 0$ (from (A.130)), and $c'' > 0$ (since c is strictly convex).

Case (II): If $\mu > a$, then $\frac{1-C^\lambda}{N^\lambda-\rho^\lambda} \in [0, 1]$ for all λ (from Lemma 30 (c)), $I^\lambda \in [0, 1]$ for all λ , $\left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda-N^\lambda} \leq 1$ for all large enough λ ; and, as $\lambda \rightarrow \infty$, $\frac{\rho^\lambda}{d_1^\lambda} = \frac{a}{\mu} \in (0, \infty)$, $\frac{\rho^\lambda}{d_2^\lambda} \rightarrow (\frac{\mu}{a} - 1)^{-1} \in (0, \infty)$, $d_1^\lambda \rightarrow \infty$, $d_2^\lambda \rightarrow \infty$, $C^\lambda \rightarrow 0$, and $\lim_{\lambda \rightarrow \infty} (k^\lambda - N^\lambda)^r \left(\frac{\rho^\lambda}{d_1^\lambda}\right)^{k^\lambda-N^\lambda} < \infty$ for any $r \in \mathbb{N}$ (noting that $\frac{\rho^\lambda}{d_1^\lambda} < 1$ for all large enough λ , and, even if $\limsup_{\lambda \rightarrow \infty} k^\lambda - N^\lambda = \infty$, exponential decay in terms of $k^\lambda - N^\lambda$ dominates its polynomial growth). Based on these facts, it can be shown that

$$\begin{aligned} \lim_{\lambda \rightarrow \infty} t_1^\lambda &:= 2 \frac{\rho^\lambda}{d_2^\lambda} \frac{C^\lambda}{d_2^\lambda} \left[1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda-N^\lambda} \right] \left[(1 - 2I^\lambda) - \frac{\frac{\mu_1}{\mu} + \rho^\lambda \left(\frac{1-C^\lambda}{N^\lambda-\rho^\lambda} \right)}{d_2^\lambda} I^\lambda \right] = 0, \\ \lim_{\lambda \rightarrow \infty} t_2^\lambda &:= \frac{\rho^\lambda}{d_2^\lambda} C^\lambda \frac{1}{d_1^\lambda} k^\lambda - N^\lambda \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda-N^\lambda} \left[2(1 - 2I^\lambda) + \left(\frac{2}{d_2^\lambda} - \frac{1}{d_1^\lambda} \right) \frac{\mu_1}{\mu} - 4 \frac{\frac{\mu_1}{\mu} + \rho^\lambda \left(\frac{1-C^\lambda}{N^\lambda-\rho^\lambda} \right)}{d_2^\lambda} I^\lambda \right] = 0, \\ \lim_{\lambda \rightarrow \infty} t_3^\lambda &:= \frac{\rho^\lambda}{d_1^\lambda} C^\lambda \frac{1}{(d_1^\lambda)^2} (k^\lambda - N^\lambda)^2 \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda-N^\lambda} \left[\frac{1}{d_2^\lambda} \frac{\mu_1}{\mu} - 2 \frac{\frac{\mu_1}{\mu} + \rho^\lambda \left(\frac{1-C^\lambda}{N^\lambda-\rho^\lambda} \right)}{d_2^\lambda} I^\lambda - \frac{2\rho^\lambda C^\lambda}{(d_2^\lambda)^2} I^\lambda \right] = 0. \end{aligned}$$

which, from (A.128), implies that

$$\lim_{\lambda \rightarrow \infty} \frac{\partial^2 I^\lambda}{\partial \mu_1^2} = -\frac{2}{\mu_1^2} \left(\lim_{\lambda \rightarrow \infty} I^\lambda \right)^2 \left(1 - \lim_{\lambda \rightarrow \infty} I^\lambda \right) \stackrel{(*)}{=} -\frac{\frac{2a}{\mu_1^3} \left(1 - \frac{a}{\mu} \right)^2}{\left(1 - \frac{a}{\mu} + \frac{a}{\mu_1} \right)^3} < 0, \quad (\text{A.131})$$

where (*) follows from (A.126), and the last inequality follows by noting that $1 - \frac{a}{\mu} + \frac{a}{\mu_1} >$

$\frac{a}{\mu_1} > 0$ (when $\mu > a$). Thus, together with (A.127),

$$\lim_{\lambda \rightarrow \infty} \frac{\partial^2 I^\lambda}{\partial \mu_1^2} + \frac{2}{\mu_1} \frac{\partial I^\lambda}{\partial \mu_1} = -\frac{\frac{2a}{\mu_1^3} \left(1 - \frac{a}{\mu}\right)^2}{\left(1 - \frac{a}{\mu} + \frac{a}{\mu_1}\right)^3} + \frac{2}{\mu_1} \frac{\frac{a}{\mu_1^2} \left(1 - \frac{a}{\mu}\right)}{\left(1 - \frac{a}{\mu} + \frac{a}{\mu_1}\right)^2} = \frac{2a^2}{\mu_1^4} \frac{1 - \frac{a}{\mu}}{\left(1 - \frac{a}{\mu} + \frac{a}{\mu_1}\right)^3} > 0, \quad (\text{A.132})$$

where the last inequality follows by noting that $1 - \frac{a}{\mu} > 0$ and $1 - \frac{a}{\mu} + \frac{a}{\mu_1} > \frac{a}{\mu_1} > 0$ (when $\mu > a$). Hence, from (A.125),

$$\lim_{\lambda \rightarrow \infty} \frac{\partial^2 U^\lambda}{\partial \mu_1^2} = \lim_{\lambda \rightarrow \infty} v \frac{\partial^2 I^\lambda}{\partial \mu_1^2} - p \mu_1 \left(\frac{\partial^2 I^\lambda}{\partial \mu_1^2} + \frac{2}{\mu_1} \frac{\partial I^\lambda}{\partial \mu_1} \right) - c''(\mu_1) < 0,$$

recalling that $\lim_{\lambda \rightarrow \infty} \frac{\partial^2 I^\lambda}{\partial \mu_1^2} < 0$ (from (A.131)), $\lim_{\lambda \rightarrow \infty} \frac{\partial^2 I^\lambda}{\partial \mu_1^2} + \frac{2}{\mu_1} \frac{\partial I^\lambda}{\partial \mu_1} > 0$ (from (A.132)), and $c'' > 0$ (since c is strictly convex).

Case (III): If $\mu = a$, it suffices to show that if $\lim_{\lambda \rightarrow \infty} N^\lambda - \frac{\lambda}{a} \in \{-\infty\} \cup [1, \infty) \cup \{\infty\}$, then $\lim_{\lambda \rightarrow \infty} \frac{\partial^2 I^\lambda}{\partial \mu_1^2} = 0$. Then, from (A.125),

$$\lim_{\lambda \rightarrow \infty} \frac{\partial^2 U^\lambda}{\partial \mu_1^2} = \lim_{\lambda \rightarrow \infty} v \frac{\partial^2 I^\lambda}{\partial \mu_1^2} - p \mu_1 \left(\frac{\partial^2 I^\lambda}{\partial \mu_1^2} + \frac{2}{\mu_1} \frac{\partial I^\lambda}{\partial \mu_1} \right) - c''(\mu_1) < 0,$$

recalling that $\lim_{\lambda \rightarrow \infty} \frac{\partial I^\lambda}{\partial \mu_1} = 0$ (from (A.127)) and $c'' > 0$ (since c is strictly convex).

To show $\lim_{\lambda \rightarrow \infty} \frac{\partial^2 I^\lambda}{\partial \mu_1^2} = 0$ when $\lim_{\lambda \rightarrow \infty} N^\lambda - \frac{\lambda}{a} \in \{-\infty\} \cup [1, \infty) \cup \{\infty\}$, we need the following auxiliary claims, whose proofs are delayed until the end.

Claim 7. Under linear staffing (1.14), if $\mu = a$ and $0 < N^\lambda - \frac{\lambda}{a} \in \omega(1)$, then

$$\lim_{\lambda \rightarrow \infty} \sqrt{\rho^\lambda} C^\lambda \left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^r \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} I^\lambda = 0 \text{ for all } r \in \mathbb{N}.$$

Claim 8. Under linear staffing (1.14), if $\mu = a$ and $\lim_{\lambda \rightarrow \infty} d_2^\lambda \in (\mathbb{R} \setminus \{0\}) \cup \{-\infty, \infty\}$, then $\lim_{\lambda \rightarrow \infty} \frac{1}{I^\lambda} \left(\frac{\partial I^\lambda}{\partial \mu_1} \right)^2 = 0$.

Claim 9. If $\lim_{\lambda \rightarrow \infty} d_2^\lambda \in (\mathbb{R} \setminus \{0\}) \cup \{-\infty, \infty\}$ and $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in \mathbb{R} \cup \{\infty\}$, then

$$\lim_{\lambda \rightarrow \infty} \left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^r \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \in [0, \infty) \text{ for all } r \in \mathbb{N}.$$

Now, we are ready to complete the proof by dividing the condition $\lim_{\lambda \rightarrow \infty} N^\lambda - \frac{\lambda}{a} \in \{-\infty\} \cup [1, \infty) \cup \{\infty\}$ into three cases. Specifically, **Case (A)** $\lim_{\lambda \rightarrow \infty} N^\lambda - \frac{\lambda}{a} = \infty$, **Case (B)** $\lim_{\lambda \rightarrow \infty} N^\lambda - \frac{\lambda}{a} \in [1, \infty)$, and **Case (C)** $\lim_{\lambda \rightarrow \infty} N^\lambda - \frac{\lambda}{a} = -\infty$.

Using Lemma 22 and after algebra, one can check that, when $\mu = a$, $\frac{\partial^2 I^\lambda}{\partial \mu_1^2}$ can be equivalently written as

$$\begin{aligned} \frac{\partial^2 I^\lambda}{\partial \mu_1^2} &= -\frac{2}{\mu_1} I^\lambda \frac{\partial I^\lambda}{\partial \mu_1} - 2 \left[\frac{1}{\mu_1} \frac{\partial I^\lambda}{\partial \mu_1} - \frac{1}{I^\lambda} \left(\frac{\partial I^\lambda}{\partial \mu_1} \right)^2 - \frac{1}{\mu_1} I^\lambda \frac{\partial I^\lambda}{\partial \mu_1} - \frac{I^\lambda}{\mu_1 a d_2^\lambda} + \frac{1}{a d_2^\lambda} \frac{\partial I^\lambda}{\partial \mu_1} + \frac{(I^\lambda)^2}{\mu_1 a d_2^\lambda} \right] \\ &\quad + \frac{1}{\mu_1 a} \left(\frac{\rho^\lambda}{d_2^\lambda} I^\lambda \right) \frac{(k^\lambda - N^\lambda)(k^\lambda - N^\lambda + 1)}{(d_1^\lambda)^2} C^\lambda \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} I^\lambda. \end{aligned} \quad (\text{A.133})$$

Case (A): If $\lim_{\lambda \rightarrow \infty} N^\lambda - \frac{\lambda}{a} = \infty$. Note that the last term of (A.133) satisfies

$$\begin{aligned} &\frac{1}{\mu_1 a} \left(\frac{\rho^\lambda}{d_2^\lambda} I^\lambda \right) \frac{(k^\lambda - N^\lambda)(k^\lambda - N^\lambda + 1)}{(d_1^\lambda)^2} C^\lambda \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} I^\lambda \\ &= \frac{1}{\mu_1 a} \left(\frac{\sqrt{\rho^\lambda} I^\lambda}{d_2^\lambda} \right) \left(\sqrt{\rho^\lambda} C^\lambda \frac{(k^\lambda - N^\lambda)(k^\lambda - N^\lambda + 1)}{(d_1^\lambda)^2} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} I^\lambda \right) \rightarrow 0, \quad \text{as } \lambda \rightarrow \infty, \end{aligned}$$

by Lemma 36 and Claim 7. Using this, together with Claim 8, and $\lim_{\lambda \rightarrow \infty} I^\lambda = 0$ and $\lim_{\lambda \rightarrow \infty} \frac{\partial I^\lambda}{\partial \mu_1} = 0$ (from Propositions 16 and 17 when $\mu = a$), one can easily check that every term in (A.133) converges to 0 as $\lambda \rightarrow \infty$, implying that $\lim_{\lambda \rightarrow \infty} \frac{\partial^2 I^\lambda}{\partial \mu_1^2} = 0$.

Case (B): If $\lim_{\lambda \rightarrow \infty} N^\lambda - \frac{\lambda}{a} \in [1, \infty)$, we further discuss cases depending on the value of $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}$.

Case (B-1): If $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} = 0$. One can easily check that every term in (A.133) converges to 0 as $\lambda \rightarrow \infty$, using $\lim_{\lambda \rightarrow \infty} I^\lambda = 0$ and $\lim_{\lambda \rightarrow \infty} \frac{\partial I^\lambda}{\partial \mu_1} = 0$ (from Propositions 16 and 17 when $\mu = a$), together with Lemma 39 and Claim 8. Hence, $\lim_{\lambda \rightarrow \infty} \frac{\partial^2 I^\lambda}{\partial \mu_1^2} = 0$.

Case (B-2): If $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in (0, \infty) \cup \{\infty\}$. Note that the last term of (A.133)

satisfies

$$\begin{aligned} & \frac{1}{\mu_1 a} \left(\frac{\rho^\lambda}{d_2^\lambda} I^\lambda \right) \frac{(k^\lambda - N^\lambda)(k^\lambda - N^\lambda + 1)}{(d_1^\lambda)^2} C^\lambda \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} I^\lambda \\ &= \frac{1}{\mu_1 a} \frac{1}{d_2^\lambda} \left(\sqrt{\rho^\lambda} I^\lambda \right)^2 \left(\frac{(k^\lambda - N^\lambda)(k^\lambda - N^\lambda + 1)}{(d_1^\lambda)^2} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right) C^\lambda \rightarrow 0, \text{ as } \lambda \rightarrow \infty, \end{aligned}$$

by noting that $\lim_{\lambda \rightarrow \infty} C^\lambda = 1$ (from Lemma 29 (b)), and by Lemma 37 (ii) and Claim 9. Using this, together with Claim 8, and $\lim_{\lambda \rightarrow \infty} I^\lambda = 0$ and $\lim_{\lambda \rightarrow \infty} \frac{\partial I^\lambda}{\partial \mu_1} = 0$ (from Propositions 16 and 17 when $\mu = a$), one can easily check that every term in (A.133) converges to 0 as $\lambda \rightarrow \infty$, implying that $\lim_{\lambda \rightarrow \infty} \frac{\partial^2 I^\lambda}{\partial \mu_1^2} = 0$.

Case (C) If $\lim_{\lambda \rightarrow \infty} N^\lambda - \frac{\lambda}{a} = -\infty$, we further discuss cases depending on the value of $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}$.

Case (C-1): If $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} = 0$. One can easily check that every term in (A.133) converges to 0 as $\lambda \rightarrow \infty$, using $\lim_{\lambda \rightarrow \infty} I^\lambda = 0$ and $\lim_{\lambda \rightarrow \infty} \frac{\partial I^\lambda}{\partial \mu_1} = 0$ (from Propositions 16 and 17 when $\mu = a$), together with Lemma 39 and Claim 8.

Case (C-2): If $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in \{-\infty\} \cup (-\infty, 0)$. Note that the term of (A.133) satisfies

$$\begin{aligned} & \frac{1}{\mu_1 a} \left(\frac{\rho^\lambda}{d_2^\lambda} I^\lambda \right) \frac{(k^\lambda - N^\lambda)(k^\lambda - N^\lambda + 1)}{(d_1^\lambda)^2} C^\lambda \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} I^\lambda \\ &= \frac{1}{\mu_1 a} \left(\frac{\rho^\lambda C^\lambda}{d_2^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} I^\lambda \right) \left(\frac{(k^\lambda - N^\lambda)(k^\lambda - N^\lambda + 1)}{(d_1^\lambda)^2} I^\lambda \right) \rightarrow 0, \text{ as } \lambda \rightarrow \infty, \end{aligned}$$

by Lemmas 34 and 35. Using this, together with Claim 8, and $\lim_{\lambda \rightarrow \infty} I^\lambda = 0$ and $\lim_{\lambda \rightarrow \infty} \frac{\partial I^\lambda}{\partial \mu_1} = 0$ (from Propositions 16 and 17 when $\mu = a$), one can easily check that every term in (A.133) converges to 0 as $\lambda \rightarrow \infty$, implying that $\lim_{\lambda \rightarrow \infty} \frac{\partial^2 I^\lambda}{\partial \mu_1^2} = 0$.

Combining Cases (A)-(C), when $\mu = a$, if $\lim_{\lambda \rightarrow \infty} N^\lambda - \frac{\lambda}{a} \in \{-\infty\} \cup [1, \infty) \cup \{\infty\}$, then $\lim_{\lambda \rightarrow \infty} \frac{\partial^2 I^\lambda}{\partial \mu_1^2} = 0$. This concludes the proof of Case (III).

Together Cases (I)-(III) establish that, under the staffing rule (1.14), except when $\lim_{\lambda \rightarrow \infty} N^\lambda -$

$\frac{\lambda}{a}$ is finite and strictly less than 1, for all large enough λ , we have $\frac{\partial^2 U^\lambda}{\partial \mu_1^2} < 0$. ■

Proof of Claim 7

When $0 < N^\lambda - \frac{\lambda}{\mu} \in \omega(1)$, it is clear that both d_2^λ and $d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}$ are non-negative for all large enough λ . Furthermore, $\lim_{\lambda \rightarrow \infty} d_2^\lambda = \infty$. Then, by multiplying and dividing by $(d_2^\lambda)^r$ and regrouping the terms, we can write

$$\lim_{\lambda \rightarrow \infty} \sqrt{\rho^\lambda} C^\lambda \left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^r \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} I^\lambda = \lim_{\lambda \rightarrow \infty} \left(\frac{\sqrt{\rho^\lambda} I^\lambda}{d_2^\lambda} \right) \cdot C^\lambda \cdot \frac{1}{(d_2^\lambda)^{r-1}} \cdot \left[\left(d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^r e^{-d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \right] = 0,$$

because $\lim_{\lambda \rightarrow \infty} \frac{\sqrt{\rho^\lambda} I^\lambda}{d_2^\lambda} = 0$ (from Lemma 36), $\lim_{\lambda \rightarrow \infty} C^\lambda \in [0, 1]$ (from Lemma 29(b)), and $\frac{1}{(d_2^\lambda)^{r-1}} \in \{0, 1\}$ and $\lim_{\lambda \rightarrow \infty} \left(d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^r e^{-d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \in [0, \infty)$ (because even if $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} = \infty$, exponential decay in terms of $d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}$ would dominate its polynomial growth) for all $r \in \mathbb{N}$. ■

Proof of Claim 8

We discuss two cases depending on the value of $\lim_{\lambda \rightarrow \infty} d_2^\lambda$.

Case (I): If $\lim_{\lambda \rightarrow \infty} d_2^\lambda \in (0, \infty) \cup \{\infty\}$, then $d_2^\lambda > 0$ for all large enough λ . Then, recalling that $\frac{1-C^\lambda}{N^\lambda-\rho^\lambda} > 0$ for all λ (from Lemma 19 (c)), it follows from (A.89) that

$$\begin{aligned} \frac{\partial I^\lambda}{\partial \mu_1} &\leq \frac{1}{\mu_1} \left[\left(1 + \frac{1}{d_2^\lambda} \frac{\mu_1}{\mu} \right) I^\lambda (1 - I^\lambda) \right] \\ \Rightarrow \lim_{\lambda \rightarrow \infty} \frac{1}{I^\lambda} \left(\frac{\partial I^\lambda}{\partial \mu_1} \right)^2 &\leq \lim_{\lambda \rightarrow \infty} \frac{1}{\mu_1^2} \left(1 + \frac{1}{d_2^\lambda} \frac{\mu_1}{\mu} \right)^2 I^\lambda (1 - I^\lambda)^2 = 0, \end{aligned}$$

recalling that $\lim_{\lambda \rightarrow \infty} I^\lambda = 0$ (from Proposition 16). Hence, $\lim_{\lambda \rightarrow \infty} \frac{1}{I^\lambda} \left(\frac{\partial I^\lambda}{\partial \mu_1} \right)^2 = 0$, by non-negativity ($I^\lambda > 0$ for all λ).

Case (II): If $\lim_{\lambda \rightarrow \infty} d_2^\lambda \in \{-\infty\} \cup (-\infty, 0)$, then $d_2^\lambda < 0$ for all large enough λ , which implies that (i) either $0 < \frac{\lambda}{a} - N^\lambda \in \omega(1)$ or $|N^\lambda - \frac{\lambda}{a}| \in \mathcal{O}(1)$ and (ii) $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \leq 0$.

It follows from (A.89) that

$$\begin{aligned}
\frac{1}{I^\lambda} \left(\frac{\partial I^\lambda}{\partial \mu_1} \right)^2 &= \frac{1}{I^\lambda} \frac{1}{\mu_1^2} \left(\left[\left(1 + \frac{1}{d_2^\lambda} \frac{\mu_1}{\mu} \right) I^\lambda (1 - I^\lambda) \right] - \left[\frac{\rho^\lambda}{d_2^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} (I^\lambda)^2 \right] - \left[\frac{\rho^\lambda C^\lambda}{d_2^\lambda} \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} (I^\lambda)^2 \right] \right)^2 \\
&= \frac{1}{\mu_1^2} \left(\left[\left(1 + \frac{1}{d_2^\lambda} \frac{\mu_1}{\mu} \right) (I^\lambda)^{\frac{1}{2}} (1 - I^\lambda) \right] - \left[\frac{\rho^\lambda}{d_2^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} (I^\lambda)^{\frac{3}{2}} \right] - \left[\frac{\rho^\lambda C^\lambda}{d_2^\lambda} \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} (I^\lambda)^{\frac{3}{2}} \right] \right)^2.
\end{aligned} \tag{A.134}$$

Note that the first term in square brackets on the right-hand side of (A.134) vanishes in the limit as $\lambda \rightarrow \infty$, by recalling that $\lim_{\lambda \rightarrow \infty} d_2^\lambda \in \{-\infty\} \cup (-\infty, 0)$ (by assumption) and $\lim_{\lambda \rightarrow \infty} I^\lambda = 0$ (from Proposition 16). To complete the proof, it suffices to show that the second and third terms in square brackets follow suit.

Second Term: We further discuss three cases depending on the asymptotic behavior of $N^\lambda - \frac{\lambda}{a}$.

- Suppose $0 < \frac{\lambda}{a} - N^\lambda \in \omega(\sqrt{\lambda})$. Then, the second term can be rearranged as follows:

$$\frac{\rho^\lambda}{d_2^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} (I^\lambda)^{\frac{3}{2}} = \left(\frac{\rho^\lambda C^\lambda I^\lambda}{d_2^\lambda} \right) \left(\frac{1}{C^\lambda} - 1 \right) \frac{1}{N^\lambda - \frac{\lambda}{a}} (I^\lambda)^{\frac{1}{2}},$$

which vanishes in the limit as $\lambda \rightarrow \infty$ by noting that $\lim_{\lambda \rightarrow \infty} \frac{\rho^\lambda C^\lambda I^\lambda}{d_2^\lambda} \in [-\frac{\mu_1}{a}, 0]$ (by Lemma 38) and $\lim_{\lambda \rightarrow \infty} C^\lambda = \infty$ (from Lemma 29 (b)) and by recalling that $\frac{\lambda}{a} - N^\lambda \in \omega(\sqrt{\lambda})$ and $\lim_{\lambda \rightarrow \infty} I^\lambda = 0$ (from Proposition 16).

- Suppose $0 < \frac{\lambda}{a} - N^\lambda \in \mathcal{O}(\sqrt{\lambda}) \cap \omega(1)$. Then, recalling that $\rho^\lambda = \frac{\lambda}{a}$ when $\mu = a$, the second term can be rearranged as follows:

$$\frac{\rho^\lambda}{d_2^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} (I^\lambda)^{\frac{3}{2}} = \left(\frac{\sqrt{\rho^\lambda} I^\lambda}{d_2^\lambda} \right) \frac{1}{\sqrt{a}} \left(\sqrt{\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \right) (I^\lambda)^{\frac{1}{2}},$$

which vanishes in the limit as $\lambda \rightarrow \infty$ by noting that $\lim_{\lambda \rightarrow \infty} \frac{\sqrt{\rho^\lambda} I^\lambda}{d_2^\lambda} = 0$ (by Lemma 36) and $\lim_{\lambda \rightarrow \infty} \sqrt{\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \in (0, \infty)$ (from Lemma 30 (c)) and by recalling that $\lim_{\lambda \rightarrow \infty} I^\lambda = 0$ (from Proposition 16).

- Suppose $\left|N^\lambda - \frac{\lambda}{a}\right| \in \mathcal{O}(1)$. Then, recalling that $\rho^\lambda = \frac{\lambda}{a}$ when $\mu = a$, the second term can be rearranged as follows:

$$\frac{\rho^\lambda}{d_2^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} (I^\lambda)^{\frac{3}{2}} = \frac{1}{d_2^\lambda} \left((\rho^\lambda)^{\frac{1}{3}} I^\lambda \right)^{\frac{3}{2}} \frac{1}{\sqrt{a}} \left(\sqrt{\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \right),$$

which vanishes in the limit as $\lambda \rightarrow \infty$ by recalling that $\lim_{\lambda \rightarrow \infty} d_2^\lambda \neq 0$ (by assumption) and by noting that $\lim_{\lambda \rightarrow \infty} (\rho^\lambda)^{\frac{1}{3}} I^\lambda = 0$ (by Lemma 37) and $\lim_{\lambda \rightarrow \infty} \sqrt{\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \in (0, \infty)$ (from Lemma 30 (c)).

Third Term: We further discuss three cases depending on $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}$ and the asymptotic behavior of $N^\lambda - \frac{\lambda}{a}$.

- Suppose $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in \{-\infty\} \cup (-\infty, 0)$. Then, the third term can be rearranged as follows:

$$\frac{\rho^\lambda C^\lambda}{d_2^\lambda} \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} (I^\lambda)^{\frac{3}{2}} = \left(\frac{\rho^\lambda C^\lambda}{d_2^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} I^\lambda \right) \left(\left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^2 I^\lambda \right)^{\frac{1}{2}},$$

which vanishes in the limit as $\lambda \rightarrow \infty$ by noting that $\lim_{\lambda \rightarrow \infty} \frac{\rho^\lambda C^\lambda}{d_2^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} I^\lambda \in (-\infty, \infty)$ (by Lemma 35) and $\lim_{\lambda \rightarrow \infty} \left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^2 I^\lambda = 0$ (by Lemma 34 (ii)).

- Suppose $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} = 0$ and $0 < \frac{\lambda}{a} - N^\lambda \in \omega(\sqrt{\lambda})$. Then, the third term can be rearranged as follows:

$$\frac{\rho^\lambda C^\lambda}{d_2^\lambda} \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} (I^\lambda)^{\frac{3}{2}} = \left(\frac{\rho^\lambda C^\lambda I^\lambda}{d_2^\lambda} \right) \left(\left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^2 I^\lambda \right)^{\frac{1}{2}} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda},$$

which vanishes in the limit as $\lambda \rightarrow \infty$ by noting that $\lim_{\lambda \rightarrow \infty} \frac{\rho^\lambda C^\lambda I^\lambda}{d_2^\lambda} \in [-\frac{\mu_1}{a}, 0]$ (by Lemma 38), $\lim_{\lambda \rightarrow \infty} \left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^2 I^\lambda = 0$ (by Lemma 34 (ii)), and $\lim_{\lambda \rightarrow \infty} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} = 1$ (from Lemma 33 (i)).

- Suppose $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} = 0$ and $\left| N^\lambda - \frac{\lambda}{a} \right| \in \mathcal{O}(\sqrt{\lambda})$. Then, recalling that $d_2^\lambda < 0$ for all large enough λ , the third term can be expressed as follows:

$$\frac{\rho^\lambda C^\lambda}{d_2^\lambda} \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} (I^\lambda)^{\frac{3}{2}} = - \left[\left(-\frac{\rho^\lambda C^\lambda}{d_2^\lambda} \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right)^{\frac{4}{3}} (I^\lambda)^2 \right]^{\frac{3}{4}}.$$

To complete the proof, we show that the expression within the square brackets above vanishes in the limit (by showing that its reciprocal diverges to ∞) as $\lambda \rightarrow \infty$. Using (A.88) and after algebra, when $\mu = a$,

$$\begin{aligned} & \left[\left(-\frac{\rho^\lambda C^\lambda}{d_2^\lambda} \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right)^{\frac{4}{3}} (I^\lambda)^2 \right]^{-1} \\ &= \left(\frac{(d_2^\lambda)^2}{\rho^\lambda C^\lambda} \left(\frac{\left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda}}{-d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \right) \right)^{\frac{2}{3}} \left(1 + \frac{a}{\mu_1} \rho^\lambda \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \right)^2 \\ &\quad + 2 \left(\frac{a}{\mu_1} \frac{\rho^\lambda C^\lambda}{d_2^\lambda} \right) \left(\frac{(d_2^\lambda)^2}{\rho^\lambda C^\lambda} \left(\frac{\left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda}}{-d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \right) \right)^{\frac{2}{3}} \left(1 + \frac{a}{\mu_1} \rho^\lambda \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \right) \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right) \\ &\quad + \left(\frac{a}{\mu_1} \frac{\rho^\lambda C^\lambda}{d_2^\lambda} \right)^2 \left(\frac{(d_2^\lambda)^2}{\rho^\lambda C^\lambda} \left(\frac{\left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda}}{-d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \right) \right)^{\frac{2}{3}} \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right)^2. \end{aligned} \tag{A.135}$$

Observe that $\frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \geq 0$ (from Lemma 19 (c)) and $\frac{1}{d_2^\lambda} \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right) \geq 0$ (recalling the definition $d_2^\lambda := d_1^\lambda - \rho^\lambda$) for all λ . As a result, each of the three terms on the right-hand side of (A.135) is non-negative for all λ . Therefore, the left-hand side of (A.135) will diverge to ∞ as $\lambda \rightarrow \infty$ even if one of these three terms grows unboundedly with λ . To complete the proof, we show that the first term on the right-hand side of (A.135) diverges to ∞ as $\lambda \rightarrow \infty$. Recalling that $\rho^\lambda = \frac{\lambda}{a}$ when $\mu = a$, this term can

be rearranged as follows:

$$\begin{aligned} & \left(\frac{(d_2^\lambda)^2}{\rho^\lambda C^\lambda} \left(\frac{\left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda}}{-d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \right) \right)^{\frac{2}{3}} \left(1 + \frac{a}{\mu_1} \rho^\lambda \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \right)^2 \\ &= \left(\frac{\sqrt{\rho^\lambda} (d_2^\lambda)^2}{C^\lambda} \left(\frac{\left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda}}{-d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \right) \right)^{\frac{2}{3}} \left(\frac{1}{\sqrt{\rho^\lambda}} + \frac{\sqrt{a}}{\mu_1} \left(\sqrt{\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \right) \right)^2, \end{aligned}$$

which diverges to ∞ as $\lambda \rightarrow \infty$ by recalling that $\lim_{\lambda \rightarrow \infty} \rho^\lambda = \infty$, $\lim_{\lambda \rightarrow \infty} d_2^\lambda \neq 0$ (by assumption), and $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} = 0$; and by noting that $\lim_{\lambda \rightarrow \infty} \left(\frac{d_1^\lambda}{\rho^\lambda} \right)^{k^\lambda - N^\lambda} = 1$ (from Lemma 33 (i)), $\lim_{\lambda \rightarrow \infty} C^\lambda \in (0, \infty)$ (from Lemma 29 (b), recalling that $\left| N^\lambda - \frac{\lambda}{a} \right| \in \mathcal{O}(\sqrt{\lambda})$), and $\lim_{\lambda \rightarrow \infty} \sqrt{\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \in (0, \infty)$ (from Lemma 30 (c)).

■

Proof of Claim 9

First, we recall the definition $d_2^\lambda := d_1^\lambda - \rho^\lambda$, where $d_1^\lambda := N^\lambda - 1 + \frac{\mu_1}{\mu} > 0$ (since $N^\lambda \geq 1$, $\mu_1 > 0$, and $\mu > 0$) and $\rho^\lambda := \frac{\lambda}{\mu} > 0$ for all $\lambda > 0$. This implies that $\frac{d_2^\lambda}{d_1^\lambda} < 1$ for all $\lambda > 0$. Next, we recall that $\lim_{\lambda \rightarrow \infty} d_2^\lambda \neq 0$ (by assumption), which implies that $d_2^\lambda \neq 0$ for all large enough λ . By multiplying and dividing by $(d_2^\lambda)^r$ and regrouping the terms, we can write

$$\begin{aligned} 0 &\leq \lim_{\lambda \rightarrow \infty} \left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^r \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} = \lim_{\lambda \rightarrow \infty} \frac{1}{(d_2^\lambda)^r} \left(d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^r \left(1 - \frac{d_2^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \\ &= \lim_{\lambda \rightarrow \infty} \frac{1}{(d_2^\lambda)^r} \left(d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^r \exp \left((k^\lambda - N^\lambda) \ln \left(1 - \frac{d_2^\lambda}{d_1^\lambda} \right) \right) \\ &\stackrel{(i)}{\leq} \lim_{\lambda \rightarrow \infty} \frac{1}{(d_2^\lambda)^r} \left(d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^r \exp \left(-d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \right) \\ &\stackrel{(ii)}{<} \infty, \end{aligned}$$

where (i) follows because $\ln(1 - x) \leq -x$ for all $x < 1$ and $\exp(x)$ is an increasing function of x for all $x \in \mathbb{R}$; and (ii) follows by recalling that $\lim_{\lambda \rightarrow \infty} d_2^\lambda \neq 0$ and $\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \in$

$\mathbb{R} \cup \{\infty\}$ (by assumption) and noting that, even if $\lim_{\lambda \rightarrow \infty} d_2^{\lambda} \frac{k^{\lambda} - N^{\lambda}}{d_1^{\lambda}} = \infty$, exponential decay in terms of $d_2^{\lambda} \frac{k^{\lambda} - N^{\lambda}}{d_1^{\lambda}}$ would dominate its polynomial growth. \blacksquare

A.7.7 Technical Details for Footnote 7

When $\mu = a$ and $\lim_{\lambda \rightarrow \infty} N^{\lambda} - \frac{\lambda}{a} =: x \in (-\infty, 1)$, we exhibit a counterexample under which $\lim_{\lambda \rightarrow \infty} \frac{\partial^2 U^{\lambda}(\mu_1, \mu)}{\partial \mu_1^2} = \infty$ for $\mu_1 = a(1 - x)$ for a range of values of the parameters a, p , and v . We inherit the shorthand notation and the setup outlined around the proof of Proposition 16 in the preliminary Section A.7.2.

Consider the sequence of staffing levels $N^{\lambda} = \frac{\lambda}{a} - \frac{1}{\lambda^2}$ and a corresponding sequence of system sizes $k^{\lambda} = N^{\lambda} + \lambda^2$. Recalling that $\rho^{\lambda} = \frac{\mu_1}{a}$ when $\mu = a$, we begin by presenting the following observations which form the building blocks of our argument:

- $0 < \frac{\lambda}{a} - N^{\lambda} \in o(1)$, implying that $\lim_{\lambda \rightarrow \infty} C^{\lambda} = 1$ (from Lemma 29 (b)), $\lim_{\lambda \rightarrow \infty} I^{\lambda} = 0$ (from Proposition 16 when $\mu = a$), and $\lim_{\lambda \rightarrow \infty} \frac{\partial I^{\lambda}}{\partial \mu_1} = 0$ (from Proposition 17 when $\mu = a$);
- $x := \lim_{\lambda \rightarrow \infty} N^{\lambda} - \frac{\lambda}{a} = 0$;
- $\mu_1 = a(1 - x) = a$;
- $d_1^{\lambda} = N^{\lambda} - \left(1 - \frac{\mu_1}{\mu}\right) = \frac{\lambda}{a} - \frac{1}{\lambda^2}$;
- $d_2^{\lambda} = d_1^{\lambda} - \frac{\lambda}{a} = -\frac{1}{\lambda^2}$;
- $d_2^{\lambda} \frac{k^{\lambda} - N^{\lambda}}{d_1^{\lambda}} = -\frac{1}{\frac{\lambda}{a} - \frac{1}{\lambda^2}} = -\frac{1}{\rho^{\lambda} - \frac{1}{\lambda^2}}$, implying that $\lim_{\lambda \rightarrow \infty} \rho^{\lambda} d_2^{\lambda} \frac{k^{\lambda} - N^{\lambda}}{d_1^{\lambda}} = -1$.

From (A.88), when $\mu = a$,

$$\begin{aligned} \lim_{\lambda \rightarrow \infty} \frac{I^\lambda}{d_2^\lambda} &= \lim_{\lambda \rightarrow \infty} \left[d_2^\lambda + \frac{\mu}{\mu_1} \rho^\lambda d_2^\lambda \left(\frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \right) + \frac{\mu}{\mu_1} C^\lambda \rho^\lambda \left(1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right) \right]^{-1} \\ &= \lim_{\lambda \rightarrow \infty} \left[d_2^\lambda + \frac{\mu}{\mu_1} \sqrt{\rho^\lambda} d_2^\lambda \left(\sqrt{\rho^\lambda} \left(\frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \right) \right) + \frac{\mu}{\mu_1} C^\lambda \rho^\lambda d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} \left(\frac{1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda}}{d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} \right) \right]^{-1} = -1, \end{aligned} \quad (\text{A.136})$$

because $\lim_{\lambda \rightarrow \infty} d_2^\lambda = 0$, $\lim_{\lambda \rightarrow \infty} \sqrt{\rho^\lambda} d_2^\lambda = 0$, $\lim_{\lambda \rightarrow \infty} \sqrt{\rho^\lambda} \frac{1 - C^\lambda}{N^\lambda - \rho^\lambda} \in (0, \infty)$ (from Lemma 30 (c)), $\lim_{\lambda \rightarrow \infty} C^\lambda = 1$, $\lim_{\lambda \rightarrow \infty} \rho^\lambda d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda} = -1$, and $\lim_{\lambda \rightarrow \infty} \frac{1 - \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda}}{d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} = 1$ (from Lemma 33 (ii)).

From (A.90) in the proof of Proposition 17,

$$\begin{aligned} &\lim_{\lambda \rightarrow \infty} \frac{\mu_1^2}{I^\lambda} \left(\frac{\partial I^\lambda}{\partial \mu_1} \right)^2 \\ &= \lim_{\lambda \rightarrow \infty} I^\lambda \left(1 - I^\lambda + \rho^\lambda C^\lambda I^\lambda \left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^2 \right)^2 \end{aligned} \quad (\text{A.137})$$

$$\begin{aligned} &= \lim_{\lambda \rightarrow \infty} I^\lambda (1 - I^\lambda)^2 + 2(I^\lambda)^2 (1 - I^\lambda) \rho^\lambda C^\lambda \left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^2 + (\rho^\lambda)^2 (C^\lambda)^2 (I^\lambda)^3 \left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^4 \\ &= \lim_{\lambda \rightarrow \infty} I^\lambda (1 - I^\lambda)^2 + 2(1 - I^\lambda) C^\lambda \left(\sqrt{\rho^\lambda} \frac{k^\lambda - N^\lambda}{d_1^\lambda} I^\lambda \right)^2 + (C^\lambda)^2 \left(\left(\frac{I^\lambda}{d_2^\lambda} \right)^3 (d_2^\lambda)^3 (\rho^\lambda)^2 \left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^4 \right) \stackrel{(\ddagger)}{=} a^2, \end{aligned} \quad (\text{A.138})$$

where (\ddagger) follows by recalling that $\lim_{\lambda \rightarrow \infty} I^\lambda = 0$, $\lim_{\lambda \rightarrow \infty} C^\lambda = 1$, $\lim_{\lambda \rightarrow \infty} \frac{I^\lambda}{d_2^\lambda} = -1$ (from (A.136)), $\lim_{\lambda \rightarrow \infty} (d_2^\lambda)^3 (\rho^\lambda)^2 \left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^4 = \lim_{\lambda \rightarrow \infty} \left(-\frac{1}{\lambda^2} \right)^3 \left(\frac{\lambda}{a} \right)^2 \left(\frac{\lambda^2}{\frac{\lambda}{a} - \frac{1}{\lambda^2}} \right)^4 = -a^2$, and $\lim_{\lambda \rightarrow \infty} \sqrt{\rho^\lambda} \frac{k^\lambda - N^\lambda}{d_1^\lambda} I^\lambda = 0$ (from (A.92) in the proof of Proposition 17).

Next, from (A.133) and recalling that $\mu_1 = a$,

$$\begin{aligned} \lim_{\lambda \rightarrow \infty} \frac{\partial^2 I^\lambda}{\partial \mu_1^2} &= \lim_{\lambda \rightarrow \infty} \left(-\frac{2}{\mu_1} I^\lambda \frac{\partial I^\lambda}{\partial \mu_1} - 2 \left[\frac{1}{\mu_1} \frac{\partial I^\lambda}{\partial \mu_1} - \frac{1}{I^\lambda} \left(\frac{\partial I^\lambda}{\partial \mu_1} \right)^2 - \frac{1}{\mu_1} I^\lambda \frac{\partial I^\lambda}{\partial \mu_1} - \frac{I^\lambda}{\mu_1 a d_2^\lambda} + \frac{1}{ad_2^\lambda} \frac{\partial I^\lambda}{\partial \mu_1} + \frac{(I^\lambda)^2}{\mu_1 a d_2^\lambda} \right] \right. \\ &\quad \left. + \frac{1}{\mu_1 a} \left(\frac{\rho^\lambda}{d_2^\lambda} I^\lambda \right) \frac{(k^\lambda - N^\lambda)(k^\lambda - N^\lambda + 1)}{(d_1^\lambda)^2} C^\lambda \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} I^\lambda \right) \end{aligned}$$

$$\begin{aligned}
&= 2 \left(1 - \frac{1}{a^2} \right) + \frac{1}{a} \lim_{\lambda \rightarrow \infty} \left(-2 \frac{I^\lambda}{d_2^\lambda} \frac{1}{I^\lambda} \frac{\partial I^\lambda}{\partial \mu_1} + \frac{1}{a} \left(\frac{\rho^\lambda}{d_2^\lambda} I^\lambda \right) \frac{(k^\lambda - N^\lambda)(k^\lambda - N^\lambda + 1)}{(d_1^\lambda)^2} C^\lambda \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} I^\lambda \right) \\
&\stackrel{(*)}{=} 2 \left(1 - \frac{1}{a^2} \right) + \frac{1}{a} \lim_{\lambda \rightarrow \infty} \left(2 \left[1 - I^\lambda + \frac{I^\lambda}{d_2^\lambda} C^\lambda \rho^\lambda d_2^\lambda \left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^2 \right] \right. \\
&\quad \left. + \frac{1}{a} \left(\frac{\rho^\lambda}{d_2^\lambda} I^\lambda \right) \frac{(k^\lambda - N^\lambda)(k^\lambda - N^\lambda + 1)}{(d_1^\lambda)^2} C^\lambda \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} I^\lambda \right) \\
&= 2 \left(1 - \frac{1}{a^2} \right) + \frac{1}{a^2} \lim_{\lambda \rightarrow \infty} \left(2a(1 - I^\lambda) + \frac{I^\lambda}{d_2^\lambda} \left(2a + \frac{I^\lambda}{d_2^\lambda} \frac{k^\lambda - N^\lambda + 1}{k^\lambda - N^\lambda} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} \right) C^\lambda \rho^\lambda d_2^\lambda \left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^2 \right) \\
&\stackrel{(**)}{=} 2 \left(1 - \frac{1}{a^2} \right) + \frac{1}{a^2} \{2a - (2a - 1)(-\infty)\},
\end{aligned}$$

where $(*)$ follows from (A.90) in the proof of Proposition 17 and $(**)$ follows by recalling that $\lim_{\lambda \rightarrow \infty} I^\lambda = 0$, $\lim_{\lambda \rightarrow \infty} \frac{I^\lambda}{d_2^\lambda} = -1$ (from (A.136)), $\lim_{\lambda \rightarrow \infty} \frac{k^\lambda - N^\lambda + 1}{k^\lambda - N^\lambda} = 1$ (since $k^\lambda - N^\lambda = \lambda^2$), $\lim_{\lambda \rightarrow \infty} \left(\frac{\rho^\lambda}{d_1^\lambda} \right)^{k^\lambda - N^\lambda} = e^{-\lim_{\lambda \rightarrow \infty} d_2^\lambda \frac{k^\lambda - N^\lambda}{d_1^\lambda}} = 1$ (from Lemma 33 (i)), $\lim_{\lambda \rightarrow \infty} C^\lambda = 1$, and $\lim_{\lambda \rightarrow \infty} \rho^\lambda d_2^\lambda \left(\frac{k^\lambda - N^\lambda}{d_1^\lambda} \right)^2 = \lim_{\lambda \rightarrow \infty} \frac{\lambda}{a} \left(-\frac{1}{\lambda^2} \right) \left(\frac{\lambda^2}{\frac{\lambda}{a} - \frac{1}{\lambda^2}} \right)^2 = -\infty$. Hence, the above display implies that

$$\lim_{\lambda \rightarrow \infty} \frac{\partial^2 I^\lambda}{\partial \mu_1^2} = \begin{cases} -\infty, & \text{if } a < \frac{1}{2}, \\ +\infty, & \text{if } a > \frac{1}{2}. \end{cases}$$

Then, from (A.125), when $\mu_1 = a$ and recalling that $\lim_{\lambda \rightarrow \infty} \frac{\partial I^\lambda}{\partial \mu_1} = 0$, it follows that, if $p > 2ap > 2v$ or $p < 2ap < 2v$, then

$$\lim_{\lambda \rightarrow \infty} \frac{\partial^2 U^\lambda}{\partial \mu_1^2} = (v - ap) \frac{\partial^2 I^\lambda}{\partial \mu_1^2} - 2p \frac{\partial I^\lambda}{\partial \mu_1} - c''(a) = \infty.$$

■

A.7.8 Proof of Proposition 5

From Lemma 6, under the same restrictions on the staffing rules, the utility function $U(\mu_1, \mu)$ is a concave function of μ_1 for all $\mu > 0$, which implies that there exists a unique maximum.

Thus, any solution to the FOC (1.6) (i.e., candidate server equilibrium) $\mu^{*\dagger} > 0$ is a global maximizer of $U(\mu_1, \mu^{*\dagger})$, that is, candidate server equilibrium $\mu^{*\dagger}$ is actually a sever equilibrium. Therefore, $\mu^{*\dagger} > 0$ is a server equilibrium if and only if it satisfies the FOC (1.6). ■

A.7.9 Proof of Theorem 4

First, we state three lemmas concerning underloaded equilibria (whose proofs appear later) that help simplify the proof of Theorem 4. Recall the definition of $p^\dagger(v)$ in Theorem 4 (also in (A.112)).

Lemma 47. *If $0 \leq p < \min\{c'(a), p^\dagger(v)\}$, then there exists a unique $\bar{a}(p, v) > 0$ such that $n_u = 0, 1$, or 2 according to whether $a > \bar{a}(p, v)$, $a = \bar{a}(p, v)$, or $a < \bar{a}(p, v)$, respectively.*

Lemma 48. *If $p > c'(a)$ or $p = c'(a) < p^\dagger(v)$, then $n_u = 1$.*

Lemma 49. *If $p^\dagger(v) \leq p \leq c'(a)$, then $n_u = 0$.*

Next, we note that an overloaded or critically loaded equilibrium must lie in $(0, a]$, and for this interval, the limiting FOC (1.16) reduces to $c'(\mu) = p$. We are now ready to prove Theorem 4.

(a) $\mathbf{p} \leq \mathbf{c}'(0)$: Here, $n_o = n_c = 0$ because $c'(\mu) > c'(0) \geq p$ for all $\mu \in (0, a]$, recalling that c' is strictly increasing (by strict convexity of c). For underloaded equilibria, we first note that $p \leq c'(0)$ implies $p < c'(a)$. Furthermore, $p^\dagger(v) > c'(0)$ by definition (see (A.112)), so, $p \leq c'(0)$ also implies $p < p^\dagger(v)$. Together, we have $p < \min\{c'(a), p^\dagger(v)\}$, so, the result follows from Lemma 47.

(b) $\mathbf{p} > \mathbf{c}'(0)$: In this case, note that $(c')^{-1}(p) > 0$ is well-defined, since c' is strictly increasing.

(i): If $a > (c')^{-1}(p)$, then $p < c'(a)$. Thus, $c'(0) < p < c'(a)$. Recalling that c' is strictly increasing, it follows that $c'(\mu) = p$ must admit a unique solution in $(0, a)$, i.e., $n_o = 1$.

On the other hand, if $a \leq (c')^{-1}(p)$, then $c'(\mu) < c'(a) \leq p$ for all $\mu \in (0, a)$. Therefore, $c'(\mu) = p$ has no solution in $(0, a)$, i.e., $n_o = 0$.

(ii): If $a = (c')^{-1}(p)$, then $c'(a) = p$, so $\mu = a$ is a solution to $c'(\mu) = p$, i.e., $n_c = 1$.

Otherwise, $\mu = a$ is not a solution to $c'(\mu) = p$, i.e., $n_c = 0$.

(iii): If $p < p^\dagger(v)$, the result follows from combining the following two subcases:

(iii-1): If $a \leq (c')^{-1}(p)$, Lemma 48 implies that $n_u = 1$.

(iii-2): If $a > (c')^{-1}(p)$, then $p < c'(a)$, so, from Lemma 47, there exists a unique $\bar{a}(p, v) > 0$ such that $n_u = 0, 1$, or 2 according to whether $a > \bar{a}(p, v)$, $a = \bar{a}(p, v)$, or $a < \bar{a}(p, v)$, respectively.

(iv): If $p \geq p^\dagger(v)$.

(iv-1): If $a < (c')^{-1}(p)$, then Lemma 48 implies that $n_u = 1$.

(iv-2): If $a \geq (c')^{-1}(p)$, then Lemma 49 implies that $n_u = 0$.

■

We now set up the necessary tools for the proofs of Lemmas 47, 48, and 49. Before proceeding, we note that, throughout these proofs, we heavily reference the definitions and results from the preliminary Sections A.7.3 and A.7.4. The reader might therefore find it useful to review these sections first.

To begin, recall, from (A.109), that the limiting FOC (1.16) can be equivalently written as:

$$c'(\mu) = \begin{cases} p, & \mu < a, \\ h(\mu; a, p, v), & \mu \geq a, \end{cases} \quad (\text{A.139})$$

where $h(\mu; a, p, v) = \frac{a^2}{\mu^2} \left(p + \frac{v}{a} - \frac{v}{\mu} \right)$ is well-defined for all $\mu > 0$. It follows from Lemma 40 (a)(b) and the fact that $\frac{3v}{2p+\frac{2v}{a}} < \frac{2v}{p+\frac{v}{a}}$ for any $a > 0$, $p \geq 0$, and $v > 0$; that $h(\mu)$ is strictly in-

creasing and strictly concave in $\mu \in \left(0, \frac{3v}{2p+\frac{2v}{a}}\right)$; strictly decreasing and strictly concave in $\mu \in \left(\frac{3v}{2p+\frac{2v}{a}}, \frac{2v}{p+\frac{v}{a}}\right)$; and then strictly decreasing and strictly convex in $\mu \in \left(\frac{2v}{p+\frac{v}{a}}, \infty\right)$. It is useful to investigate how these properties of $h(\mu)$ affect the behavior of the right-hand side of (A.139). This is because, in order to study the number of underloaded equilibria, it is important to understand how the right-hand side of (A.139) interacts with $c'(\mu)$, a strictly convex and strictly increasing function, in the interval (a, ∞) . We consider the following two scenarios, illustrated in Figure A.6:

- When $p < \frac{v}{2a}$, it is easy to see that $\frac{3v}{2p+\frac{2v}{a}} > a$, implying that $h(\mu)$ is strictly increasing and strictly concave in $\mu \in \left(a, \frac{3v}{2p+\frac{2v}{a}}\right)$, strictly decreasing and strictly concave in $\mu \in \left(\frac{3v}{2p+\frac{2v}{a}}, \frac{2v}{p+\frac{v}{a}}\right)$, and then strictly decreasing and strictly convex in $\mu \in \left(\frac{2v}{p+\frac{v}{a}}, \infty\right)$, as shown in Figure A.6 (I).
- When $p \geq \frac{v}{2a}$, it is easy to see that $\frac{3v}{2p+\frac{2v}{a}} \leq a$, implying that $h(\mu)$ is strictly decreasing in $\mu \in (a, \infty)$, as shown in Figure A.6 (II).

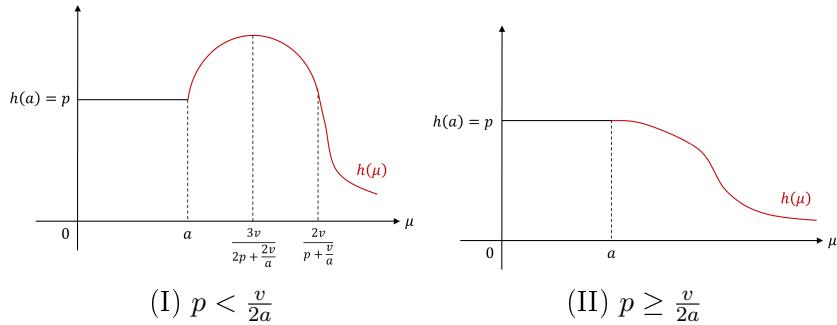


Figure A.6: Illustration of the right-hand side (solid black and solid red curves) of the limiting FOC (A.139).

To simplify the proof of Lemmas 47 and 48, we leverage the following generic result.

Lemma 50. *Suppose $f(x)$ is a convex function of $x \in (0, \bar{x}]$. For any $0 < x_1 < x_2 \leq \bar{x}$,*

- If $f(x_1)f(x_2) < 0$, then $f(x) = 0$ is obtained at exactly one $x \in (x_1, x_2)$.*
- If $f(x_1) < 0$ and $f(x_2) < 0$, then $f(x) < 0$ for all $x \in (x_1, x_2)$.*

A.7.9.1 Proof of Lemma 50

(a): Since $f(x_1)f(x_2) < 0$, the intermediate value theorem implies that $f(x) = 0$ is achieved at at least one $x \in (x_1, x_2)$. We prove uniqueness by contradiction. Suppose there exist two distinct x_0, x'_0 such that $x_1 < x_0 < x'_0 < x_2$ and $f(x_0) = f(x'_0) = 0$. Since $x_1 < x_0 < x'_0$, there exists $\alpha_0 \in (0, 1)$ such that $x_0 = \alpha_0 x_1 + (1 - \alpha_0)x'_0$. Similarly, since $x_0 < x'_0 < x_2$, there exists $\alpha'_0 \in (0, 1)$ such that $x'_0 = \alpha'_0 x_0 + (1 - \alpha'_0)x_2$.

Case (I): Suppose $f(x_1) < 0$ and $f(x_2) > 0$. Since f is a convex function, we have $\alpha_0 f(x_1) + (1 - \alpha_0)f(x'_0) \geq f(\alpha_0 x_1 + (1 - \alpha_0)x'_0) = f(x_0)$, which implies that $f(x_1) \geq 0$ (recalling from the assumption that $f(x_0) = f(x'_0) = 0$), a contradiction.

Case (II): Suppose $f(x_1) > 0$ and $f(x_2) < 0$. Since f is a convex function, we have $\alpha'_0 f(x_0) + (1 - \alpha'_0)f(x_2) \geq f(\alpha'_0 x_0 + (1 - \alpha'_0)x_2) = f(x'_0)$, which implies that $f(x_2) \geq 0$ (recalling from the assumption that $f(x_0) = f(x'_0) = 0$), a contradiction.

(b): Note that $x = \frac{x_2 - x}{x_2 - x_1}x_1 + \left(1 - \frac{x_2 - x}{x_2 - x_1}\right)x_2$, then since f is a convex function, it follows that $f(x) = f\left(\frac{x_2 - x}{x_2 - x_1}x_1 + \left(1 - \frac{x_2 - x}{x_2 - x_1}\right)x_2\right) \leq \frac{x_2 - x}{x_2 - x_1}f(x_1) + \left(1 - \frac{x_2 - x}{x_2 - x_1}\right)f(x_2) < 0$.

■

Now, we are ready to prove Lemmas 47, 48, and 49.

A.7.9.2 Proof of Lemma 47

Suppose $0 < p < \min\{c'(a), p^\dagger(v)\}$. Since $p < p^\dagger(v)$, due to Remark 20, we can divide the interval $a \in (\bar{a}(p, v), \infty)$ into two subintervals $a \in (\bar{a}(p, v), \frac{v}{2p})$ and $a \in [\frac{v}{2p}, \infty)$. In the latter subinterval, $p \geq \frac{v}{2a}$, and so $h(\mu)$ is strictly decreasing in $\mu \in (a, \infty)$ (recall Figure A.6 (II)); then, since $p < c'(a)$, we have $c'(\mu) > c'(a) > p = h(a) > h(\mu)$ for all $\mu > a$, resulting in no underloaded equilibria, i.e., $n_u = 0$. Therefore, for the remainder of the proof, we need only focus on the case of $0 \leq p < \frac{v}{2a}$ (recall Figure A.6 (I)) in studying underloaded equilibria.

For any $p < c'(a)$, Lemma 46 states that there exists a unique $v > 0$ such that $p < p^\dagger(v)$, given by $v^\dagger(a, p)$. Recall the definitions of $\mu^\dagger(a, p)$ and $v^\dagger(a, p)$ from Definition 15, $p^\dagger(v)$ from Definition 16, and $\bar{a}(p, v)$ from Definition 17.

When $p < p^\dagger(v)$ for all $v > 0$, by Corollary 6, $a > \bar{a}(p, v)$, $a = \bar{a}(p, v)$ and $a < \bar{a}(p, v)$ are equivalent to $v < v^\dagger(a, p)$, $v = v^\dagger(a, p)$ and $v > v^\dagger(a, p)$, respectively. In the remainder of the proof, we use conditions in terms of $v^\dagger(a, p)$ instead of $\bar{a}(p, v)$, for ease of presentation.

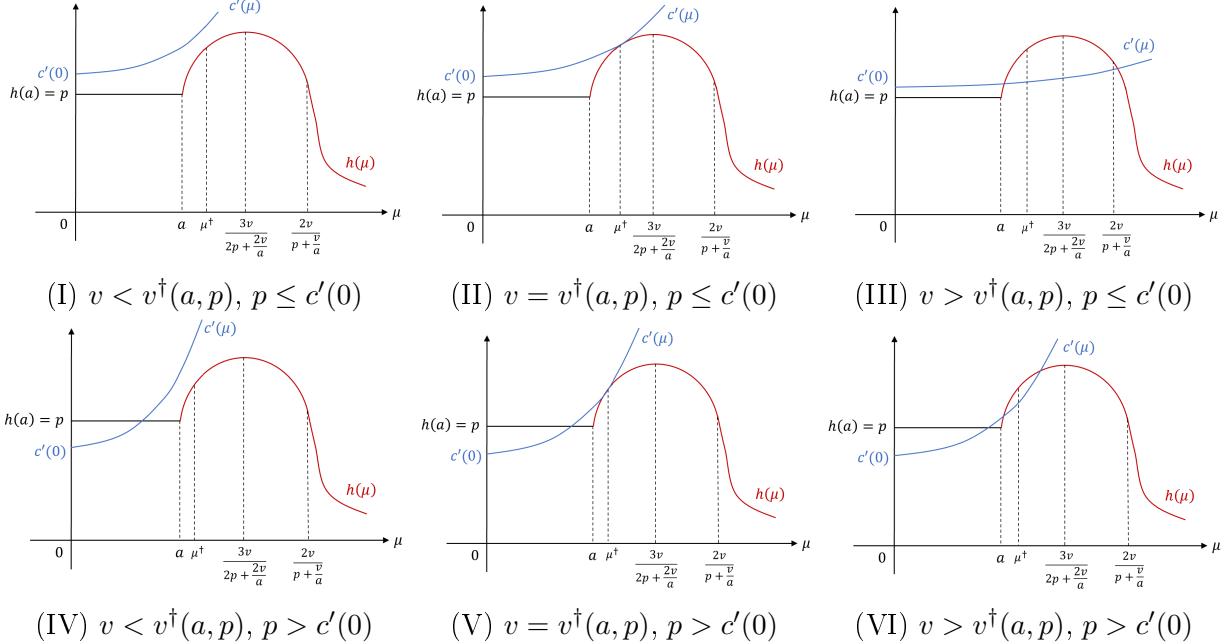


Figure A.7: Illustration of the left-hand side (solid blue curve) and the right-hand side (solid black and solid red curves) of the limiting FOC (A.139) when $p < \frac{v}{2a}$ and $0 \leq p < \min \{c'(a), p^\dagger(v)\}$.

- $v < v^\dagger(a, p)$ (Figures A.7 (I)(IV)): First, recall from Lemma 40 (e) that $h(\mu; a, p, v)$ is strictly increasing in v for all $\mu > a$. Therefore, when $v < v^\dagger(a, p)$, $h(\mu; a, p, v) < h(\mu; a, p, v^\dagger(a, p))$ for all $\mu > a$. Moreover, from Lemma 43, when $p < c'(a)$, $h(\mu; a, p, v^\dagger(a, p)) \leq c'(\mu)$ for all $\mu > a$. Thus, it follows that when $v < v^\dagger(a, p)$, $h(\mu; a, p, v) < c'(\mu)$ for all $\mu > a$, i.e., $n_u = 0$.

- $v = v^\dagger(a, p)$ (Figure A.7 (II)(V)): From Lemma 43, when $p < c'(a)$, $h(\mu; a, p, v^\dagger(a, p)) =$

$c'(\mu)$ has exactly one solution in (a, ∞) , namely, $\mu^\dagger(a, p)$, i.e., $n_u = 1$.

- $v > v^\dagger(a, p)$ (Figure A.7 (III)(VI)): First, recall from Lemma 40 (e) that $h(\mu; a, p, v)$ is strictly increasing in v for all $\mu > a$. Therefore, when $v > v^\dagger(a, p)$, $h(\mu; a, p, v) > h(\mu; a, p, v^\dagger(a, p))$ for all $\mu > a$. Thus, $h(\mu^\dagger(a, p); a, p, v) > h(\mu^\dagger(a, p); a, p, v^\dagger(a, p))$ (since $\mu^\dagger(a, p) > a$). Moreover, from Lemma 43, when $p < c'(a)$, $h(\mu^\dagger(a, p); a, p, v^\dagger(a, p)) = c'(\mu^\dagger(a, p))$. Thus, it follows that, when $v > v^\dagger(a, p)$, $h(\mu^\dagger(a, p); a, p, v) > c'(\mu^\dagger(a, p))$. Since $h(a) = p < c'(a)$ and $\lim_{\mu \rightarrow \infty} h(\mu) = 0 < \infty = \lim_{\mu \rightarrow \infty} c'(\mu)$, by the intermediate value theorem, there must exist at least one underloaded equilibrium in $(a, \mu^\dagger(a, p))$ and at least one underloaded equilibrium in $(\mu^\dagger(a, p), \infty)$, i.e., $n_u \geq 2$.

Recalling that $h(\mu)$ is strictly concave in μ for $\mu \in \left(0, \frac{2v}{p+\frac{v}{a}}\right)$ and strictly decreasing in μ for $\mu \in \left[\frac{2v}{p+\frac{v}{a}}, \infty\right)$, we discuss the following two cases.

Case (I): If there exists an underloaded equilibrium in $\left(\frac{2v}{p+\frac{v}{a}}, \infty\right)$ (e.g., Figure A.7 (III)), then it is unique in $\left(\frac{2v}{p+\frac{v}{a}}, \infty\right)$ because $c'(\mu)$ is a strictly increasing function of μ and $h(\mu)$ is a strictly decreasing function of μ in this range. Moreover, $c'\left(\frac{2v}{p+\frac{v}{a}}\right) < h\left(\frac{2v}{p+\frac{v}{a}}\right)$, by the intermediate value theorem and by noting that $\lim_{\mu \rightarrow \infty} c'(\mu) = \infty > 0 = \lim_{\mu \rightarrow \infty} h(\mu)$. Given that $c'\left(\frac{2v}{p+\frac{v}{a}}\right) < h\left(\frac{2v}{p+\frac{v}{a}}\right)$ and $h(a) = p < c'(a)$, which implies that $\left(c'\left(\frac{2v}{p+\frac{v}{a}}\right) - h\left(\frac{2v}{p+\frac{v}{a}}\right)\right)(c'(a) - h(a)) < 0$, together with the fact that $c' - h$ is convex in $\left(a, \frac{2v}{p+\frac{v}{a}}\right)$, Lemma 50 (a) implies that $c'(\mu) - h(\mu) = 0$ at exactly one $\mu \in \left(a, \frac{2v}{p+\frac{v}{a}}\right)$. Hence, $n_u = 2$, where the smaller underloaded equilibrium lies in $(a, \mu^\dagger(a, p))$.

Case (II): If all the underloaded equilibria lie in $\left(a, \frac{2v}{p+\frac{v}{a}}\right]$ (e.g., Figure A.7 (VI)), then there are exactly two underloaded equilibria, because $c'(\mu)$, a strictly convex function and $h(\mu)$, a strictly concave function cannot intersect more than twice, i.e., $n_u = 2$, where the smaller underloaded equilibrium lies in $(a, \mu^\dagger(a, p))$ and the larger underloaded equilibrium lies in $\left(\mu^\dagger(a, p), \frac{2v}{p+\frac{v}{a}}\right]$.

Therefore, $n_u = 0$ if $a > \bar{a}(p, v)$; $n_u = 1$ if $a = \bar{a}(p, v)$; and $n_u = 2$ if $a < \bar{a}(p, v)$. ■

A.7.9.3 Proof of Lemma 48

We address the cases $p > c'(a)$ and $p = c'(a) < p^\dagger(v)$ separately.

Case (I): When $p > c'(a)$, we consider two further subcases $p \geq \frac{v}{2a}$ and $p < \frac{v}{2a}$ separately.

Case (I-1): If $p \geq \frac{v}{2a}$ (Figure A.8), then $h(\mu)$ is strictly decreasing in $\mu \in (a, \infty)$. Since $h(a) = p > c'(a)$, and $\lim_{\mu \rightarrow \infty} h(\mu) = 0 < \infty = \lim_{\mu \rightarrow \infty} c'(\mu)$, together with the fact that h is strictly decreasing and c' is strictly increasing, it follows that there exists exactly one intersection point in (a, ∞) , i.e., $n_u = 1$.

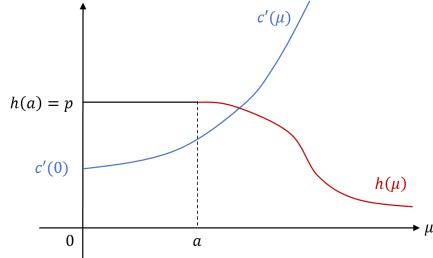


Figure A.8: Illustration of the left-hand side (solid blue curve) and the right-hand side (solid black and solid red curves) of the limiting FOC (A.139) when $p \geq \frac{v}{2a}$ and $p > c'(a)$.

Case (I-2): If $p < \frac{v}{2a}$ (Figure A.9), since $h(a) = p > c'(a)$ and $\lim_{\mu \rightarrow \infty} h(\mu) = 0 < \infty = \lim_{\mu \rightarrow \infty} c'(\mu)$, there must exist at least one underloaded equilibria in (a, ∞) , i.e., $n_u \geq 1$, by the intermediate value theorem. Recalling that $h(\mu)$ is strictly concave in μ for $\mu \in (0, \frac{2v}{p+\frac{v}{a}})$ and strictly decreasing in μ for $\mu \in [\frac{2v}{p+\frac{v}{a}}, \infty)$, we discuss the following two cases.

Case (I-2-a): If there exists an underloaded equilibrium in $(\frac{2v}{p+\frac{v}{a}}, \infty)$ (Figure A.9 (I)), then it is unique in $(\frac{2v}{p+\frac{v}{a}}, \infty)$ because $c'(\mu)$ is a strictly increasing function of μ and $h(\mu)$ is a strictly decreasing function of μ in this range. Moreover, $c'(\frac{2v}{p+\frac{v}{a}}) < h(\frac{2v}{p+\frac{v}{a}})$, because, otherwise, $c'(\mu)$ and $h(\mu)$ would not intersect in $(\frac{2v}{p+\frac{v}{a}}, \infty)$, which contradicts the assumption. Then, Lemma 50 (b) implies no equilibrium in $(a, \frac{2v}{p+\frac{v}{a}}]$, given that $c'(a) - h(a) < 0$, $c'(\frac{2v}{p+\frac{v}{a}}) - h(\frac{2v}{p+\frac{v}{a}}) < 0$ and the fact that $c' - h$ is convex in $(a, \frac{2v}{p+\frac{v}{a}})$. Therefore, $n_u = 1$.

Case (I-2-b): If all the underloaded equilibria lie in $\left(a, \frac{2v}{p+\frac{v}{a}}\right]$ (Figure A.9 (II)), then $c'\left(\frac{2v}{p+\frac{v}{a}}\right) \geq h\left(\frac{2v}{p+\frac{v}{a}}\right)$, because, otherwise, by the intermediate value theorem and the fact that $\lim_{\mu \rightarrow \infty} c'(\mu) = \infty > 0 = \lim_{\mu \rightarrow \infty} h(\mu)$, there would exist at least one underloaded equilibrium in $\left(\frac{2v}{p+\frac{v}{a}}, \infty\right)$, which contradicts the assumption. Since $c'\left(\frac{2v}{p+\frac{v}{a}}\right) \geq h\left(\frac{2v}{p+\frac{v}{a}}\right)$ and $h'\left(\frac{2v}{p+\frac{v}{a}}\right) < 0 < c''\left(\frac{2v}{p+\frac{v}{a}}\right)$, it follows that $c'\left(\frac{2v}{p+\frac{v}{a}} + \epsilon\right) > h\left(\frac{2v}{p+\frac{v}{a}} + \epsilon\right)$ for all small enough $\epsilon > 0$. Then, given $h(a) = p > c'(a)$ and $c'\left(\frac{2v}{p+\frac{v}{a}} + \epsilon\right) > h\left(\frac{2v}{p+\frac{v}{a}} + \epsilon\right)$ for all small enough $\epsilon > 0$, which implies that $(c'\left(\frac{2v}{p+\frac{v}{a}} + \epsilon\right) - h\left(\frac{2v}{p+\frac{v}{a}} + \epsilon\right)) (c'(a) - h(a)) < 0$, together with the fact that $c' - h$ is convex in $\left(a, \frac{2v}{p+\frac{v}{a}}\right)$, Lemma 50 (a) implies that $c'(\mu) - h(\mu) = 0$ at exactly one $\mu \in \left(a, \frac{2v}{p+\frac{v}{a}} + \epsilon\right)$ for all small enough $\epsilon > 0$. Hence, $n_u = 1$.

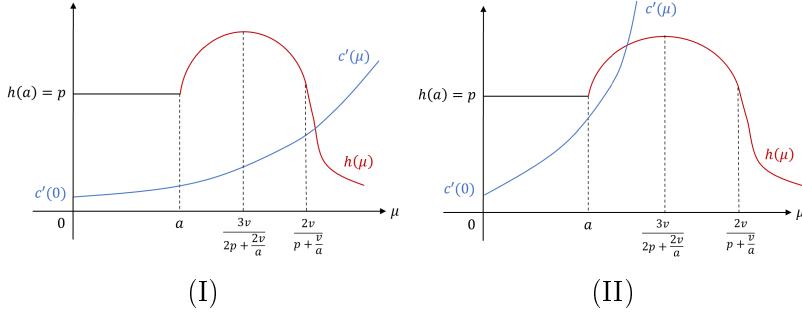


Figure A.9: Illustration of the left-hand side (solid blue curve) and the right-hand side (solid black and solid red curves) of the limiting FOC (A.139) when $p < \frac{v}{2a}$ and $p > c'(a)$.

Case (II): If $p = c'(a) < p^\dagger(v)$ (Figure A.10), then, recalling the definition of $p^\dagger(v)$ in (A.112), $p < p^\dagger(v)$ becomes

$$\begin{aligned} p(c')^{-1}(p) + \frac{1}{2} ((c')^{-1}(p))^2 c''((c')^{-1}(p)) &< \frac{v}{2} \\ \Rightarrow \quad pa + \frac{1}{2} a^2 c''(a) &< \frac{v}{2} \\ \Rightarrow \quad c''(a) &< -\frac{2p}{a} + \frac{v}{a^2} \stackrel{(*)}{=} h'(a), \end{aligned}$$

where $(*)$ follows by substituting for $\mu = a$ into $h'(\mu) = -\frac{2a^2}{\mu^3} (p + \frac{v}{a}) + \frac{3a^2v}{\mu^4}$. Given $c'(a) = h(a)$ and $c''(a) < h'(a)$, $c'(a + \epsilon) < h(a + \epsilon)$ for all small enough $\epsilon > 0$. Then, the same proof from Case (I-2) is applicable here, yielding a unique intersection of $c'(\mu)$ and

$h(\mu)$ in $(a + \epsilon, \infty)$ for all small enough $\epsilon > 0$, i.e., $n_u = 1$.

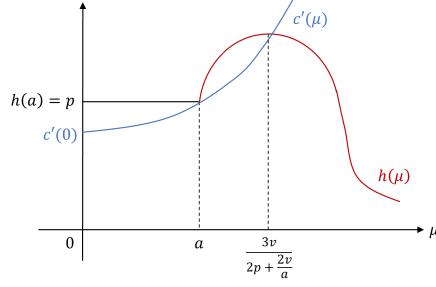


Figure A.10: Illustration of the left-hand side (solid blue curve) and the right-hand side (solid black and solid red curves) of the limiting FOC (A.139) when $c'(0) < p < p^\dagger(v)$ and $a = (c')^{-1}(p)$. ■

A.7.9.4 Proof of Lemma 49

If $p^\dagger(v) \leq p \leq c'(a)$, then, recalling the definition of $p^\dagger(v)$ in (A.112), $p \geq p^\dagger(v)$ implies

$$p(c')^{-1}(p) + \frac{1}{2} ((c')^{-1}(p))^2 c''((c')^{-1}(p)) \geq \frac{v}{2},$$

which implies

$$pa + \frac{1}{2}a^2c''(a) \geq p(c')^{-1}(p) + \frac{1}{2} ((c')^{-1}(p))^2 c''((c')^{-1}(p)) \geq \frac{v}{2},$$

This above inequality implies that

$$\begin{aligned} pa + \frac{1}{2}a^2c''(a) &\geq \frac{v}{2} \\ \Rightarrow c''(a) &\geq -\frac{2p}{a} + \frac{v}{a^2} \stackrel{(*)}{=} h'(a), \end{aligned}$$

where $(*)$ follows by substituting for $\mu = a$ into $h'(\mu) = -\frac{2a^2}{\mu^3} (p + \frac{v}{a}) + \frac{3a^2v}{\mu^4}$. Thus, $c''(a) \geq h'(a) \geq h'(\mu)$, for all $\mu \geq a$, recalling that c' is strictly increasing, and $h(\mu)$ is either (a) strictly decreasing in $\mu \in (a, \infty)$ (Figure A.11 (I)), or (b) strictly increasing and strictly

concave on $\mu \in \left(0, \frac{3v}{2p+\frac{2v}{a}}\right)$, and strictly decreasing on $\mu \in \left(\frac{3v}{2p+\frac{2v}{a}}, \infty\right)$ (Figure A.11 (II)). Together $c''(\mu) \geq h'(\mu)$ for all $\mu \geq a$, and the fact that $c'(a) \geq p = h(a)$, it follows that $c'(\mu) \geq h(\mu)$ for all $\mu \geq a$, with equality held only possible at $\mu = a$. Thus, $n_u = 0$.

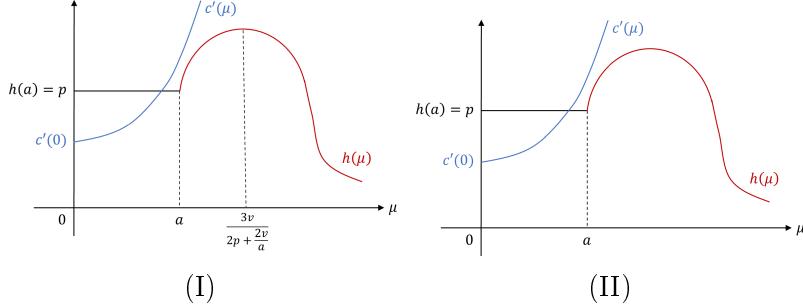


Figure A.11: Illustration of the left-hand side (solid blue curve) and the right-hand side (solid black and solid red curves) of the limiting FOC (A.139) when $p \geq p^\dagger(v)$ and $a \geq (c')^{-1}(p)$. ■

A.7.10 Proof of Proposition 6

Suppose μ_1^* , μ_2^* are symmetric limiting equilibrium service rates such that $\mu_1^* > \mu_2^* > 0$.

From (1.15), the limiting utility function is given by

$$U^\infty(\mu, \mu; a, p, v) = \begin{cases} p\mu - c(\mu), & \mu \leq a, \\ pa + v \left(1 - \frac{a}{\mu}\right) - c(\mu), & \mu > a. \end{cases}$$

The limiting equilibrium service rates μ_1^* and μ_2^* satisfy the limiting FOC (1.16):

$$p \left(1 - \left[1 - \frac{a^2}{\mu^2}\right]^+\right) + v \frac{a}{\mu^2} \left[1 - \frac{a}{\mu}\right]^+ = c'(\mu), \text{ for } \mu = \mu_1^*, \mu_2^*. \quad (\text{A.140})$$

Moreover, strict convexity of c implies that

$$c(\mu_1^*) - c(\mu_2^*) \leq (\mu_1^* - \mu_2^*) c'(\mu_1^*) < \mu_1^* c'(\mu_1^*) - \mu_2^* c'(\mu_2^*),$$

where the second inequality follows from $c'(\mu_1^*) > c'(\mu_2^*)$ for $\mu_1^* > \mu_2^*$. Thus,

$$0 < c(\mu_1^*) - c(\mu_2^*) < \mu_1^* c'(\mu_1^*) - \mu_2^* c'(\mu_2^*). \quad (\text{A.141})$$

Recalling from Theorem 4 that there can be at most one overloaded or critically loaded equilibrium, we consider the following two cases.

Case (I): If μ_1^*, μ_2^* satisfy $\mu_1^* \geq a \geq \mu_2^*$, then

$$U^\infty(\mu_1^*, \mu_1^*) = pa + v \left(1 - \frac{a}{\mu_1^*} \right) - c(\mu_1^*) \quad \text{and} \quad U^\infty(\mu_2^*, \mu_2^*) = p\mu_2^* - c(\mu_2^*), \quad (\text{A.142})$$

and, from (A.140),

$$c'(\mu_1^*) = \frac{a}{(\mu_1^*)^2} \left(v \left(1 - \frac{a}{\mu_1^*} \right) + pa \right) \quad \text{and} \quad c'(\mu_2^*) = p. \quad (\text{A.143})$$

Substituting for $c'(\mu_1^*)$ and $c'(\mu_2^*)$ into (A.141) using (A.143):

$$0 < c(\mu_1^*) - c(\mu_2^*) < \frac{a}{\mu_1^*} \left(v \left(1 - \frac{a}{\mu_1^*} \right) + pa \right) - p\mu_2^* < v \left(1 - \frac{a}{\mu_1^*} \right) + pa - p\mu_2^*,$$

where the second inequality follows because $\mu_1^* > a$. Thus,

$$p\mu_2^* - c(\mu_2^*) < pa + v \left(1 - \frac{a}{\mu_1^*} \right) - c(\mu_1^*),$$

which, from (A.142), is equivalent to

$$U^\infty(\mu_2^*, \mu_2^*) < U^\infty(\mu_1^*, \mu_1^*).$$

Case (II): If $\mu_1^* \geq \mu_2^*$ satisfy $\mu_1^* > \mu_2^* \geq a$, then

$$U^\infty(\mu_1^*, \mu_1^*) = pa + v\left(1 - \frac{a}{\mu_1^*}\right) - c(\mu_1^*) \quad \text{and} \quad U^\infty(\mu_2^*, \mu_2^*) = pa + v\left(1 - \frac{a}{\mu_2^*}\right) - c(\mu_2^*), \quad (\text{A.144})$$

and, from (A.140),

$$c'(\mu_1^*) = \frac{a}{(\mu_1^*)^2} \left(v\left(1 - \frac{a}{\mu_1^*}\right) + pa \right) \quad \text{and} \quad c'(\mu_2^*) = \frac{a}{(\mu_2^*)^2} \left(v\left(1 - \frac{a}{\mu_2^*}\right) + pa \right). \quad (\text{A.145})$$

Substituting for $c'(\mu_1^*)$ and $c'(\mu_2^*)$ into (A.141) using (A.145):

$$\begin{aligned} 0 &< c(\mu_1^*) - c(\mu_2^*) && < \frac{a}{\mu_1^*} \left(v\left(1 - \frac{a}{\mu_1^*}\right) + pa \right) - \frac{a}{\mu_2^*} \left(v\left(1 - \frac{a}{\mu_2^*}\right) + pa \right) \\ &&& < \frac{a}{\mu_1^*} \left[\left(v\left(1 - \frac{a}{\mu_1^*}\right) + pa \right) - \left(v\left(1 - \frac{a}{\mu_2^*}\right) + pa \right) \right] \\ &&& < \left(v\left(1 - \frac{a}{\mu_1^*}\right) + pa \right) - \left(v\left(1 - \frac{a}{\mu_2^*}\right) + pa \right), \end{aligned}$$

where the second inequality follows because $\mu_1^* > \mu_2^*$, and the third inequality follows because $\mu_1^* > a$. Thus,

$$pa + v\left(1 - \frac{a}{\mu_2^*}\right) - c(\mu_2^*) < pa + v\left(1 - \frac{a}{\mu_1^*}\right) - c(\mu_1^*),$$

which, from (A.144), is equivalent to

$$U^\infty(\mu_2^*, \mu_2^*) < U^\infty(\mu_1^*, \mu_1^*).$$

■

A.7.11 Proof of Proposition 7

Throughout the proof, we heavily reference the definitions in Section A.7.4. The reader might therefore find it useful to review it first. Fix $v > 0$.

(a) Overloaded: From (1.16), the limiting overloaded equilibrium $\mu_o^*(a, p; v) \in (0, a)$, when it exists, satisfies $c'(\mu) = p$, implying that $\mu_o^*(a, p; v) = (c')^{-1}(p)$, which exists when $c'(0) < p < c'(a)$. Thus, $\mu_o(a, p; v)$ does not depend on a , and is strictly increasing in p , recalling that c' is strictly increasing (by strict convexity of c) and thus $(c')^{-1}$ is strictly increasing.

(b) Underloaded:

Recall from (A.109) that limiting underloaded equilibria, when they exist, satisfy

$$c'(\mu) = h(\mu; a, p, v), \quad (\text{A.146})$$

where $h(\mu; a, p, v) = \frac{a^2}{\mu^2} \left(p + \frac{v}{a} - \frac{v}{\mu} \right)$.

- From Theorem 4 (a) and (b)(iii), when $p < \min\{c'(a), p^\dagger(v)\}$ and $a < \bar{a}(p, v)$ (corresponding to Areas 1,2, and 12 in Figure 1.6), $n_u = 2$. Denote the two underloaded equilibria by $\mu_1^*(a, p; v) > \mu_2^*(a, p; v) > a$. Notice that $p < \frac{v}{2a}$, because $a < \bar{a}(p, v) < \frac{v}{2p}$ (from Remark 20). We assert that $\mu_1^*(a, p; v)$ and $\mu_2^*(a, p; v)$ are continuously differentiable functions of a and p in this region.

Remark 21. Although $\mu_1^*(a, p; v)$ exists when $a = \bar{a}(p, v)$ (from Theorem 4 (b)(iii)), we exclude it because $\mu_1^*(a, p; v)$ is not defined for $a > \bar{a}(p, v)$ and therefore its monotonicity in terms of a at this boundary is meaningless. Also, although $\mu_1^*(a, p; v)$ exists when $p = (\bar{a})^{-1}(a, v)$, we choose to exclude it when discussing its monotonicity in terms of p , noting that $\mu_1^*(a, p; v)$ is not defined for $p < (\bar{a})^{-1}(a, v)$.

Claim 10. For any $p < \min\{c'(a), p^\dagger(v)\}$ and $a < \bar{a}(p, v)$, $a < \mu_2^*(a, p; v) < \mu^\dagger(a, p) < \mu_1^*(a, p; v)$.

- From Theorem 4 (b)(iii)-(iv), when $p \geq c'(a)$ and either (i) $p < p^\dagger(v)$ or (ii) $c'(a) \neq p \geq p^\dagger(v)$ (corresponding to Areas 5 and 51 in Figure 1.6), $n_u = 1$. Denote this unique underloaded equilibrium by $\mu_1^*(a, p; v)$. We assert that $\mu_1^*(a, p; v)$ is a continuously differentiable function of a and p in this region.

Later arguments will require that $\mu_1^*(a, p; v)$ is continuous in a and p across the boundary $a = (c')^{-1}(p)$ for $p \in (c'(0), p^\dagger(v))$ when transitioning between Area 1 and Area 5. Specifically, we show the following claim, whose proof appears later.

Claim 11. Fixing any $a \in (0, \bar{a}(p^\dagger(v), v))$, $\lim_{p \uparrow c'(a)} \mu_1^*(a, p; v) = \mu_1^*(a, c'(a); v)$. Fixing any $p \in (c'(0), p^\dagger(v))$, $\lim_{a \downarrow (c')^{-1}(p)} \mu_1^*(a, p; v) = \mu_1^*((c')^{-1}(p), p; v)$.

We might suppress the dependence of $\mu_1^*(a, p; v)$ and $\mu_2^*(a, p; v)$ on a , p and v henceforth, when the context is clear.

Since $\mu_1^*(a, p; v)$ is the larger or the unique underloaded equilibrium, it must satisfy $h(\mu_1^*) = c'(\mu_1^*)$ and $h'(\mu_1^*) < c''(\mu_1^*)$ (see Figure A.12 (I)(II)), which can be equivalently written as

$$\frac{a^2}{(\mu_1^*)^2} \left(p + \frac{v}{a} - \frac{v}{\mu_1^*} \right) = c'(\mu_1^*) \quad \text{and} \quad \frac{2a^2}{(\mu_1^*)^3} \left(\frac{3v}{2\mu_1^*} - \frac{v}{a} - p \right) < c''(\mu_1^*). \quad (\text{A.147})$$

Since $\mu_2^*(a, p; v)$ is the smaller underloaded equilibrium, it satisfies $h(\mu_2^*) = c'(\mu_2^*)$ and $h'(\mu_2^*) > c''(\mu_2^*)$ (see Figure A.12 (I)), which can be equivalently written as

$$\frac{a^2}{(\mu_2^*)^2} \left(p + \frac{v}{a} - \frac{v}{\mu_2^*} \right) = c'(\mu_2^*) \quad \text{and} \quad \frac{2a^2}{(\mu_2^*)^3} \left(\frac{3v}{2\mu_2^*} - \frac{v}{a} - p \right) > c''(\mu_2^*). \quad (\text{A.148})$$

Before proceeding, the next result is useful for the rest of the proof.

Claim 12. $c''(\mu^*) \neq h'(\mu^*)$ for any limiting underloaded equilibrium $\mu^* > a$.

Monotonicity in p :

(b)(i): Taking the partial derivative with respect to p on both sides of (A.146) and due to

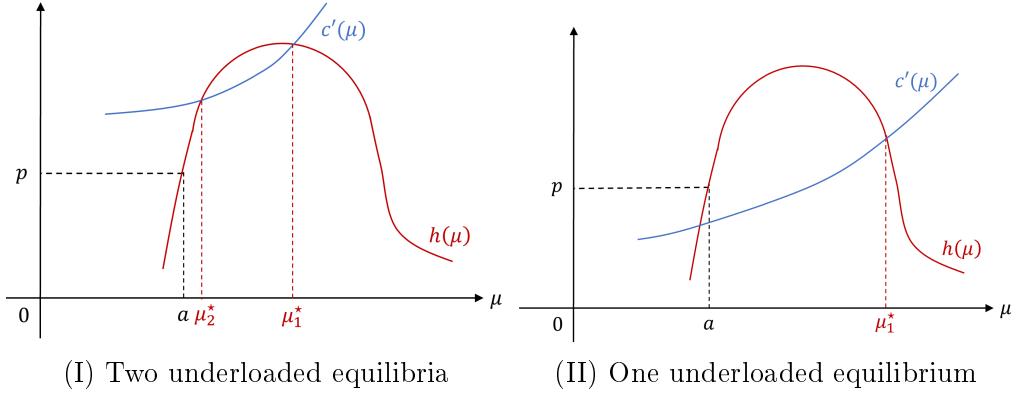


Figure A.12: Illustration of the left-hand side (solid blue curve) and the right-hand side (solid black and solid red curves) of the limiting FOC (A.146) when underloaded equilibria exist.

Claim 12:

$$c''(\mu) \frac{\partial \mu}{\partial p} = h'(\mu) \frac{\partial \mu}{\partial p} + \frac{\partial h}{\partial p} \Rightarrow \frac{\partial \mu}{\partial p} = \frac{\frac{\partial h}{\partial p}}{c''(\mu) - h'(\mu)},$$

where

$$\frac{\partial h}{\partial p} = \frac{a^2}{\mu^2} \quad \text{and} \quad h'(\mu) = \frac{a^2}{\mu^2} \cdot \frac{v}{\mu^2} - \frac{2a^2}{\mu^3} \left(p + \frac{v}{a} - \frac{v}{\mu} \right) = \frac{va^2}{\mu^4} - \frac{2}{\mu} h(\mu) = \frac{va^2}{\mu^4} - \frac{2}{\mu} c'(\mu).$$

Substituting for $\frac{\partial h}{\partial p}$ and $h'(\mu)$ using the above expressions yields

$$\frac{\partial \mu}{\partial p} = \frac{\frac{a^2}{\mu^2}}{c''(\mu) - \left(\frac{va^2}{\mu^4} - \frac{2}{\mu} c'(\mu) \right)} = \frac{1}{\frac{\mu}{a^2} (2c'(\mu) + \mu c''(\mu)) - \frac{v}{\mu^2}},$$

which implies that

$$\mu^2 \left(\frac{\partial \mu}{\partial p} \right)^{-1} = \frac{\mu^3}{a^2} (2c'(\mu) + \mu c''(\mu)) - v. \quad (\text{A.149})$$

Substituting $c'(\mu_1^*)$ and $c''(\mu_1^*)$ using (A.147) into (A.149) yields

$$\begin{aligned} (\mu_1^*)^2 \left(\frac{\partial \mu_1^*(a, p; v)}{\partial p} \right)^{-1} &= \frac{(\mu_1^*)^3}{a^2} (2c'(\mu_1^*) + \mu_1^* c''(\mu_1^*)) - v \\ &> \frac{(\mu_1^*)^3}{a^2} \frac{2a^2}{(\mu_1^*)^2} \left(p + \frac{v}{a} - \frac{v}{\mu_1^*} \right) + \frac{(\mu_1^*)^3}{a^2} \frac{2a^2}{(\mu_1^*)^2} \left(\frac{3v}{2\mu_1^*} - \frac{v}{a} - p \right) - v = 0. \end{aligned}$$

Hence, $\frac{\partial \mu_1^*(a, p; v)}{\partial p} > 0$, i.e., $\mu_1^*(a, p; v)$ is strictly increasing in p .

Similarly, when $\mu_2^*(a, p; v)$ exists, substituting $c'(\mu_2^*)$ and $c''(\mu_2^*)$ using (A.148) into (A.149) yields

$$\begin{aligned} (\mu_2^*)^2 \left(\frac{\partial \mu_2^*(a, p; v)}{\partial p} \right)^{-1} &= \frac{(\mu_2^*)^3}{a^2} (2c'(\mu_2^*) + \mu_2^* c''(\mu_2^*)) - v \\ &< \frac{(\mu_2^*)^3}{a^2} \frac{2a^2}{(\mu_2^*)^2} \left(p + \frac{v}{a} - \frac{v}{\mu_2^*} \right) + \frac{(\mu_2^*)^3}{a^2} \frac{2a^2}{(\mu_2^*)^2} \left(\frac{3v}{2\mu_2^*} - \frac{v}{a} - p \right) - v = 0. \end{aligned}$$

Hence, $\frac{\partial \mu_2^*(a, p; v)}{\partial p} < 0$, i.e., $\mu_2^*(a, p; v)$ is strictly decreasing in p .

(Non-)monotonicity in a :

Taking the partial derivative with respect to a on both sides of (A.146) and due to Claim 12:

$$c''(\mu) \frac{\partial \mu}{\partial a} = h'(\mu) \frac{\partial \mu}{\partial a} + \frac{\partial h}{\partial a} \quad \Rightarrow \quad \frac{\partial \mu}{\partial a} = \frac{\frac{\partial h}{\partial a}}{c''(\mu) - h'(\mu)}, \quad (\text{A.150})$$

where

$$\begin{aligned} \frac{\partial h}{\partial a} &= \frac{a^2}{\mu^2} \left(-\frac{v}{a^2} \right) + \frac{2a}{\mu^2} \left(p + \frac{v}{a} - \frac{v}{\mu} \right) = \frac{2}{a} h(\mu) - \frac{v}{\mu^2} = \frac{2}{a} c'(\mu) - \frac{v}{\mu^2}, \text{ and} \\ h'(\mu) &= \frac{a^2}{\mu^2} \cdot \frac{v}{\mu^2} - \frac{2a^2}{\mu^3} \left(p + \frac{v}{a} - \frac{v}{\mu} \right) = \frac{va^2}{\mu^4} - \frac{2}{\mu} h(\mu) = \frac{va^2}{\mu^4} - \frac{2}{\mu} c'(\mu). \end{aligned} \quad (\text{A.151})$$

Substituting for $\frac{\partial h}{\partial a}$ and $h'(\mu)$ using the above expressions yields

$$\frac{\partial \mu}{\partial a} = \frac{\frac{2}{a} c'(\mu) - \frac{v}{\mu^2}}{c''(\mu) - \left(\frac{va^2}{\mu^4} - \frac{2}{\mu} c'(\mu) \right)} = \frac{\frac{2\mu^2 c'(\mu)}{a} - v}{\frac{a^2}{\mu^2} \left[\frac{\mu^3}{a^2} (2c'(\mu) + \mu c''(\mu)) - v \right]},$$

which, together with (A.149), implies that

$$a^2 \frac{\partial \mu}{\partial a} = \frac{\frac{2\mu^2 c'(\mu)}{a} - v}{\frac{1}{\mu^2} \left[\frac{\mu^3}{a^2} (2c'(\mu) + \mu c''(\mu)) - v \right]} = \varphi(\mu; a, v) \cdot \frac{\partial \mu}{\partial p}, \quad (\text{A.152})$$

where $\varphi(\mu; a, v) := \frac{2\mu^2 c'(\mu)}{a} - v$. Note that, for all $\mu > a$,

$$\varphi(\mu; a, v) = \frac{2\mu^2 c'(\mu)}{a} - v < \frac{\mu}{a} \cdot \frac{2\mu^2 c'(\mu)}{a} - v < \mu^2 \left(\frac{\partial \mu}{\partial p} \right)^{-1}, \quad (\text{A.153})$$

where the last inequality follows from (A.149). Thus, one can leverage the results about $\frac{\partial \mu}{\partial p}$ from (b)(i) to study $\varphi(\mu; a, v)$.

(b)(ii): When a smaller underloaded equilibrium $\mu_2^*(a, p; v)$ exists, recall from (b)(i) that $\frac{\partial \mu_2^*(a, p; v)}{\partial p} < 0$. Then, (A.153) implies

$$\varphi(\mu_2^*(a, p; v); a, v) < (\mu_2^*(a, p; v))^2 \left(\frac{\partial \mu_2^*(a, p; v)}{\partial p} \right)^{-1} < 0, \quad (\text{A.154})$$

which, from (A.152), implies that $\frac{\partial \mu_2^*(a, p; v)}{\partial a} = \frac{1}{a^2} \varphi(\mu_2^*(a, p; v); a, v) \frac{\partial \mu_2^*(a, p; v)}{\partial p} > 0$, i.e., $\mu_2^*(a, p; v)$, when it exists, is strictly increasing in a .

(b)(iii): Throughout the proof of (b)(iii), recall from (b)(i) that $\frac{\partial \mu_1^*(a, p; v)}{\partial p} > 0$.

Definition 18. For any $v > 0$, and $p \in [0, p^\dagger(v)]$, $a^\dagger(p; v)$ is the unique solution for $a \in \left(0, \frac{v}{2p}\right]$ to $\varphi(\mu_1^*(a, p; v); a, v) = 0$, which simplifies to

$$c' \left(\frac{v}{p + \frac{v}{2a}} \right) = \frac{av}{2} \left(\frac{p}{v} + \frac{1}{2a} \right)^2. \quad (\text{A.155})$$

Remark 22. For any $v > 0$ and $p \in [0, p^\dagger(v)]$, $a^\dagger(p; v) = \frac{v}{2p}$ if and only if $p = p^\dagger(v)$.

Remark 23. $\varphi(\mu_1^*(a, p, v); a, v) > 0$ if $a \in \left(0, a^\dagger(p; v)\right)$, and $\varphi(\mu_1^*(a, p, v); a, v) < 0$ if $a \in \left(a^\dagger(p; v), \frac{v}{2p}\right]$.

Note that $a^\dagger(p; v)$ depends on p and v , but for simplicity, we may expose or suppress the

dependence. The following lemma provides conditions for the existence of a^\dagger .

Lemma 51 (Validating Definition 18). *There exists a solution for $a \in \left(0, \frac{v}{2p}\right]$ that solves (A.155) if and only if $p \in [0, +p^\ddagger(v)]$, where $p^\ddagger(v)$ is defined in Definition 16 (b). Furthermore, if such a solution exists, it is unique, and is denoted by $a^\dagger(p; v)$.*

Lemma 52. *For any $v > 0$, $a^\dagger(p; v)$ is strictly increasing in $p \in [0, p^\ddagger(v)]$.*

Let

$$g(a; p, v) := \begin{cases} 0, & a = 0, \\ \frac{2v^2}{a(p+\frac{v}{2a})^2} c' \left(\frac{v}{p+\frac{v}{2a}} \right), & a > 0. \end{cases} \quad (\text{A.156})$$

Then, it is easy to see that (A.155) is equivalent to $g(a; p, v) = v$; that is, $a^\dagger(p; v) \in \left(0, \frac{v}{2p}\right]$, if it exists, solves $g(a; p, v) = v$. Note that g is continuous at $a = 0$, and it is straightforward to see that it is a strictly increasing function of a for $a \in \left(0, \frac{v}{2p}\right]$.

(b)(iii)(1): Using Remark 22 to exclude $p = p^\ddagger(v)$ from Lemma 51, when $0 \leq p < p^\ddagger(v)$, $a^\dagger(p; v) \in \left(0, \frac{v}{2p}\right)$. Recall from (A.152) that $\frac{\partial \mu_1^*(a, p; v)}{\partial a} = \frac{1}{a^2} \varphi(\mu_1^*(a, p; v); a, v) \frac{\partial \mu_1^*(a, p; v)}{\partial p}$, where $\frac{\partial \mu_1^*(a, p; v)}{\partial p} > 0$ from (b)(i).

Case (I): If $0 \leq p < p^\ddagger(v)$, considering Remark 23, it suffices to show $a^\dagger(p; v) < \bar{a}(p, v)$, where $\bar{a}(p, v) < \frac{v}{2p}$ from Remark 20. (Recall from Theorem 4 (b)(iii) that $\mu_1^*(a, p; v)$ exists only on $(0, \bar{a}(p, v))$.) We first establish the following claim:

Claim 13. *If $0 \leq p < p^\ddagger(v)$, then $g(\bar{a}(p, v); p, v) > v$, where g is defined in (A.156) and $\bar{a}(p, v)$ is defined in Theorem 4.*

Recalling that $g(a; p, v)$ is a strictly increasing function of a for $0 < a \leq \bar{a}(p, v) < \frac{v}{2p}$ with $g(0; p, v) = 0$ and $g(\bar{a}(p, v); p, v) > v$ (from Claim 13), it follows that there exists a unique $a^\dagger(p; v) \in (0, \bar{a}(p, v))$ such that $g(a^\dagger(p; v); p, v) = v$. Therefore, $a^\dagger(p; v) < \bar{a}(p, v)$.

Case (II): If $p^\ddagger(v) \leq p < p^\ddagger(v)$, considering Remark 23, it suffices to show that $a^\dagger(p; v) < (c')^{-1}(p)$, where $(c')^{-1}(p) < \frac{v}{2p}$ when $p < p^\ddagger(v)$ from Definition 16 (b). (Recall

from Theorem 4 (b)(iv) that $\mu_1^\star(a, p; v)$ exists only on $(0, (c')^{-1}(p))$.) It is straightforward to see that the right-hand side of (A.155) is a strictly decreasing function of a for $0 < a \leq (c')^{-1}(p) < \frac{v}{2p}$, and the left-hand side of (A.155) is strictly increasing in $a > 0$. Thus, in order to show $a^\dagger(p; v) < (c')^{-1}(p)$, it suffices to show that the left-hand side of (A.155) is strictly greater than its right-hand side at $a = (c')^{-1}(p)$, i.e.,

$$p^\dagger(v) \leq p = c'(a) < p^\ddagger(v) \quad \text{implies} \quad c' \left(\frac{v}{c'(a) + \frac{v}{2a}} \right) > \frac{av}{2} \left(\frac{c'(a)}{v} + \frac{1}{2a} \right)^2. \quad (\text{A.157})$$

By definition of $p^\dagger(v)$ in Definition 16 (a) and $p^\ddagger(v)$ in Definition 16 (b), $p^\dagger(v) \leq p < p^\ddagger(v)$ is equivalent to

$$p(c')^{-1}(p) < \frac{v}{2} \leq p(c')^{-1}(p) + \frac{1}{2} ((c')^{-1}(p))^2 c''((c')^{-1}(p)). \quad (\text{A.158})$$

Using (A.158), (A.157) is equivalent to

$$c'(a) < \frac{v}{2a} \leq c'(a) + \frac{1}{2} ac''(a) \quad \text{implies} \quad c' \left(\frac{v}{c'(a) + \frac{v}{2a}} \right) > \frac{av}{2} \left(\frac{c'(a)}{v} + \frac{1}{2a} \right)^2. \quad (\text{A.159})$$

Note that $a < \frac{v}{p+\frac{v}{2a}}$ for any $a \in \left(0, \frac{v}{2p}\right)$. Then, by convexity of c' (Assumption 1), for any $a \in \left(0, \frac{v}{2p}\right)$,

$$c''(a) \leq \frac{c' \left(\frac{v}{p+\frac{v}{2a}} \right) - c'(a)}{\frac{v}{p+\frac{v}{2a}} - a} \Leftrightarrow c' \left(\frac{v}{p+\frac{v}{2a}} \right) \geq c'(a) + c''(a) \left(\frac{v}{p+\frac{v}{2a}} - a \right). \quad (\text{A.160})$$

Moreover, $0 < \frac{v}{2a} - c'(a) \leq \frac{1}{2} ac''(a)$ implies

$$c''(a) \geq \frac{2}{a} \left(\frac{v}{2a} - c'(a) \right) \stackrel{(*)}{>} \frac{1}{2a} \left(\frac{v}{2a} - c'(a) \right) > 0, \quad (\text{A.161})$$

where $(*)$ follows by noting that $\frac{v}{2a} - c'(a) > 0$. Then, combining (A.160) and (A.161)

implies that, when $a = (c')^{-1}(p)$,

$$\begin{aligned} c' \left(\frac{v}{p + \frac{v}{2a}} \right) &> c'(a) + \frac{1}{2a} \left(\frac{v}{2a} - c'(a) \right) \left(\frac{v}{p + \frac{v}{2a}} - a \right) = c'(a) + \frac{1}{2} \frac{\left(\frac{v}{2a} - c'(a) \right)^2}{\frac{v}{2a} + c'(a)} \\ &\stackrel{(*)}{=} \frac{a}{2v} \left[\frac{\left(\frac{v}{2a} - c'(a) \right)^3}{\frac{v}{2a} + c'(a)} + \left(c'(a) + \frac{v}{2a} \right)^2 \right] > \frac{a}{2v} \left(c'(a) + \frac{v}{2a} \right)^2, \end{aligned}$$

where $(*)$ follows from algebra. Hence, (A.159) is established.

(b)(iii)(2): When $p \geq p^\ddagger(v)$, it suffices to show that $\varphi(\mu_1^*(a, p; v)) > 0$ for all $a \in (0, (c')^{-1}(p))$, recalling from (A.152) that $\frac{\partial \mu_1^*(a, p; v)}{\partial a} = \frac{1}{a^2} \varphi(\mu_1^*(a, p; v); a, v) \frac{\partial \mu_1^*(a, p; v)}{\partial p}$, where $\frac{\partial \mu_1^*(a, p; v)}{\partial p} > 0$ from (b)(i). (Recall from Theorem 4 (b)(iv) that $\mu_1^*(a, p; v)$ exists only for $a \in (0, (c')^{-1}(p))$.) We discuss two cases depending on whether there exists $a > 0$ that solves (A.155).

- **Case (I):** If there does not exist $a > 0$ that solves (A.155), then $\varphi(\mu_1^*(a, p; v)) > 0$ for all $a \in (0, (c')^{-1}(p))$, because $\lim_{a \downarrow 0} \varphi(\mu_1^*(a, p; v); a, v) = \infty > 0$.

- **Case (II):** If there exists $a > 0$ that solves (A.155), then denote the smallest solution by $a^\dagger(p; v)$. Then, it suffices to show that $a^\dagger(p; v) \geq (c')^{-1}(p)$. Then, $\varphi(\mu_1^*(a, p; v)) > 0$ for all $a \in (0, (c')^{-1}(p))$, because $\lim_{a \downarrow 0} \varphi(\mu_1^*(a, p; v); a, v) = \infty > 0$.

- **Case (II-1):** If $p = p^\ddagger(v)$, Remark 22 implies that $a^\dagger(p; v) = \frac{v}{2p} = (c')^{-1}(p)$ by definition of $p^\ddagger(v)$ in Definition 16 (b).

- **Case (II-2):** If $p > p^\ddagger(v)$, then it is straightforward that $(c')^{-1}(p) > \frac{v}{2p}$ (by definition of $p^\ddagger(v)$ in Definition 16 (b)). By Lemma 51 and Remark 22, when $p > p^\ddagger(v)$, $a^\dagger(p; v)$, if exists, must satisfy $a^\dagger(p; v) > \frac{v}{2p}$. This can be equivalently written as

$$a^\dagger(p; v) > \frac{v}{p + \frac{v}{2a^\dagger(p; v)}},$$

which implies that

$$c' \left(a^\dagger(p; v) \right) > c' \left(\frac{v}{p + \frac{v}{2a^\dagger(p; v)}} \right),$$

recalling that c' is a strictly increasing function (by strict convexity of c). Note that

$$c' \left(\frac{v}{p + \frac{v}{2a^\dagger(p; v)}} \right) \stackrel{(i)}{=} \frac{a^\dagger(p; v)v}{2} \left(\frac{p}{v} + \frac{1}{2a^\dagger(p; v)} \right)^2 \stackrel{(ii)}{>} \frac{\frac{v}{2p}v}{2} \left(\frac{p}{v} + \frac{1}{2\frac{v}{2p}} \right)^2 = p,$$

where (i) follows because $a^\dagger(p; v)$ solves (A.155), and (ii) follows because $a^\dagger(p; v) > \frac{v}{2p}$ and $\frac{a(p+\frac{v}{2a})^2}{2v}$ is increasing in $a \in \left(\frac{v}{2p}, \infty\right)$. Hence, from the above two displays,

$$c'(a^\dagger(p; v)) > p \Leftrightarrow a^\dagger(p; v) > (c')^{-1}(p),$$

which is desired. ■

Below we prove Lemmas 51-52, Remarks 22-23, and Claims 10-13.

A.7.11.1 Proof of Lemma 51

By definition of a^\dagger in Definition 18, $\varphi(\mu_1^*; a^\dagger, v) = 0$ implies

$$\frac{2(\mu_1^*)^2 c'(\mu_1^*)}{a^\dagger} = v. \quad (\text{A.162})$$

Since μ_1^* satisfies the limiting FOC (A.146), substituting for $c'(\mu_1^*)$ using (A.146) in the above display yields

$$\mu_1^*(a^\dagger, p; v) = \frac{v}{p + \frac{v}{2a^\dagger}}. \quad (\text{A.163})$$

Recall from (A.154) that when $\mu_2^*(a, p; v)$ exists, $\varphi(\mu_2^*; a, v) < 0$, including when $\mu_2^*(a^\dagger, p; v)$ exists.

Substitution into (A.162) using (A.163) shows that a^\dagger satisfies

$$\frac{2v^2}{a(p + \frac{v}{2a})^2} c' \left(\frac{v}{p + \frac{v}{2a}} \right) = v \Leftrightarrow c' \left(\frac{v}{p + \frac{v}{2a}} \right) = \frac{av}{2} \left(\frac{p}{v} + \frac{1}{2a} \right)^2,$$

which establishes (A.155).

It is straightforward to see that $g(a; p, v)$ is a strictly increasing function of a for $a \in (0, \frac{v}{2p}]$ with $g(0; p, v) = 0$. Thus, $a^\dagger \in (0, \frac{v}{2p}]$ that solves $g(a; p, v) = v$ exists if and only if $g\left(\frac{v}{2p}; p, v\right) \geq v$, i.e.,

$$g\left(\frac{v}{2p}; p, v\right) = \frac{v}{p} c'\left(\frac{v}{2p}\right) \geq v,$$

which holds if and only if $0 \leq p \leq p^\ddagger(v)$, recalling the definition of $p^\ddagger(v)$ in Definition 16 (b).

From Remark 22, $a^\dagger(p; v) = \frac{v}{2p}$ when $p = p^\ddagger(v)$. Moreover, from Lemma 52, $a^\dagger(p; v)$ is strictly increasing in $p \in [0, p^\ddagger(v)]$. Hence, $a^\dagger(p; v) \leq \frac{v}{2p}$ for $p \in [0, p^\ddagger(v)]$.

Finally, since $g(a; p, v)$ is a strictly increasing function of a , a^\dagger , which is a solution to $g(a; p, v) = v$, must be unique, if it exists. ■

A.7.11.2 Proof of Lemma 52

By definition, $a^\dagger(p; v)$ solves (A.155). Denote the left-hand side of (A.155) by $LHS(a, p; v)$, and the right-hand side of (A.155) by $RHS(a, p; v)$. Note that

- (1) $LHS(a, p; v)$ is strictly increasing in $a \in (0, \infty)$, and $RHS(a, p; v)$ is strictly decreasing in $a \in (0, \frac{v}{2p})$ and strictly increasing in $a \in (\frac{v}{2p}, \infty)$;
- (2) $LHS(a, p; v)$ is strictly decreasing in p , and $RHS(a, p; v)$ is strictly increasing in p .

Then, for any $0 \leq p_1 < p_2 \leq p^\ddagger(v)$, we have $LHS(a^\dagger(p_1; v), p_1; v) = RHS(a^\dagger(p_1; v), p_1; v)$.

Then, (2) implies that $LHS(a^\dagger(p_1; v), p_2; v) < LHS(a^\dagger(p_1; v), p_1; v)$ and $RHS(a^\dagger(p_1; v), p_2; v) > RHS(a^\dagger(p_1; v), p_1; v)$, implying that $LHS(a^\dagger(p_1; v), p_2; v) < RHS(a^\dagger(p_1; v), p_2; v)$. Then, (1) implies that $a^\dagger(p_1; v) < a^\dagger(p_2; v)$. Therefore, $a^\dagger(p; v)$ is strictly increasing in $p \in [0, p^\ddagger(v)]$. ■

A.7.11.3 Proof of Remark 22

Plugging in $a = \frac{v}{2p}$ into (A.155) yields $c' \left(\frac{v}{2p} \right) = p$, which implies that $p = p^\dagger(v)$ by uniqueness and by the definition of $p^\dagger(v)$ (in Definition 16 (b)). \blacksquare

A.7.11.4 Proof of Remark 23

We evaluate the partial derivative of $\varphi(\mu_1^*(a, p; v); a, v)$ with respect to a at $a = a^\dagger(p; v)$:

$$\begin{aligned} & \frac{\partial}{\partial a} \varphi(\mu_1^*(a, p; v); a, v) \Big|_{a=a^\dagger} = \frac{\partial}{\partial a} \left(\frac{2\mu_1^*(a, p; v)^2 c'(\mu_1^*(a, p; v))}{a} - v \right) \Big|_{a=a^\dagger} \\ &= \frac{1}{(a^\dagger)^2} \left[a^\dagger \cdot \left(4\mu_1^* c'(\mu_1^*) \frac{\partial \mu_1^*(a, p; v)}{\partial a} \Big|_{a=a^\dagger} + 2(\mu_1^*)^2 c''(\mu_1^*) \frac{\partial \mu_1^*(a, p; v)}{\partial a} \Big|_{a=a^\dagger} \right) - 2(\mu_1^*)^2 c'(\mu_1^*) \right] \\ &= -\frac{2}{(a^\dagger)^2} (\mu_1^*)^2 c'(\mu_1^*) < 0, \end{aligned}$$

which implies that $\varphi(\mu_1^*(a, p; v); a, v)$ is strictly decreasing in a at $a = a^\dagger$. Since a^\dagger is the unique $a \in (0, \frac{v}{2p}]$ that satisfies $\varphi(\mu_1^*(a, p; v); a, v) = 0$, it follows that $\varphi(\mu_1^*(a, p; v); a, v) > 0$ for all $a \in (0, a^\dagger(p; v))$, and $\varphi(\mu_1^*(a, p; v); a, v) < 0$ for all $a \in (a^\dagger(p; v), \frac{v}{2p}]$. \blacksquare

A.7.11.5 Proof of Claim 10

We first note that $0 < a < \bar{a}(p, v)$ is equivalent to $v > v^\dagger(a, p)$ (from Corollary 6). When $v > v^\dagger(a, p)$, we have $h(\mu^\dagger(a, p); a, p, v) > c'(\mu^\dagger(a, p))$, because $h(\mu^\dagger(a, p); a, p, v) > h(\mu^\dagger(a, p); a, p, v^\dagger(a, p)) = c'(\mu^\dagger(a, p))$ (from Lemma 43), recalling that $h(\mu; a, p, v)$ is strictly increasing in v when $\mu > a$ (from Lemma 40 (e)). Then, since $h(a) = p < c'(a)$, $h(\mu^\dagger(a, p); a, p, v) > c'(\mu^\dagger(a, p))$ and $\lim_{\mu \rightarrow \infty} h(\mu) = 0 < \infty = \lim_{\mu \rightarrow \infty} c'(\mu)$, the intermediate value theorem implies that there exists $\mu_2^*(a, p; v)$ in $(a, \mu^\dagger(a, p))$ and $\mu_1^*(a, p; v)$ in $(\mu^\dagger(a, p), \infty)$; that is, $a < \mu_2^*(a, p; v) < \mu^\dagger(a, p) < \mu_1^*(a, p; v)$. \blacksquare

A.7.11.6 Proof of Claim 11

We formally prove the first part of the statement, and the second part follows by similar arguments. The proof follows three steps:

- (i) $\mu_1^*(a, p; v)$ has a limit as $p \uparrow c'(a)$;
- (ii) $\lim_{p \uparrow c'(a)} \mu_1^*(a, p; v)$ solves the limiting FOC (A.146);
- (iii) $\lim_{p \uparrow c'(a)} \mu_1^*(a, p; v) = \mu_1^*(a, c'(a); v)$.

Proof of (i): Recall that $\mu_1^*(a, p; v)$ and $\mu_2^*(a, p; v)$ are continuous in p for $p < \min\{c'(a), p^\dagger(v)\}$ when $a \in (0, \bar{a}(p^\dagger(v), v))$ (i.e., within Area 1). This implies that $\mu_1^*(a, p; v)$ and $\mu_2^*(a, p; v)$ have limits as $p \uparrow c'(a)$, denoted by $L_1 := \lim_{p \uparrow c'(a)} \mu_1^*(a, p; v)$ and $L_2 := \lim_{p \uparrow c'(a)} \mu_2^*(a, p; v)$.

Proof of (ii): Let $\varphi(\mu, a, p; v) := c'(\mu) - \frac{a^2}{\mu^2} \left(p + \frac{v}{a} - \frac{v}{\mu} \right)$. Any limiting equilibrium must satisfy the limiting FOC (A.146), and so

$$\varphi\left(\mu_j^*(a, p; v), a, p; v\right) = 0, \quad j \in \{1, 2\},$$

which implies

$$\lim_{p \uparrow c'(a)} \varphi\left(\mu_j^*(a, p; v), a, p; v\right) = 0, \quad j \in \{1, 2\}. \quad (\text{A.164})$$

Since $\varphi(\mu, a, p; v)$ is a continuous function of p , it follows that

$$\lim_{p \uparrow c'(a)} \varphi(\mu_j^*(a, p; v), a, p; v) = \varphi(L_j, a, c'(a); v), \quad j \in \{1, 2\}. \quad (\text{A.165})$$

From (A.164) and (A.165), we conclude that $\varphi(L_j, a, c'(a); v) = 0$, $j \in \{1, 2\}$, which implies that both L_1 and L_2 solve the limiting FOC (A.146).

Proof of (iii): From Theorem 4 (b), when $p = c'(a)$ and $p < p^\dagger(v)$ (i.e., Area 51), there exist an underloaded equilibrium, namely, $\mu_1^*(a, c'(a); v) > a$, and a critically loaded equilibrium a . Therefore, L_1 and L_2 must be either $\mu_1^*(a, c'(a); v)$ or a .

- Suppose $L_1 = a$ and $L_2 = \mu_1^*(a, c'(a); v)$. Then, since $\mu_1^*(a, c'(a); v) > a$, there must exist $p < c'(a)$ such that $\mu_2^*(a, p; v) > \mu_1^*(a, p; v)$, which contradicts the definition of μ_1^* as the larger limiting equilibrium. Hence, $L_1 = a$ and $L_2 = \mu_1^*(a, c'(a); v)$ cannot happen.
- Suppose $L_1 = L_2 = a$. Then, $\lim_{p \uparrow c'(a)} \mu_o^*(a, p; v) = \mu_1^*(a, c'(a); v)$, recalling from Theorem 4 (b) that there exists an overloaded equilibrium $\mu_o^*(a, p; v) < a$ when $p < \min\{c'(a), p^\dagger(v)\}$ and $a \in (0, \bar{a}(p^\dagger(v), v))$ (i.e., within Area 1). Since $\mu_1^*(a, c'(a); v) > a$, there must exist $p < c'(a)$ such that $\mu_o^*(a, p; v) > \mu_1^*(a, p; v)$ and $\mu_o^*(a, p; v) > \mu_2^*(a, p; v)$, which contradicts the definition of $\mu_o^*(a, p; v)$ as an overloaded equilibrium, and $\mu_1^*(a, p; v)$ and $\mu_2^*(a, p; v)$ as underloaded equilibria. Hence, $L_1 = L_2 = a$ cannot happen.

Therefore, we conclude that $L_1 = \mu_1^*(a, c'(a); v)$ and $L_2 = a$, or $L_1 = L_2 = \mu_1^*(a, c'(a); v)$.

In either case, $L_1 = \mu_1^*(a, c'(a); v)$, which completes Step (iii). ■

A.7.11.7 Proof of Claim 12

We prove by contradiction. Suppose that $c''(\mu) - h'(\mu) = 0$ for some $\mu_o > a$. Then, from the left equation in (A.150), $\frac{\partial h}{\partial a} = 0$ at μ_o . Using the expressions for $h'(\mu)$ and $\frac{\partial h}{\partial a}$ in (A.151), it follows that

$$c''(\mu_o) - \frac{va^2}{\mu_o^4} + \frac{2}{\mu_o} c'(\mu_o) = 0 \Leftrightarrow \mu_o^3 c''(\mu_o) = \frac{va^2}{\mu_o} - 2\mu_o^2 c'(\mu_o), \text{ and} \quad (\text{A.166})$$

$$\frac{2}{a} c'(\mu_o) - \frac{v}{\mu_o^2} = 0 \Leftrightarrow \mu_o^2 c'(\mu_o) = \frac{1}{2}av. \quad (\text{A.167})$$

Substituting for $\mu_o^2 c'(\mu_o)$ using (A.167) into (A.166) yields

$$\mu_o^3 c''(\mu_o) = \frac{va^2}{\mu_o} - av = va \left(\frac{a}{\mu_o} - 1 \right) < 0,$$

which contradicts $c'' > 0$ (by strict convexity of c). Hence, $c''(\mu) \neq h'(\mu)$ for all $\mu > a$. \blacksquare

A.7.11.8 Proof of Claim 13

Substituting for $\mu c''(\mu)$ using (A.110) in Definition 15 (a) into (A.111) yields

$$\begin{aligned} v^\dagger(a, p) &= \frac{(\mu^\dagger)^3}{a^2} \left(2c'(\mu^\dagger) + \frac{(3a - 2\mu^\dagger) c'(\mu^\dagger) - \frac{a^3 p}{(\mu^\dagger)^2}}{\mu^\dagger - a} \right) \\ &= \frac{\mu^\dagger}{a(\mu^\dagger - a)} \left((\mu^\dagger)^2 c'(\mu^\dagger) - a^2 p \right), \end{aligned}$$

which implies

$$\frac{(\mu^\dagger)^2 c'(\mu^\dagger)}{a} + \frac{av^\dagger(a, p)}{\mu^\dagger} = v^\dagger(a, p) + ap, \quad \forall a > 0, p \geq 0. \quad (\text{A.168})$$

Let $\phi(\mu; a, v) := \frac{\mu^2 c'(\mu)}{a} + \frac{av}{\mu}$. Then, note that

$$\frac{\mu^2}{a} \phi'(\mu; a, v) = \frac{\mu^3}{a^2} (2c'(\mu) + \mu c''(\mu)) - v.$$

Plugging in $a = \bar{a}$ into the above equation:

$$\begin{aligned} \frac{\mu^2}{\bar{a}} \phi'(\mu; \bar{a}, v) &= \frac{\mu^3}{\bar{a}^2} (2c'(\mu) + \mu c''(\mu)) - v \\ &= \frac{\mu^3}{\bar{a}^2} (2c'(\mu) + \mu c''(\mu)) - \frac{(\mu^\dagger(\bar{a}, p))^3}{\bar{a}^2} \left(2c'(\mu^\dagger(\bar{a}, p)) + \mu^\dagger(\bar{a}, p) c''(\mu^\dagger(\bar{a}, p)) \right), \end{aligned}$$

where the second equality follows from $v = v^\dagger(\bar{a}, p)$ by the definition of $\bar{a}(p, v)$ when $0 \leq p < p^\dagger(v)$ (see Definition 17) and from the definition of $v^\dagger(a, p)$ in Definition 15 (b). Since $\frac{\mu^3}{a^2} (2c'(\mu) + \mu c''(\mu))$ is a strictly increasing function of μ , it is clear from the above display

that $\phi'(\mu; \bar{a}, v)$ is strictly negative when $\mu < \mu^\dagger(\bar{a}, p)$, and is strictly positive when $\mu > \mu^\dagger(\bar{a}, p)$. That is, $\phi(\mu; \bar{a}, v)$ is strictly decreasing in $\mu \in (0, \mu^\dagger(\bar{a}, p))$, strictly increasing in $\mu \in (\mu^\dagger(\bar{a}, p), \infty)$, and thus, is minimized at $\mu = \mu^\dagger(\bar{a}, p)$. Hence,

$$\phi(\mu; \bar{a}, v) > \phi(\mu^\dagger(\bar{a}, p); \bar{a}, v), \quad \forall \mu \neq \mu^\dagger(\bar{a}, p). \quad (\text{A.169})$$

Recall from Lemma 41 and Definition 17 that $\mu^\dagger(\bar{a}, p)$ satisfies $h'(\mu^\dagger(\bar{a}, p); \bar{a}, p, v) = c''(\mu^\dagger(\bar{a}, p)) > 0$, hence $\mu^\dagger(\bar{a}, p)$ lies on the increasing portion of $h(\mu; \bar{a}, p, v)$; that is, $\mu^\dagger(\bar{a}, p) < \frac{3v}{2p + \frac{2v}{\bar{a}}}$, recalling from Lemma 40 (a). Moreover, it is easy to see that $\frac{3v}{2p + \frac{2v}{\bar{a}}} < \frac{v}{p + \frac{v}{2\bar{a}}}$ given that $\bar{a} < \frac{v}{2p}$ from Remark 20. Thus, $\mu^\dagger(\bar{a}, p) > \frac{v}{p + \frac{v}{2\bar{a}}}$, and hence, from (A.169),

$$\begin{aligned} & \phi\left(\frac{v}{p + \frac{v}{2\bar{a}}}; \bar{a}, v\right) > \phi(\mu^\dagger(\bar{a}, v); \bar{a}, v) \\ \Leftrightarrow & \frac{\left(\frac{v}{p + \frac{v}{2\bar{a}}}\right)^2 c'\left(\frac{v}{p + \frac{v}{2\bar{a}}}\right)}{\bar{a}} + \frac{\bar{a}v}{\frac{v}{p + \frac{v}{2\bar{a}}}} > \frac{(\mu^\dagger(\bar{a}, p))^2 c'(\mu^\dagger(\bar{a}, p))}{\bar{a}} + \frac{\bar{a}v}{\mu^\dagger(\bar{a}, p)}. \end{aligned}$$

Then, from (A.168) by plugging in $a = \bar{a}$, the above display is equivalent to

$$\frac{1}{2} \cdot g(\bar{a}; p, v) + \bar{a}p + \frac{v}{2} > v^\dagger(\bar{a}, p) + \bar{a}p = v + \bar{a}p \quad \Leftrightarrow \quad g(\bar{a}; p, v) > v.$$

■

A.7.12 Proof of Proposition 8

By definition in (1.17), $\mu_{\max}^*(p, v)$ is the supremum (limiting) equilibrium service rate over all possible values of a , given p and v . To determine this, we derive the supremum underloaded equilibrium and the supremum overloaded equilibrium separately; then, $\mu_{\max}^*(p, v)$ is the maximum of these two. (We do not consider the critically loaded equilibrium separately, because every critically loaded equilibrium at $a = (c')^{-1}(p)$ is also an overloaded equilibrium for all $a > (c')^{-1}(p)$.)

(I) $0 \leq \mathbf{p} < \mathbf{p}^\dagger(\mathbf{v})$: From Proposition 7 (b)(iii)(1), the maximum underloaded equilibrium

is obtained at $a = a^\dagger(p; v) \in (0, \frac{v}{2p})$, which satisfies

$$c' \left(\frac{v}{p + \frac{v}{2a^\dagger}} \right) = \frac{a^\dagger v}{2} \left(\frac{p}{v} + \frac{1}{2a^\dagger} \right)^2. \quad (\text{A.170})$$

Note that plugging $\mu = \frac{v}{p + \frac{v}{2a^\dagger}}$ into the limiting FOC (A.146) with $a = a^\dagger$ yields (A.170), which means that $\mu = \frac{v}{p + \frac{v}{2a^\dagger}}$ is an underloaded equilibrium. This can only be μ_1^* because even if μ_2^* exists, $\frac{\partial \mu_2^*(a, p; v)}{\partial a}|_{a=a^\dagger} > 0$ (from Theorem 7 (b)(ii)), contradicting the definition of a^\dagger for which this partial derivative is zero. Thus, the maximum underloaded equilibrium is given by

$$\mu_1^*(a^\dagger, p; v) = \frac{v}{p + \frac{v}{2a^\dagger}}. \quad (\text{A.171})$$

From Proposition 7 (a)(ii), the overloaded equilibrium $\mu_o^*(a, p; v)$, when it exists, does not depend on a , and is given by $\mu_o^*(a, p; v) = (c')^{-1}(p)$ (from the limiting FOC (1.16)).

It remains to compare the maximum underloaded equilibrium and the overloaded equilibrium. We argue that the former is larger. To see this, it suffices to show that

$$(c')^{-1}(p) < \frac{v}{p + \frac{v}{2a^\dagger}} \quad \stackrel{(i)}{\Leftrightarrow} \quad p < c' \left(\frac{v}{p + \frac{v}{2a^\dagger}} \right) \quad \stackrel{(ii)}{\Leftrightarrow} \quad p < \frac{a^\dagger v}{2} \left(\frac{p}{v} + \frac{1}{2a^\dagger} \right)^2, \quad (\text{A.172})$$

where (i) follows because c' is strictly increasing (by strict convexity of c), and (ii) follows from (A.170). Using Remark 22 to exclude $p = p^\ddagger(v)$ from Lemma 51, $a^\dagger \in (0, \frac{v}{2p})$ when $0 \leq p < p^\ddagger(v)$. Additionally, it is straightforward to see that $\frac{av}{2} \left(\frac{p}{v} + \frac{1}{2a} \right)^2$ is strictly decreasing in a for $a \in (0, \frac{v}{2p})$. Thus, it follows that

$$\frac{a^\dagger v}{2} \left(\frac{p}{v} + \frac{1}{2a^\dagger} \right)^2 > \frac{\frac{v}{2p} v}{2} \left(\frac{p}{v} + \frac{1}{2\frac{v}{2p}} \right)^2 = p,$$

which establishes (A.172).

Hence, the maximum equilibrium is the maximum underloaded equilibrium, namely,

$$\mu_{\max}^*(p, v) = \mu_1^*(a^\dagger, p; v) = \frac{v}{p + \frac{v}{2a^\dagger}} = \left(\frac{p}{v} + \frac{1}{2a^\dagger} \right)^{-1}.$$

(II) $p \geq p^\dagger(v)$: From Proposition 7 (b)(iii)(2), the underloaded equilibrium $\mu_1^*(a, p; v)$, when it exists, is strictly increasing in $a \in (0, (c')^{-1}(p))$. Using similar arguments that prove the first part of Claim 11, it can be shown that $\lim_{a \rightarrow (c')^{-1}(p)} \mu_1^*(a, p; v)$ solves the limiting FOC, which has the unique solution $\mu = a$ by Theorem 4 (b)(ii). Thus, $\mu_1^*(a, p; v) \rightarrow (c')^{-1}(p)$ when $a \rightarrow (c')^{-1}(p)$. On the other hand, identical to the analysis in (I) above, the overloaded equilibrium $\mu_o^*(a, p; v)$, when it exists, is given by $\mu_o^*(a, p; v) = (c')^{-1}(p)$. Hence, the maximum equilibrium is given by $\mu_{\max}^*(p, v) = (c')^{-1}(p)$.

Finally, we verify the equivalence of this statement to Proposition 3.

- When $0 \leq p < p^\dagger(v)$,

$$\mu_{\max}^* c'(\mu_{\max}^*) = \frac{v}{p + \frac{v}{2a^\dagger}} c' \left(\frac{v}{p + \frac{v}{2a^\dagger}} \right) \stackrel{(*)}{=} \frac{v}{p + \frac{v}{2a^\dagger}} \frac{a^\dagger (p + \frac{v}{2a^\dagger})^2}{2v} = \frac{a^\dagger}{2} \left(p + \frac{v}{2a^\dagger} \right).$$

where (*) follows from (A.170). Moreover,

$$\frac{v^2}{4(v - p\mu_{\max}^*)} = \frac{v^2}{4 \left(v - p \frac{v}{p + \frac{v}{2a^\dagger}} \right)} = \frac{a^\dagger}{2} \left(p + \frac{v}{2a^\dagger} \right).$$

The above two displays imply that $\mu_{\max}^* c'(\mu_{\max}^*) = \frac{v^2}{4(v - p\mu_{\max}^*)}$ when $0 \leq p < p^\dagger(v)$.

- When $p \geq p^\dagger(v)$, $c'(\mu_{\max}^*) = p$.

Therefore, (1.18) satisfies the equations that characterize μ_{\max}^* for the loss system in Proposition 3. ■

A.7.13 Proofs of Theorems 5A and 5B

From (1.6), the prelimit FOC is given by

$$c'(\mu) = p \left(1 - I^\lambda(\mu, \mu) \right) + (v - p\mu) \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu}. \quad (\text{A.173})$$

Set $LHS(\mu)$ and $RHS^\lambda(\mu)$ equal to the left-hand side and right-hand side of the above display, respectively. From (1.16), the limiting FOC is given by

$$c'(\mu) = p \left(1 - \left[1 - \frac{a}{\mu} \right]^+ \right) + (v - p\mu) \frac{a}{\mu^2} \left[1 - \frac{a}{\mu} \right]^+. \quad (\text{A.174})$$

Recall the definition of $h(\mu)$ from (A.108) in Section A.7.3:

$$h(\mu) := \frac{a^2}{\mu^2} \left(p + \frac{v}{a} - \frac{v}{\mu} \right) = p \frac{a}{\mu} + (v - p\mu) \frac{a}{\mu^2} \left(1 - \frac{a}{\mu} \right), \quad \mu > 0.$$

Then, (A.174) can be written as

$$c'(\mu) = \begin{cases} p, & \text{when } \mu < a, \\ h(\mu), & \text{when } \mu \geq a. \end{cases} \quad (\text{A.175})$$

Set $RHS(\mu)$ equal to the right-hand side of the above display. Observe that LHS , RHS , and RHS^λ are all continuous functions of μ for $\mu \in (0, \infty)$.

In order to analyze the convergence of prelimit equilibria (i.e., solutions to (A.173)) to limiting equilibria (i.e., solutions to (A.174)), we rely on the uniform convergence of RHS^λ to RHS . Formally, we need the next lemma, whose proof appears at the end.

Lemma 53. $RHS^\lambda(\mu)$ converges uniformly on $[0, \infty)$ to $RHS(\mu)$.

Proof of Theorem 5A (a): $n_c = n_o = 0$ implies that $LHS(\mu) \neq RHS(\mu)$ for all $\mu \in (0, a]$. Moreover, $LHS(0) = c'(0) \neq p = RHS(0)$. Therefore, $LHS(\mu) \neq RHS(\mu)$ for all $\mu \in [0, a]$. Then, due to the uniform convergence of RHS^λ to RHS (from Lemma 53), for all large

enough λ , $LHS(\mu) \neq RHS^\lambda(\mu)$ for all $\mu \in [0, a]$, implying that there does not exist a solution to the prelimit FOC (A.173) in $(0, a]$, i.e., $n_o^\lambda = 0$.

Proof of Theorem 5A (b): Recall, from Lemma 3, that $\mu_{\max}^*(p, v)$ is a finite, constant upper bound on pre-limit equilibria for all λ . $n_u = n_c = 0$ implies that $LHS(\mu) \neq RHS(\mu)$ for all $\mu \in [a, \mu_{\max}^*(p, v)]$. Then, due to the uniform convergence of RHS^λ to RHS (from Lemma 53), for all large enough λ , $LHS(\mu) \neq RHS^\lambda(\mu)$ for all $\mu \in [a, \mu_{\max}^*(p, v)]$, implying that there does not exist a solution to the prelimit FOC (A.173) in $(a, \mu_{\max}^*(p, v)]$, i.e., $n_u^\lambda = 0$.

Proof of Theorem 5A (c): $n_c = 0$ implies that $LHS(a) \neq RHS(a)$. Then, by continuity of LHS and RHS , there exists $\delta \in (0, a]$ such that $LHS(\mu) \neq RHS(\mu)$ for all $\mu \in [a - \delta, a + \delta]$. Then, due to the uniform convergence of RHS^λ to RHS (from Lemma 53), for all large enough λ , $LHS(\mu) \neq RHS^\lambda(\mu)$ for all $\mu \in [a - \delta, a + \delta]$. Moreover, $\lim_{\lambda \rightarrow \infty} \frac{\lambda}{N^\lambda} = a$ implies that for all large enough λ , $\frac{\lambda}{N^\lambda} \in [a - \delta, a + \delta]$. Therefore, for all large enough λ , the critically loaded service rate $\frac{\lambda}{N^\lambda}$ is not a solution to the prelimit FOC (A.173), i.e., $n_c^\lambda = 0$.

Proof of Theorem 5A (d): First, it follows from Theorem 4 (a) and (b)(iii)(iv) that when $p < c'(a)$, $n_u = 1$ implies that (i) $p < p^\dagger(v)$, and (ii) $a = \bar{a}$, or, equivalently, from Definition 17, $v = v^\dagger$. Next, recall, from Definition 15 and Remark 16, that when $p < c'(a)$, $\mu^\dagger \in (a, \frac{3}{2}a)$ is the unique limiting underloaded equilibrium for which $c'(\mu^\dagger)$ (i.e., $LHS(\mu^\dagger)$) and $h(\mu^\dagger)$ (i.e., $RHS(\mu^\dagger)$) are tangent and v^\dagger is the unique value of v for which this phenomenon occurs; see Figure A.5 (I) for an illustration.

In order to prove Theorem 5A (d), it suffices to show that if $\lim_{\lambda \rightarrow \infty} N^\lambda - \frac{\lambda}{a} < 0$, then there exists some $\delta \in (0, \mu^\dagger - a)$ such that $RHS^\lambda(\mu) - RHS(\mu) < 0$ for all $\mu \in [\mu^\dagger - \delta, \mu^\dagger + \delta]$ for all large enough λ . This is because, from Lemma 43 (noting that, when $\mu \geq a$, $c'(\mu)$ and $h(\mu)$ are the same as $LHS(\mu)$ and $RHS(\mu)$ respectively), it would then follow that

- $LHS(\mu) \geq RHS(\mu) > RHS^\lambda(\mu)$ for all $\mu \in [\mu^\dagger - \delta, \mu^\dagger + \delta]$ for all large enough λ ; and
- $LHS(\mu) > RHS(\mu)$ for all $\mu \in [a, \mu^\dagger - \delta] \cup [\mu^\dagger + \delta, \infty)$, implying that, due to the

uniform convergence of RHS^λ to RHS (from Lemma 53), $LHS(\mu) > RHS^\lambda(\mu)$ for all $\mu \in [a, \mu^\dagger - \delta] \cup [\mu^\dagger + \delta, \infty)$ for all large enough λ .

To proceed, we need the following result, whose proof appears at the end.

Claim 14. *If $\lim_{\lambda \rightarrow \infty} N^\lambda - \frac{\lambda}{a} < 0$, then there exists $\delta \in (0, \mu^\dagger - a)$ such that the following hold for all $\mu \in [\mu^\dagger - \delta, \mu^\dagger + \delta]$, where μ^\dagger is defined in Definition 15 (a):*

$$(i) \quad I^\lambda(\mu, \mu) - \left(1 - \frac{a}{\mu}\right) < 0 \text{ for all large enough } \lambda, \text{ and}$$

$$(ii) \quad \lim_{\lambda \rightarrow \infty} \frac{I^\lambda(\mu, \mu)^2 ErlC\left(N^\lambda, \frac{\lambda}{\mu}\right) \sum_{i=1}^{k^\lambda - N^\lambda} i \left(\frac{\lambda}{N^\lambda \mu}\right)^i}{N^\lambda \left(1 - \frac{a}{\mu}\right) - I^\lambda(\mu, \mu)} = 0.$$

Claim 14 guarantees the existence of some $\delta_1 \in (0, \mu^\dagger - a)$ such that for all $\mu \in [\mu^\dagger - \delta_1, \mu^\dagger + \delta_1]$, the statements Claim 14 (i)(ii) are true. In addition, the following statements are also true:

- $\mu^\dagger - \delta_1 > \frac{\lambda}{N^\lambda}$ for all large enough λ , and
- $\mu^\dagger + \delta_1 < 2\mu^\dagger - a < 2\left(\frac{3}{2}a\right) - a = 2a < \frac{v^\dagger}{p}$, where the last two inequalities follow from Lemma 42 and Remark 17, respectively.

In other words, for all large enough λ , in addition to the two statements guaranteed by Claim 14 (i)(ii), it is also true that

$$\frac{\lambda}{N^\lambda} < \mu < \frac{v^\dagger}{p} \quad \forall \mu \in [\mu^\dagger - \delta_1, \mu^\dagger + \delta_1]. \quad (\text{A.176})$$

Next, by definition of RHS^λ and RHS ,

$$RHS^\lambda(\mu) - RHS(\mu) = -p \left[I^\lambda(\mu, \mu) - \left(1 - \frac{a}{\mu}\right) \right] + \left(\frac{v^\dagger}{\mu} - p \right) \left[\mu \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} - \frac{a}{\mu} \left(1 - \frac{a}{\mu}\right) \right]. \quad (\text{A.177})$$

Using (A.21) from Corollary 2, we can evaluate the term multiplying $\left(\frac{v^\dagger}{\mu} - p\right)$ in the above display as follows:

$$\begin{aligned}
& \mu \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} - \frac{a}{\mu} \left(1 - \frac{a}{\mu}\right) \\
&= I^\lambda(\mu, \mu) \left(1 - I^\lambda(\mu, \mu)\right) + I^\lambda(\mu, \mu)^2 \frac{ErlC(N^\lambda, \frac{\lambda}{\mu})}{N^\lambda} \sum_{i=1}^{k^\lambda-N^\lambda} i \left(\frac{\lambda}{N^\lambda \mu}\right)^i - \frac{a}{\mu} \left(1 - \frac{a}{\mu}\right) \\
&= \left[I^\lambda(\mu, \mu) - \left(1 - \frac{a}{\mu}\right)\right] \left(\frac{a}{\mu} - I^\lambda(\mu, \mu)\right) + I^\lambda(\mu, \mu)^2 \frac{ErlC(N^\lambda, \frac{\lambda}{\mu})}{N^\lambda} \sum_{i=1}^{k^\lambda-N^\lambda} i \left(\frac{\lambda}{N^\lambda \mu}\right)^i \\
&= \left[I^\lambda(\mu, \mu) - \left(1 - \frac{a}{\mu}\right)\right] \left[\left(\frac{2a}{\mu} - 1\right) - \left(I^\lambda(\mu, \mu) - \left(1 - \frac{a}{\mu}\right)\right)\right] \\
&\quad + I^\lambda(\mu, \mu)^2 \frac{ErlC(N^\lambda, \frac{\lambda}{\mu})}{N^\lambda} \sum_{i=1}^{k^\lambda-N^\lambda} i \left(\frac{\lambda}{N^\lambda \mu}\right)^i. \tag{A.178}
\end{aligned}$$

Substituting (A.178) into (A.177) yields

$$\begin{aligned}
RHS^\lambda(\mu) - RHS(\mu) &\leq -p \left[I^\lambda(\mu, \mu) - \left(1 - \frac{a}{\mu}\right)\right] \\
&\quad + \left(\frac{v^\dagger}{\mu} - p\right) \left[I^\lambda(\mu, \mu) - \left(1 - \frac{a}{\mu}\right)\right] \left[\left(\frac{2a}{\mu} - 1\right) - \left(I^\lambda(\mu, \mu) - \left(1 - \frac{a}{\mu}\right)\right)\right] \\
&\quad + \left(\frac{v^\dagger}{\mu} - p\right) I^\lambda(\mu, \mu)^2 \frac{ErlC(N^\lambda, \frac{\lambda}{\mu})}{N^\lambda} \sum_{i=1}^{k^\lambda-N^\lambda} i \left(\frac{\lambda}{N^\lambda \mu}\right)^i \\
&= \left[I^\lambda(\mu, \mu) - \left(1 - \frac{a}{\mu}\right)\right] \left[\frac{2a}{\mu} \left(\frac{v^\dagger}{\mu} - \frac{v^\dagger}{2a} - p\right) - \left(\frac{v^\dagger}{\mu} - p\right) \left[I^\lambda(\mu, \mu) - \left(1 - \frac{a}{\mu}\right)\right]\right] \\
&\quad + \left(\frac{v^\dagger}{\mu} - p\right) I^\lambda(\mu, \mu)^2 \frac{ErlC(N^\lambda, \frac{\lambda}{\mu})}{N^\lambda} \sum_{i=1}^{k^\lambda-N^\lambda} i \left(\frac{\lambda}{N^\lambda \mu}\right)^i.
\end{aligned}$$

Claim 14 (i) guarantees that $\left(1 - \frac{a}{\mu}\right) - I^\lambda(\mu, \mu) > 0$; dividing throughout by this term

yields

$$\begin{aligned}
\frac{RHS^\lambda(\mu) - RHS(\mu)}{\left(1 - \frac{a}{\mu}\right) - I^\lambda(\mu, \mu)} &\leq -\frac{2a}{\mu} \left(\frac{v^\dagger}{\mu} - \frac{v^\dagger}{2a} - p \right) + \left(\frac{v^\dagger}{\mu} - p \right) \left[I^\lambda(\mu, \mu) - \left(1 - \frac{a}{\mu}\right) \right] \\
&\quad + \left(\frac{v^\dagger}{\mu} - p \right) \frac{I^\lambda(\mu, \mu)^2 ErlC \left(N^\lambda, \frac{\lambda}{\mu} \right) \sum_{i=1}^{k^\lambda - N^\lambda} i \left(\frac{\lambda}{N^\lambda \mu} \right)^i}{N^\lambda \left(\left(1 - \frac{a}{\mu}\right) - I^\lambda(\mu, \mu) \right)} \\
&\stackrel{(*)}{<} -\frac{2a}{\mu} \left(\frac{v^\dagger}{\mu} - \frac{v^\dagger}{2a} - p \right) + \left(\frac{v^\dagger}{\mu} - p \right) \frac{I^\lambda(\mu, \mu)^2 ErlC \left(N^\lambda, \frac{\lambda}{\mu} \right) \sum_{i=1}^{k^\lambda - N^\lambda} i \left(\frac{\lambda}{N^\lambda \mu} \right)^i}{N^\lambda \left(\left(1 - \frac{a}{\mu}\right) - I^\lambda(\mu, \mu) \right)},
\end{aligned}$$

where $(*)$ follows from Claim 14 (i). Taking the limit as $\lambda \rightarrow \infty$ and applying Claim 14 (ii), we obtain, for all $\mu \in [\mu^\dagger - \delta_1, \mu^\dagger + \delta_1]$,

$$\begin{aligned}
\lim_{\lambda \rightarrow \infty} \frac{RHS^\lambda(\mu) - RHS(\mu)}{\left(1 - \frac{a}{\mu}\right) - I^\lambda(\mu, \mu)} &\leq -\frac{2a}{\mu} \left(\frac{v^\dagger}{\mu} - \frac{v^\dagger}{2a} - p \right) \\
&= -\frac{2av^\dagger}{\mu} \left(\frac{1}{\mu} - \frac{\frac{2v^\dagger}{a} + 2p}{3v^\dagger} \right) - \frac{a}{3\mu} \left(\frac{v^\dagger}{a} - 2p \right) \\
&= -\frac{2av^\dagger}{\mu} f(\mu) - \frac{a}{3} g(\mu),
\end{aligned} \tag{A.179}$$

where $f(\mu) = \frac{1}{\mu} - \frac{\frac{2v^\dagger}{a} + 2p}{3v^\dagger}$ and $g(\mu) = \frac{1}{\mu} \left(\frac{v^\dagger}{a} - 2p \right)$. We resolve the signs of $f(\mu)$ and $g(\mu)$ as follows:

- From Lemma 42, we know that $f(\mu^\dagger) > 0$, which, due to the continuity of f , implies that there exists a small enough $\delta \in (0, \delta_1)$ such that $f(\mu) > 0$ for all $\mu \in [\mu^\dagger - \delta, \mu^\dagger + \delta]$.
- From Remark 17, we know that $v^\dagger > 2ap$, which implies that $g(\mu) > 0$ for all $\mu > 0$.

Therefore, (A.179) implies that

$$\lim_{\lambda \rightarrow \infty} \frac{RHS^\lambda(\mu) - RHS(\mu)}{\left(1 - \frac{a}{\mu}\right) - I^\lambda(\mu, \mu)} < 0, \quad \forall \mu \in [\mu^\dagger - \delta, \mu^\dagger + \delta].$$

Then, using the guarantee from Claim 14 (i) that $\left(1 - \frac{a}{\mu}\right) - I^\lambda(\mu, \mu) > 0$ for all $\mu \in$

$[\mu^\dagger - \delta, \mu^\dagger + \delta]$ for all large enough λ , it follows that

$$RHS^\lambda(\mu) - RHS(\mu) < 0, \quad \forall \mu \in [\mu^\dagger - \delta, \mu^\dagger + \delta] \quad \text{for all large enough } \lambda,$$

as desired.

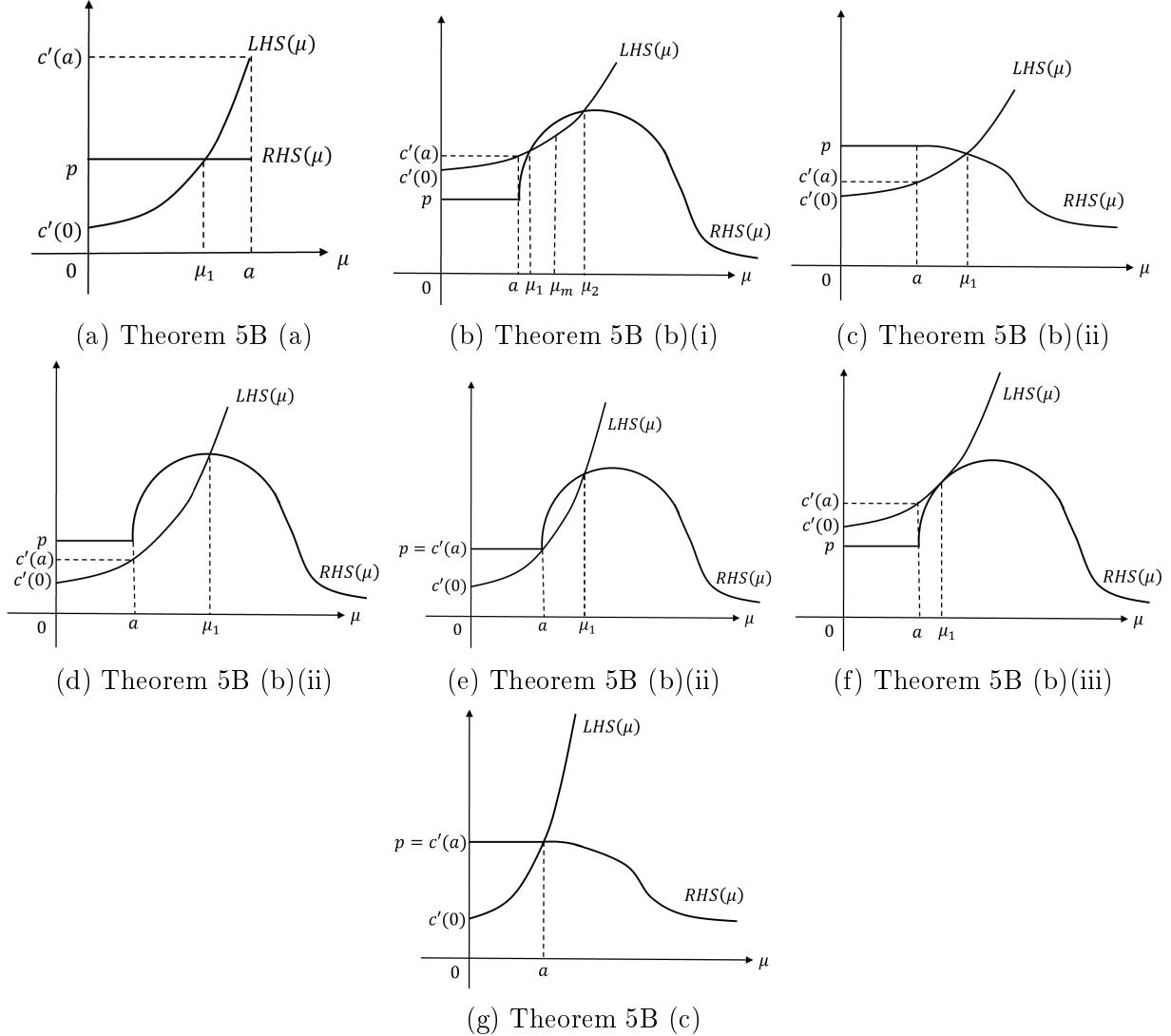


Figure A.13: Existence of the solutions to the prelimit FOC (A.173) in Theorem 5B.

Proof of Theorem 5B (a): From Theorem 4 (a) and (b)(i), $n_o = 1$ if and only if $c'(0) < p < c'(a)$ (Figure A.13 (a)); that is, $LHS(0) = c'(0) < p = RHS(0)$ and $LHS(a) = c'(a) > p = RHS(a)$. Then, due to the uniform convergence of RHS^λ to RHS (from Lemma 53), for

all large enough λ , $LHS(0) < RHS^\lambda(0)$ and $LHS(a) > RHS^\lambda(a)$; the intermediate value theorem then guarantees the existence of at least one solution to the prelimit FOC (A.173) in $(0, a)$, i.e., $n_o^\lambda \geq 1$.

Proof of Theorem 5B (b)(i): From Theorem 4 (a) and (b)(iii)(iv), $n_u = 2$ only if $LHS(a) = c'(a) > p = RHS(a)$ (Figure A.13 (b)). Then, because there is more than one underloaded equilibrium, Lemmas 41 and 43 together imply that LHS and RHS cannot be tangent at any $\mu \in (a, \infty)$. Let $\mu_2 > \mu_1 > a$ be the two distinct limiting underloaded equilibria that solve $LHS(\mu) = RHS(\mu)$. Since $\lim_{\mu \rightarrow \infty} LHS(\mu) = \infty > 0 = \lim_{\mu \rightarrow \infty} RHS(\mu)$, it must be that $LHS(\mu) > RHS(\mu)$ for all $\mu > \mu_2$, and in particular, for some $\bar{\mu} > \mu_2$. Together with $LHS(a) > RHS(a)$, this means that there must exist some $\mu_m \in (\mu_1, \mu_2)$ for which $LHS(\mu_m) < RHS(\mu_m)$. Then, due to the uniform convergence of RHS^λ to RHS (from Lemma 53), for all large enough λ , $LHS(a) > RHS^\lambda(a)$, $LHS(\mu_m) < RHS^\lambda(\mu_m)$, and $LHS(\bar{\mu}) > RHS^\lambda(\bar{\mu})$; the intermediate value theorem then guarantees the existence of at least two solutions to the prelimit FOC (A.173) in $(a, \bar{\mu})$, i.e., $n_u^\lambda \geq 2$.

Proof of Theorem 5B (b)(ii): Here, we are given that $p \geq c'(a)$ and $n_u = 1$ (Figures A.13 (c)-(e)). Let $\mu_1 > a$ denote the unique limiting underloaded equilibrium. Recall, from Lemmas 41 and 43 (noting that, when $\mu \geq a$, $c'(\mu)$ and $h(\mu)$ are the same as $LHS(\mu)$ and $RHS(\mu)$ respectively), that $LHS(\mu)$ and $RHS(\mu)$ are tangent for some $\mu \geq a$ if and only if $p \leq c'(a)$ and $\mu = \mu^\dagger(a, p)$; moreover, when this happens, there are no solutions to $LHS(\mu) = RHS(\mu)$ in $[a, \infty)$ other than μ^\dagger . Also recall, from Remark 16, that $p = c'(a)$ if and only if $\mu^\dagger = a$. Therefore, we can conclude that if $p \geq c'(a)$, then $LHS(\mu)$ and $RHS(\mu)$ are never tangent at any $\mu \in (a, \infty)$, and in particular, at μ_1 . Since $\lim_{\mu \rightarrow \infty} LHS(\mu) = \infty > 0 = \lim_{\mu \rightarrow \infty} RHS(\mu)$, it must be that $LHS(\mu) > RHS(\mu)$ for all $\mu > \mu_1$, and in particular, for some $\bar{\mu} > \mu_1$. Therefore, there exists a small enough $\delta \in (0, \mu_1 - a)$ such that $LHS(\mu_1 - \delta) < RHS(\mu_1 - \delta)$. Then, due to the uniform convergence of RHS^λ to RHS (from Lemma 53), for all large enough λ , $LHS(\mu_1 - \delta) < RHS^\lambda(\mu_1 - \delta)$

and $LHS(\bar{\mu}) > RHS^\lambda(\bar{\mu})$; the intermediate value theorem then guarantees the existence of at least one solution to the prelimit FOC (A.173) in $(\mu_1 - \delta, \bar{\mu})$, i.e., $n_u^\lambda \geq 1$.

Proof of Theorem 5B (b)(iii): The proof technique is similar to that of Theorem 5A (d).

First, it follows from Theorem 4 (a) and (b)(iii)(iv) that when $p < c'(a)$, $n_u = 1$ implies that (i) $p < p^\dagger(v)$, and (ii) $a = \bar{a}$, or, equivalently, from Definition 17, $v = v^\dagger$. Next, recall, from Definition 15 and Remark 16, that when $p < c'(a)$, $\mu^\dagger \in (a, \frac{3}{2}a)$ is the unique limiting underloaded equilibrium for which $c'(\mu^\dagger)$ (i.e., $LHS(\mu^\dagger)$) and $h(\mu^\dagger)$ (i.e., $RHS(\mu^\dagger)$) are tangent and v^\dagger is the unique value of v for which this phenomenon occurs; see Figure A.5 (I) for an illustration. Also recall, from Remark 17, that $v^\dagger > 2ap$.

In order to prove Theorem 5B (b)(iii), it suffices to show that if $N^\lambda - \frac{\lambda}{a} \geq 0$ for all large enough λ , then $RHS^\lambda(\mu^\dagger) - RHS(\mu^\dagger) > 0$ for all large enough λ . This is because, from Lemmas 43 (noting that, when $\mu \geq a$, $c'(\mu)$ and $h(\mu)$ are the same as $LHS(\mu)$ and $RHS(\mu)$ respectively), it would then follow that

- $LHS(\mu) > RHS(\mu)$ for all $\mu > \mu^\dagger$, and in particular, for some $\bar{\mu} > \mu^\dagger$;
- $LHS(\mu^\dagger) = RHS(\mu^\dagger) < RHS^\lambda(\mu^\dagger)$ for all large enough λ ; and
- $LHS(a) = c'(a) > p = RHS(a)$.

Then, due to the uniform convergence of RHS^λ to RHS (from Lemma 53), for all large enough λ , $LHS(a) > RHS^\lambda(a)$ and $LHS(\bar{\mu}) > RHS^\lambda(\bar{\mu})$; together with $LHS(\mu^\dagger) < RHS^\lambda(\mu^\dagger)$, the intermediate value theorem then guarantees the existence of at least two solutions to the prelimit FOC (A.173) in $(a, \bar{\mu})$, i.e., $n_u^\lambda \geq 2$.

First, we observe that $\mu^\dagger < \left(\frac{3}{2}\right)a < \frac{3}{4}\frac{v^\dagger}{p} < \frac{v^\dagger}{p}$, where the first two inequalities follow from Lemma 42 and Remark 17, respectively. In other words, we have

$$\frac{v^\dagger}{\mu^\dagger} - p > 0. \quad (\text{A.180})$$

Next, by definition of RHS^λ and RHS ,

$$RHS^\lambda(\mu^\dagger) - RHS(\mu^\dagger) = -p \left[I^\lambda(\mu^\dagger, \mu^\dagger) - \left(1 - \frac{a}{\mu^\dagger}\right) \right] + \left(\frac{v^\dagger}{\mu^\dagger} - p \right) \left[\mu^\dagger \frac{\partial I^\lambda(\mu_1, \mu^\dagger)}{\partial \mu_1} \Big|_{\mu_1=\mu^\dagger} - \frac{a}{\mu^\dagger} \left(1 - \frac{a}{\mu^\dagger}\right) \right]. \quad (\text{A.181})$$

Using (A.21) from Corollary 2, we can infer that $\mu \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} \geq I^\lambda(\mu, \mu) \left(1 - I^\lambda(\mu, \mu)\right)$ for all $\mu \in (0, \infty)$ and for all λ . This observation can be used to bound the term multiplying $\left(\frac{v^\dagger}{\mu^\dagger} - p\right)$ in the above display as follows:

$$\begin{aligned} \mu^\dagger \frac{\partial I^\lambda(\mu_1, \mu^\dagger)}{\partial \mu_1} \Big|_{\mu_1=\mu^\dagger} - \frac{a}{\mu^\dagger} \left(1 - \frac{a}{\mu^\dagger}\right) &\geq I^\lambda(\mu^\dagger, \mu^\dagger) \left(1 - I^\lambda(\mu^\dagger, \mu^\dagger)\right) - \frac{a}{\mu^\dagger} \left(1 - \frac{a}{\mu^\dagger}\right) \\ &= \left[I^\lambda(\mu^\dagger, \mu^\dagger) - \left(1 - \frac{a}{\mu^\dagger}\right) \right] \left(\frac{a}{\mu^\dagger} - I^\lambda(\mu^\dagger, \mu^\dagger) \right) \\ &= \left[I^\lambda(\mu^\dagger, \mu^\dagger) - \left(1 - \frac{a}{\mu^\dagger}\right) \right] \left[\left(\frac{2a}{\mu^\dagger} - 1 \right) - \left(I^\lambda(\mu^\dagger, \mu^\dagger) - \left(1 - \frac{a}{\mu^\dagger}\right) \right) \right]. \end{aligned} \quad (\text{A.182})$$

Substituting (A.182) into (A.181) and given (A.180), we obtain:

$$\begin{aligned} RHS^\lambda(\mu^\dagger) - RHS(\mu^\dagger) &\geq -p \left[I^\lambda(\mu^\dagger, \mu^\dagger) - \left(1 - \frac{a}{\mu^\dagger}\right) \right] \\ &\quad + \left(\frac{v^\dagger}{\mu^\dagger} - p \right) \left[I^\lambda(\mu^\dagger, \mu^\dagger) - \left(1 - \frac{a}{\mu^\dagger}\right) \right] \left[\left(\frac{2a}{\mu^\dagger} - 1 \right) - \left(I^\lambda(\mu^\dagger, \mu^\dagger) - \left(1 - \frac{a}{\mu^\dagger}\right) \right) \right] \\ &= \left[I^\lambda(\mu^\dagger, \mu^\dagger) - \left(1 - \frac{a}{\mu^\dagger}\right) \right] \left[\frac{2a}{\mu^\dagger} \left(\frac{v^\dagger}{\mu^\dagger} - \frac{v^\dagger}{2a} - p \right) - \left(\frac{v^\dagger}{\mu^\dagger} - p \right) \left[I^\lambda(\mu^\dagger, \mu^\dagger) - \left(1 - \frac{a}{\mu^\dagger}\right) \right] \right]. \end{aligned} \quad (\text{A.183})$$

In what follows, we want to divide both sides by $I^\lambda(\mu^\dagger, \mu^\dagger) - \left(1 - \frac{a}{\mu^\dagger}\right)$. To single out the case when $I^\lambda(\mu^\dagger, \mu^\dagger) - \left(1 - \frac{a}{\mu^\dagger}\right) = 0$, we consider a subsequence on which $k^\lambda = \infty$ for all large enough λ , and a subsequence on which $k^\lambda < \infty$ for all large enough λ , separately.

Case(I): If $k^\lambda = \infty$ for all large enough λ , then Lemma 25 implies that $I^\lambda(\mu^\dagger, \mu^\dagger) - \left(1 - \frac{a}{\mu^\dagger}\right) = 0$ for all large enough λ (when $k^\lambda = \infty$). Moreover, from (A.21) in Corollary 2,

$\mu \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} > I^\lambda(\mu, \mu) \left(1 - I^\lambda(\mu, \mu)\right)$ for all $\mu \in (0, \infty)$ and for all large enough λ (when $k^\lambda = \infty$). This implies that (A.183) holds with strict inequality for all large enough λ . Therefore, $RHS^\lambda(\mu^\dagger) - RHS(\mu^\dagger) > 0$ for all large enough λ .

Case(II): If $k^\lambda < \infty$ for all large enough λ , the proof proceeds as follows. Note that, for all large enough λ ,

$$\begin{aligned} I^\lambda(\mu^\dagger, \mu^\dagger) - \left(1 - \frac{a}{\mu^\dagger}\right) &= I^\lambda(\mu^\dagger, \mu^\dagger) - \left(1 - \frac{\lambda}{N^\lambda \mu^\dagger}\right) + \left(\frac{a}{\mu^\dagger} - \frac{\lambda}{N^\lambda \mu^\dagger}\right) \\ &= I^\lambda(\mu^\dagger, \mu^\dagger) - \left(1 - \frac{\lambda}{N^\lambda \mu^\dagger}\right) + \frac{a}{N^\lambda \mu^\dagger} \left(N^\lambda - \frac{\lambda}{a}\right) > 0, \end{aligned} \quad (\text{A.184})$$

where $I^\lambda(\mu^\dagger, \mu^\dagger) - \left(1 - \frac{\lambda}{N^\lambda \mu^\dagger}\right) \geq 0$ for all λ (from Lemma 25) and $N^\lambda - \frac{\lambda}{a} \geq 0$ for all large enough λ (by assumption). Then, (A.183) implies that, for all large enough λ ,

$$\frac{RHS^\lambda(\mu^\dagger) - RHS(\mu^\dagger)}{I^\lambda(\mu^\dagger, \mu^\dagger) - \left(1 - \frac{a}{\mu^\dagger}\right)} \geq \frac{2a}{\mu^\dagger} \left(\frac{v^\dagger}{\mu^\dagger} - \frac{v^\dagger}{2a} - p\right) - \left(\frac{v^\dagger}{\mu^\dagger} - p\right) \left[I^\lambda(\mu^\dagger, \mu^\dagger) - \left(1 - \frac{a}{\mu^\dagger}\right)\right]$$

Taking the limit as $\lambda \rightarrow \infty$ and applying Lemma 5, we obtain

$$\begin{aligned} \lim_{\lambda \rightarrow \infty} \frac{RHS^\lambda(\mu^\dagger) - RHS(\mu^\dagger)}{I^\lambda(\mu^\dagger, \mu^\dagger) - \left(1 - \frac{a}{\mu^\dagger}\right)} &\geq \frac{2a}{\mu^\dagger} \left(\frac{v^\dagger}{\mu^\dagger} - \frac{v^\dagger}{2a} - p\right) \\ &\stackrel{(i)}{>} \frac{2a}{\mu^\dagger} \left(v^\dagger \cdot \frac{2p + \frac{2v^\dagger}{a}}{3v^\dagger} - \frac{v^\dagger}{2a} - p\right) = \frac{2a}{3\mu^\dagger} \left(\frac{v^\dagger}{2a} - p\right) \\ &\stackrel{(ii)}{>} 0, \end{aligned}$$

where (i) follows from Lemma 42 and (ii) follows from Remark 17. Hence, recalling (A.184), $RHS^\lambda(\mu^\dagger) - RHS(\mu^\dagger) > 0$ for all large enough λ .

Proof of Theorem 5B (c): From Theorem 4 (a) and (b)(ii), $n_c = 1$ if and only if $p = c'(a)$. Then, $LHS(0) = c'(0) < c'(a) = p = RHS(0)$. Moreover, since $\lim_{\mu \rightarrow \infty} RHS(\mu) = \infty > 0 = \lim_{\mu \rightarrow \infty} RHS(\mu)$, it must be that $LHS(\bar{\mu}) > RHS(\bar{\mu})$ for some large enough $\bar{\mu} > a$. Then, due to the uniform convergence of RHS^λ to RHS (from Lemma 53), for all large

enough λ , $LHS(0) < RHS^\lambda(0)$ and $LHS(\bar{\mu}) > RHS^\lambda(\bar{\mu})$; the intermediate value theorem then guarantees the existence of at least one solution to the prelimit FOC (A.173) in $(0, \bar{\mu})$, i.e., $n_u^\lambda + n_c^\lambda + n_o^\lambda \geq 1$. \blacksquare

A.7.13.1 Proof of Lemma 53

It suffices to show that $\lim_{\lambda \rightarrow \infty} \sup_{\mu \in [0, \infty)} |RHS^\lambda(\mu) - RHS(\mu)| = 0$. Recall from (A.173) that

$$\begin{aligned}
& \sup_{\mu \in [0, \infty)} |RHS^\lambda(\mu) - RHS(\mu)| \\
&= \sup_{\mu \in [0, \infty)} \left| -p \cdot I^\lambda(\mu, \mu) + (v - p\mu) \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} + p \left[1 - \frac{a}{\mu} \right]^+ - (v - p\mu) \frac{a}{\mu^2} \left[1 - \frac{a}{\mu} \right]^+ \right| \\
&= \sup_{\mu \in [0, \infty)} \left| -p \left(I^\lambda(\mu, \mu) - \left[1 - \frac{a}{\mu} \right]^+ \right) + (v - p\mu) \left(\frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} - \frac{a}{\mu^2} \left[1 - \frac{a}{\mu} \right]^+ \right) \right| \\
&\leq \sup_{\mu \in [0, \infty)} \left| -p \left(I^\lambda(\mu, \mu) - \left[1 - \frac{a}{\mu} \right]^+ \right) \right| + \left| (v - p\mu) \left(\frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} - \frac{a}{\mu^2} \left[1 - \frac{a}{\mu} \right]^+ \right) \right| \\
&\leq \sup_{\mu \in [0, \infty)} \left| p \left(I^\lambda(\mu, \mu) - \left[1 - \frac{a}{\mu} \right]^+ \right) \right| + |v + p\mu| \left| \left(\frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} - \frac{a}{\mu^2} \left[1 - \frac{a}{\mu} \right]^+ \right) \right| \\
&= \max \left\{ \sup_{\mu \in [0, a]} \left| p \left(I^\lambda(\mu, \mu) - \left[1 - \frac{a}{\mu} \right]^+ \right) \right|, \sup_{\mu \in [a, \infty)} \left| p \left(I^\lambda(\mu, \mu) - \left[1 - \frac{a}{\mu} \right]^+ \right) \right| \right\} \\
&\quad + \max \left\{ \sup_{\mu \in [0, a]} |v + p\mu| \left| \left(\frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} - \frac{a}{\mu^2} \left[1 - \frac{a}{\mu} \right]^+ \right) \right|, \right. \\
&\quad \left. \sup_{\mu \in [a, \infty)} \left| \frac{v}{\mu} + p \right| \left| \left(\mu \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} - \frac{a}{\mu} \left[1 - \frac{a}{\mu} \right]^+ \right) \right| \right\} \\
&\leq p \cdot \max \left\{ \sup_{\mu \in [0, a]} \left| I^\lambda(\mu, \mu) - \left[1 - \frac{a}{\mu} \right]^+ \right|, \sup_{\mu \in [a, \infty)} \left| I^\lambda(\mu, \mu) - \left[1 - \frac{a}{\mu} \right]^+ \right| \right\} \\
&\quad + \max \left\{ (v + pa) \sup_{\mu \in [0, a]} \left| \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} - \frac{a}{\mu^2} \left[1 - \frac{a}{\mu} \right]^+ \right|, \right. \\
&\quad \left. \left(\frac{v}{a} + p \right) \sup_{\mu \in [a, \infty)} \left| \mu \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} - \frac{a}{\mu} \left[1 - \frac{a}{\mu} \right]^+ \right| \right\}.
\end{aligned}$$

Thus, it suffices to show that all four suprema in the above display converge to zero, as

$\lambda \rightarrow \infty$:

$$\begin{aligned}
(A) \lim_{\lambda \rightarrow \infty} \sup_{\mu \in [0, a]} \left| I^\lambda(\mu, \mu) - \left[1 - \frac{a}{\mu} \right]^+ \right| &= 0 \Leftrightarrow \lim_{\lambda \rightarrow \infty} \sup_{\mu \in [0, a]} \left| I^\lambda(\mu, \mu) \right| = 0 \\
(B) \lim_{\lambda \rightarrow \infty} \sup_{\mu \in [0, a]} \left| \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} - \frac{a}{\mu^2} \left[1 - \frac{a}{\mu} \right]^+ \right| &= 0 \Leftrightarrow \lim_{\lambda \rightarrow \infty} \sup_{\mu \in [0, a]} \left| \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} \right| = 0 \\
(C) \lim_{\lambda \rightarrow \infty} \sup_{\mu \in [a, \infty)} \left| I^\lambda(\mu, \mu) - \left[1 - \frac{a}{\mu} \right]^+ \right| &= 0 \Leftrightarrow \lim_{\lambda \rightarrow \infty} \sup_{\mu \in [a, \infty)} \left| I^\lambda(\mu, \mu) - \left(1 - \frac{a}{\mu} \right) \right| = 0 \\
(D) \lim_{\lambda \rightarrow \infty} \sup_{\mu \in [a, \infty)} \left| \mu \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} - \frac{a}{\mu} \left[1 - \frac{a}{\mu} \right]^+ \right| &= 0 \Leftrightarrow \lim_{\lambda \rightarrow \infty} \sup_{\mu \in [a, \infty)} \left| \mu \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} - \frac{a}{\mu} \left(1 - \frac{a}{\mu} \right) \right| = 0.
\end{aligned}$$

Proof of (A):

$$\lim_{\lambda \rightarrow \infty} \sup_{\mu \in [0, a]} \left| I^\lambda(\mu, \mu) \right| \stackrel{(i)}{=} \lim_{\lambda \rightarrow \infty} \sup_{\mu \in [0, a]} I^\lambda(\mu, \mu) \stackrel{(ii)}{=} \lim_{\lambda \rightarrow \infty} I^\lambda(a, a) \stackrel{(iii)}{=} 0,$$

where (i) follows from the fact that $I^\lambda(\mu, \mu) \geq 0$ for all $\lambda, \mu > 0$; (ii) follows because $I^\lambda(\mu, \mu) \geq 0$ is strictly increasing in $\mu \in (0, \infty)$ for all λ by applying Lemma 21 (a) (N times iterating on each μ_i , $i \in [N]$); and (iii) follows from Lemma 5.

Proof of (B): The notation $I^\lambda(\mu_1, \mu)$ is a shorthand for $I(\mu_1, \mu; \lambda, k^\lambda, N^\lambda)$.

$$\begin{aligned}
&\lim_{\lambda \rightarrow \infty} \sup_{\mu \in [0, a]} \left| \frac{\partial I(\mu_1, \mu; \lambda, k^\lambda, N^\lambda)}{\partial \mu_1} \Big|_{\mu_1=\mu} \right| \\
&\stackrel{(i)}{=} \lim_{\lambda \rightarrow \infty} \sup_{\mu \in [0, a]} \frac{\partial I(\mu_1, \mu; \lambda, k^\lambda, N^\lambda)}{\partial \mu_1} \Big|_{\mu_1=\mu} \\
&\stackrel{(ii)}{\leq} \lim_{\lambda \rightarrow \infty} \sup_{\mu \in [0, a]} \frac{I(\mu, \mu; \lambda, k^\lambda, N^\lambda)}{\mu} + \frac{2\sqrt{N^\lambda}}{\lambda} \\
&\stackrel{(iii)}{\leq} \lim_{\lambda \rightarrow \infty} \sup_{\mu \in [0, a]} \frac{I(\mu, \mu; \lambda, N^\lambda, N^\lambda)}{\mu} + \frac{2\sqrt{N^\lambda}}{\lambda} \\
&\stackrel{(iv)}{\leq} \lim_{\lambda \rightarrow \infty} \frac{I(a, a; \lambda, N^\lambda, N^\lambda)}{a} + \frac{2\sqrt{N^\lambda}}{\lambda} \\
&\stackrel{(v)}{=} 0,
\end{aligned}$$

where (i) follows from the fact that $\frac{\partial I(\mu_1, \mu; \lambda, k^\lambda, N^\lambda)}{\partial \mu_1} \Big|_{\mu_1=\mu} \geq 0$ for all $\lambda, \mu > 0$ (Lemma 21 (a)),

(ii) follows from Corollary 3 (b), (iii) follows from Lemma 21 (b), (iv) follows from Lemma 27, noting that, for all large enough λ , $a < \frac{5\lambda}{4N^\lambda}$ and $N^\lambda \geq 6$, and (v) follows from Lemma 5 and because $\frac{\sqrt{N^\lambda}}{\lambda} \rightarrow 0$ as $\lambda \rightarrow \infty$.

Hence,

$$\lim_{\lambda \rightarrow \infty} \sup_{\mu \in [0, a]} \left| \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} \right| = 0.$$

Proof of (C): Consider a subsequence λ' on which $\lim_{\lambda' \rightarrow \infty} \operatorname{sgn} \left(N^{\lambda'} - \frac{\lambda'}{a} \right)$ is either ≥ 0 or < 0 . We simply use λ rather than λ' to denote the subsequence. We discuss the following two cases based on $\lim_{\lambda \rightarrow \infty} \operatorname{sgn} \left(N^\lambda - \frac{\lambda}{a} \right)$. For ease of presentation, let $f^\lambda(\mu) := I^\lambda(\mu, \mu) - \left(1 - \frac{a}{\mu} \right)$ and $g^\lambda(\mu) := I^\lambda(\mu, \mu) - \left(1 - \frac{\lambda}{N^\lambda \mu} \right)$ for $\mu \in (0, \infty)$.

Case (I): Suppose $\lim_{\lambda \rightarrow \infty} \operatorname{sgn} \left(N^\lambda - \frac{\lambda}{a} \right) \geq 0$, i.e., $N^\lambda - \frac{\lambda}{a} \geq 0$ for all $\lambda > \Lambda$ for some $\Lambda \in (0, \infty)$, then it follows that, for all $\mu \in (0, \infty)$,

$$f^\lambda(\mu) - g^\lambda(\mu) = \frac{a}{\mu} - \frac{\lambda}{N^\lambda \mu} = \frac{N^\lambda - \frac{\lambda}{a}}{N^\lambda} \frac{a}{\mu} \geq 0, \quad \forall \lambda > \Lambda,$$

which implies that

$$f^\lambda(\mu) \geq g^\lambda(\mu) \geq 0, \quad \forall \lambda > \Lambda,$$

where the last inequality follows from Lemma 25 for all λ . In addition, note that

$$(f^\lambda(\mu))' - (g^\lambda(\mu))' = -\frac{a}{\mu^2} + \frac{\lambda}{N^\lambda \mu^2} = -\frac{N^\lambda - \frac{\lambda}{a}}{N^\lambda} \frac{a}{\mu^2} \leq 0, \quad \forall \lambda > \Lambda,$$

which implies that

$$(f^\lambda(\mu))' \leq (g^\lambda(\mu))' \leq 0, \quad \forall \lambda > \Lambda,$$

where the last inequality follows from Lemma 25 for all λ .

In summary, for all $\lambda > \Lambda$, $f^\lambda(\mu) \geq 0$ and is (not necessarily strictly) decreasing in μ for $\mu \in (0, \infty)$. Hence,

$$\sup_{\mu \in [a, \infty)} \left| I^\lambda(\mu, \mu) - \left(1 - \frac{a}{\mu}\right) \right| = \sup_{\mu \in [a, \infty)} \left| f^\lambda(\mu) \right| = f^\lambda(a) = I^\lambda(a, a) - \left(1 - \frac{a}{a}\right),$$

which implies that

$$\lim_{\lambda \rightarrow \infty} \sup_{\mu \in [a, \infty)} \left| I^\lambda(\mu, \mu) - \left(1 - \frac{a}{\mu}\right) \right| = \lim_{\lambda \rightarrow \infty} I^\lambda(a, a) = 0,$$

where the last step follows from Lemma 5.

Case (II): Suppose $\lim_{\lambda \rightarrow \infty} \operatorname{sgn} \left(N^\lambda - \frac{\lambda}{a} \right) < 0$, i.e., $N^\lambda - \frac{\lambda}{a} < 0$ for all $\lambda > \Lambda$ for some $\Lambda \in (0, \infty)$. For all $\lambda > \Lambda$,

$$\begin{aligned} \sup_{\mu \in [a, \infty)} \left| I^\lambda(\mu, \mu) - \left(1 - \frac{a}{\mu}\right) \right| &= \sup_{\mu \in [a, \infty)} \left| g^\lambda(\mu) + \left(1 - \frac{\lambda}{N^\lambda \mu}\right) - \left(1 - \frac{a}{\mu}\right) \right| \\ &= \sup_{\mu \in [a, \infty)} \left| g^\lambda(\mu) + \frac{1}{\mu} \left(a - \frac{\lambda}{N^\lambda} \right) \right| \\ &\leq \sup_{\mu \in [a, \infty)} \left| g^\lambda(\mu) \right| + \sup_{\mu \in [a, \infty)} \frac{1}{\mu} \left| a - \frac{\lambda}{N^\lambda} \right| \\ &= g^\lambda(a) + \frac{1}{a} \left(\frac{\lambda}{N^\lambda} - a \right), \end{aligned}$$

where the last inequality follows because $g^\lambda(\mu)$ is (not necessarily strictly) decreasing in $\mu \in (0, \infty)$ for all λ from Lemma 25, $\frac{1}{\mu}$ is strictly decreasing in μ , and also $a < \frac{\lambda}{N^\lambda}$ for all $\lambda > \Lambda$. Then,

$$\lim_{\lambda \rightarrow \infty} g^\lambda(a) + \frac{1}{a} \left(\frac{\lambda}{N^\lambda} - a \right) = \lim_{\lambda \rightarrow \infty} I^\lambda(a, a) - \left(1 - \frac{\lambda}{N^\lambda a}\right) + \left(\frac{\lambda}{N^\lambda a} - 1 \right) = 0,$$

from Lemma 5 and because $\lim_{\lambda \rightarrow \infty} \frac{\lambda}{N^\lambda} = a$.

Hence, we conclude

$$\lim_{\lambda \rightarrow \infty} \sup_{\mu \in [a, \infty)} \left| I^\lambda(\mu, \mu) - \left(1 - \frac{a}{\mu}\right) \right| = 0. \quad (\text{A.185})$$

Proof of (D):

$$\begin{aligned}
& \sup_{\mu \in [a, \infty)} \left| \mu \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} - \frac{a}{\mu} \left(1 - \frac{a}{\mu} \right) \right| \\
& \stackrel{(i)}{=} \sup_{\mu \in [a, \infty)} \left| I^\lambda(\mu, \mu) \left(1 - I^\lambda(\mu, \mu) \right) + I^\lambda(\mu, \mu)^2 \frac{ErlC \left(N^\lambda, \frac{\lambda}{\mu} \right)}{N^\lambda} \sum_{i=1}^{k^\lambda - N^\lambda} i \left(\frac{\lambda}{N^\lambda \mu} \right)^i - \frac{a}{\mu} \left(1 - \frac{a}{\mu} \right) \right| \\
& \stackrel{(ii)}{\leq} \sup_{\mu \in [a, \infty)} \left| I^\lambda(\mu, \mu)^2 \frac{ErlC \left(N^\lambda, \frac{\lambda}{\mu} \right)}{N^\lambda} \sum_{i=1}^{k^\lambda - N^\lambda} i \left(\frac{\lambda}{N^\lambda \mu} \right)^i \right| + \left| I^\lambda(\mu, \mu) \left(1 - I^\lambda(\mu, \mu) \right) - \frac{a}{\mu} \left(1 - \frac{a}{\mu} \right) \right| \\
& = \sup_{\mu \in [a, \infty)} I^\lambda(\mu, \mu)^2 \frac{ErlC \left(N^\lambda, \frac{\lambda}{\mu} \right)}{N^\lambda} \sum_{i=1}^{k^\lambda - N^\lambda} i \left(\frac{\lambda}{N^\lambda \mu} \right)^i + \left| I^\lambda(\mu, \mu) \left(1 - I^\lambda(\mu, \mu) \right) - \frac{a}{\mu} \left(1 - \frac{a}{\mu} \right) \right| \\
& \stackrel{(iii)}{\leq} \sup_{\mu \in [a, \infty)} \frac{2\sqrt{N^\lambda}}{N^\lambda} + \left| I^\lambda(\mu, \mu) \left(1 - I^\lambda(\mu, \mu) \right) - \frac{a}{\mu} \left(1 - \frac{a}{\mu} \right) \right| \\
& = \frac{2}{\sqrt{N^\lambda}} + \sup_{\mu \in [a, \infty)} \left| I^\lambda(\mu, \mu) \left(1 - I^\lambda(\mu, \mu) \right) - \frac{a}{\mu} \left(1 - \frac{a}{\mu} \right) \right|,
\end{aligned}$$

where (i) follows from (A.21) in Corollary 2, (ii) follows from the triangle inequality, and (iii) follows from Lemma 23 (a). Since $\frac{2}{\sqrt{N^\lambda}} \rightarrow 0$ as $\lambda \rightarrow \infty$, it suffices to show that the second term in the above display converges to 0 as $\lambda \rightarrow \infty$. From (A.185),

$$\lim_{\lambda \rightarrow \infty} \sup_{\mu \in [a, \infty)} \left| I^\lambda(\mu, \mu) - \left(1 - \frac{a}{\mu} \right) \right| = 0.$$

Let $\zeta(x) := x(1-x)$ for $x \in (0, 1)$. Since $\zeta(x)$ is a continuous function,

$$\lim_{\lambda \rightarrow \infty} \sup_{\mu \in [a, \infty)} \left| \zeta \left(I^\lambda(\mu, \mu) \right) - \zeta \left(1 - \frac{a}{\mu} \right) \right| = 0,$$

which can be equivalently written as

$$\lim_{\lambda \rightarrow \infty} \sup_{\mu \in [a, \infty)} \left| I^\lambda \left(1 - I^\lambda(\mu, \mu) \right) - \frac{a}{\mu} \left(1 - \frac{a}{\mu} \right) \right| = \infty,$$

as required. Therefore,

$$\begin{aligned} & \lim_{\lambda \rightarrow \infty} \sup_{\mu \in [a, \infty)} \left| \mu \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} - \frac{a}{\mu} \left(1 - \frac{a}{\mu} \right) \right| \\ & \leq \lim_{\lambda \rightarrow \infty} \left\{ \frac{2}{\sqrt{N^\lambda}} + \sup_{\mu \in [a, \infty)} \left| I^\lambda(\mu, \mu) \left(1 - I^\lambda(\mu, \mu) \right) - \frac{a}{\mu} \left(1 - \frac{a}{\mu} \right) \right| \right\} = 0, \end{aligned}$$

noting that $N^\lambda \rightarrow \infty$ as $\lambda \rightarrow \infty$.

Hence,

$$\lim_{\lambda \rightarrow \infty} \sup_{\mu \in [a, \infty)} \left| \mu \frac{\partial I^\lambda(\mu_1, \mu)}{\partial \mu_1} \Big|_{\mu_1=\mu} - \frac{a}{\mu} \left(1 - \frac{a}{\mu} \right) \right| = 0.$$

■

A.7.13.2 Proof of Claim 14

(i): By definition of μ^\dagger (in Definition 15 (a)), $\mu^\dagger > a$. Then, there exists $\delta \in (0, \mu^\dagger - a)$ such that $\mu > a$ for all $\mu \in [\mu^\dagger - \delta, \mu^\dagger + \delta]$. Then, for all large enough λ , $\mu > \frac{\lambda}{N^\lambda}$ for all $\mu \in [\mu^\dagger - \delta, \mu^\dagger + \delta]$.

Note that

$$\begin{aligned} I^\lambda(\mu, \mu) - \left(1 - \frac{a}{\mu} \right) &= I^\lambda(\mu, \mu) - \left(1 - \frac{\lambda}{N^\lambda \mu} \right) + \left(\frac{a}{\mu} - \frac{\lambda}{N^\lambda \mu} \right) \\ &= I^\lambda(\mu, \mu) - \left(1 - \frac{\lambda}{N^\lambda \mu} \right) + \frac{a}{N^\lambda \mu} \left(N^\lambda - \frac{\lambda}{a} \right) \\ &\stackrel{(i)}{=} \frac{\left(1 - \frac{\lambda}{N^\lambda \mu} \right)}{1 - ErlC \left(N^\lambda, \frac{\lambda}{\mu} \right) \left(\frac{\lambda}{N^\lambda \mu} \right)^{k^\lambda - N^\lambda + 1}} - \left(1 - \frac{\lambda}{N^\lambda \mu} \right) + \frac{a}{N^\lambda \mu} \left(N^\lambda - \frac{\lambda}{a} \right) \\ &= \frac{\left(1 - \frac{\lambda}{N^\lambda \mu} \right)}{1 - ErlC \left(N^\lambda, \frac{\lambda}{\mu} \right) \left(\frac{\lambda}{N^\lambda \mu} \right)^{k^\lambda - N^\lambda + 1}} ErlC \left(N^\lambda, \frac{\lambda}{\mu} \right) \left(\frac{\lambda}{N^\lambda \mu} \right)^{k^\lambda - N^\lambda + 1} + \frac{a}{N^\lambda \mu} \left(N^\lambda - \frac{\lambda}{a} \right) \\ &\stackrel{(ii)}{=} I^\lambda(\mu, \mu) ErlC \left(N^\lambda, \frac{\lambda}{\mu} \right) \left(\frac{\lambda}{N^\lambda \mu} \right)^{k^\lambda - N^\lambda + 1} + \frac{a}{N^\lambda \mu} \left(N^\lambda - \frac{\lambda}{a} \right), \end{aligned}$$

where (i) and (ii) follow from (A.20) in Corollary 2. Thus, for all large enough λ , for all $\mu \in [\mu^\dagger - \delta, \mu^\dagger + \delta]$,

$$\begin{aligned} I^\lambda(\mu, \mu) - \left(1 - \frac{a}{\mu}\right) &= I^\lambda(\mu, \mu) \text{ErlC} \left(N^\lambda, \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{N^\lambda \mu}\right)^{k^\lambda - N^\lambda + 1} + \frac{a}{N^\lambda \mu} \left(N^\lambda - \frac{\lambda}{a}\right) \\ &\stackrel{(*)}{\leq} I^\lambda(\mu, \mu) \text{ErlC} \left(N^\lambda, \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{N^\lambda \mu}\right) + \frac{a}{N^\lambda \mu} \left(N^\lambda - \frac{\lambda}{a}\right) \\ &= \frac{1}{N^\lambda} \left(\frac{\lambda}{a} - N^\lambda\right) \left[I^\lambda(\mu, \mu) \frac{\lambda}{N^\lambda \mu} \frac{\text{ErlC} \left(N^\lambda, \frac{\lambda}{\mu}\right)}{\frac{1}{N^\lambda} \left(\frac{\lambda}{a} - N^\lambda\right)} - \frac{a}{\mu} \right], \end{aligned} \quad (\text{A.186})$$

where (*) follows from the fact that $k^\lambda \geq N^\lambda$ and $\frac{\lambda}{N^\lambda \mu} < 1$ for all $\mu \in [\mu^\dagger - \delta, \mu^\dagger + \delta]$, for all large enough λ . Since $\frac{\lambda}{a} - N^\lambda > 0$ for all large enough λ (by assumption), it suffices to show that $I^\lambda(\mu, \mu) \frac{\lambda}{N^\lambda \mu} \frac{\text{ErlC} \left(N^\lambda, \frac{\lambda}{\mu}\right)}{\frac{1}{N^\lambda} \left(\frac{\lambda}{a} - N^\lambda\right)} < \frac{a}{\mu}$ for all large enough λ , for which, in turn, it suffices to show that $\lim_{\lambda \rightarrow \infty} I^\lambda(\mu, \mu) \frac{\lambda}{N^\lambda \mu} \frac{\text{ErlC} \left(N^\lambda, \frac{\lambda}{\mu}\right)}{\frac{1}{N^\lambda} \left(\frac{\lambda}{a} - N^\lambda\right)} < \frac{a}{\mu}$. Note that

$$\begin{aligned} \lim_{\lambda \rightarrow \infty} I^\lambda(\mu, \mu) \frac{\lambda}{N^\lambda \mu} \frac{\text{ErlC} \left(N^\lambda, \frac{\lambda}{\mu}\right)}{\frac{1}{N^\lambda} \left(\frac{\lambda}{a} - N^\lambda\right)} &= \frac{a}{\mu} \left(1 - \frac{a}{\mu}\right) \left(\lim_{\lambda \rightarrow \infty} \frac{\text{ErlC} \left(N^\lambda, \frac{\lambda}{\mu}\right)}{\frac{1}{N^\lambda} \left(\frac{\lambda}{a} - N^\lambda\right)}\right) < \frac{a}{\mu} \\ \Leftrightarrow \lim_{\lambda \rightarrow \infty} \frac{\frac{\lambda}{a} - N^\lambda}{N^\lambda \text{ErlC} \left(N^\lambda, \frac{\lambda}{\mu}\right)} &> 1 - \frac{a}{\mu}, \end{aligned} \quad (\text{A.187})$$

which is true because the left-hand side is $+\infty$, noting that $N^\lambda \text{ErlC} \left(N^\lambda, \frac{\lambda}{\mu}\right) \rightarrow 0$ for all $\mu \geq \mu^\dagger - \delta > a$ (from Lemma 30 (b)) and $\lim_{\lambda \rightarrow \infty} \frac{\lambda}{a} - N^\lambda > 0$ (by assumption).

(ii): From (A.186), for all $\mu \in [\mu^\dagger - \delta, \mu^\dagger + \delta]$ and for all large enough λ ,

$$\begin{aligned} \left(1 - \frac{a}{\mu}\right) - I^\lambda(\mu, \mu) &\geq \frac{a}{N^\lambda \mu} \left(\frac{\lambda}{a} - N^\lambda\right) - I^\lambda(\mu, \mu) \text{ErlC} \left(N^\lambda, \frac{\lambda}{\mu}\right) \frac{\lambda}{N^\lambda \mu} \\ \Leftrightarrow \frac{I^\lambda(\mu, \mu)^2 \text{ErlC} \left(N^\lambda, \frac{\lambda}{\mu}\right) \sum_{i=1}^{k^\lambda - N^\lambda} i \left(\frac{\lambda}{N^\lambda \mu}\right)^i}{N^\lambda \left(\left(1 - \frac{a}{\mu}\right) - I^\lambda(\mu, \mu)\right)} &\leq \frac{I^\lambda(\mu, \mu)^2 \text{ErlC} \left(N^\lambda, \frac{\lambda}{\mu}\right) \sum_{i=1}^{k^\lambda - N^\lambda} i \left(\frac{\lambda}{N^\lambda \mu}\right)^i}{\frac{a}{\mu} \left(\frac{\lambda}{a} - N^\lambda\right) - I^\lambda(\mu, \mu) N^\lambda \text{ErlC} \left(N^\lambda, \frac{\lambda}{\mu}\right) \frac{\lambda}{N^\lambda \mu}}. \end{aligned} \quad (\text{A.188})$$

Taking the limit as $\lambda \rightarrow \infty$ and observing that $\lim_{\lambda \rightarrow \infty} \frac{\lambda}{N^\lambda \mu} = \frac{a}{\mu}$, we obtain, for all $\mu \in [\mu^\dagger - \delta, \mu^\dagger + \delta]$,

$$\begin{aligned} \lim_{\lambda \rightarrow \infty} \frac{I^\lambda(\mu, \mu)^2 ErlC(N^\lambda, \frac{\lambda}{\mu}) \sum_{i=1}^{k^\lambda - N^\lambda} i \left(\frac{\lambda}{N^\lambda \mu} \right)^i}{N^\lambda \left(\left(1 - \frac{a}{\mu} \right) - I^\lambda(\mu, \mu) \right)} &\leq \lim_{\lambda \rightarrow \infty} \frac{\frac{\mu}{a} I^\lambda(\mu, \mu)^2 ErlC(N^\lambda, \frac{\lambda}{\mu}) \sum_{i=1}^{k^\lambda - N^\lambda} i \left(\frac{\lambda}{N^\lambda \mu} \right)^i}{\left(\frac{\lambda}{a} - N^\lambda \right) - I^\lambda(\mu, \mu) N^\lambda ErlC(N^\lambda, \frac{\lambda}{\mu})} \\ &\stackrel{(*)}{=} \frac{\left(\lim_{\lambda \rightarrow \infty} \frac{1}{N^\lambda} \right) \left(\frac{\mu}{a} \left(1 - \frac{a}{\mu} \right)^2 \lim_{\lambda \rightarrow \infty} \sum_{i=1}^{k^\lambda - N^\lambda} i \left(\frac{\lambda}{N^\lambda \mu} \right)^i \right)}{\left(\lim_{\lambda \rightarrow \infty} \frac{\frac{\lambda}{a} - N^\lambda}{N^\lambda ErlC(N^\lambda, \frac{\lambda}{\mu})} - \left(1 - \frac{a}{\mu} \right) \right)} \\ &= \frac{T_1 T_2}{T_3}, \end{aligned} \quad (\text{A.189})$$

where $(*)$ follows because $I^\lambda(\mu, \mu) \rightarrow 1 - \frac{a}{\mu}$ as $\lambda \rightarrow \infty$ (from Lemma 5), $T_1 = \lim_{\lambda \rightarrow \infty} \frac{1}{N^\lambda}$, $T_2 = \frac{\mu}{a} \left(1 - \frac{a}{\mu} \right)^2 \lim_{\lambda \rightarrow \infty} \sum_{i=1}^{k^\lambda - N^\lambda} i \left(\frac{\lambda}{N^\lambda \mu} \right)^i$, and $T_3 = \lim_{\lambda \rightarrow \infty} \frac{\frac{\lambda}{a} - N^\lambda}{N^\lambda ErlC(N^\lambda, \frac{\lambda}{\mu})} - \left(1 - \frac{a}{\mu} \right)$.

Clearly, $T_3 > 0$ from (A.187) and $T_1 = 0$. Using the finite summation formula,

$$\begin{aligned} 0 \leq T_2 &= \frac{\mu}{a} \left(1 - \frac{a}{\mu} \right)^2 \lim_{\lambda \rightarrow \infty} \frac{\frac{\lambda}{N^\lambda \mu}}{\left(1 - \frac{\lambda}{N^\lambda \mu} \right)^2} \left[1 - \left(\frac{\lambda}{N^\lambda \mu} \right)^{k^\lambda - N^\lambda} \left(1 + (k^\lambda - N^\lambda) \left(1 - \frac{\lambda}{N^\lambda \mu} \right) \right) \right] \\ &= \lim_{\lambda \rightarrow \infty} \left[1 - \left(\frac{\lambda}{N^\lambda \mu} \right)^{k^\lambda - N^\lambda} \left(1 + (k^\lambda - N^\lambda) \left(1 - \frac{\lambda}{N^\lambda \mu} \right) \right) \right] \\ &< \infty, \end{aligned}$$

since $\left(\frac{\lambda}{N^\lambda \mu} \right)^{k^\lambda - N^\lambda} < 1$ for all large enough λ and $\lim_{\lambda \rightarrow \infty} (k^\lambda - N^\lambda) \left(\frac{\lambda}{N^\lambda \mu} \right)^{k^\lambda - N^\lambda + 1} < \infty$ because, even if $\lim_{\lambda \rightarrow \infty} k^\lambda - N^\lambda = \infty$, exponential decay would dominate linear growth in terms of $k^\lambda - N^\lambda$. As a result, (A.189) implies that

$$\lim_{\lambda \rightarrow \infty} \frac{I^\lambda(\mu, \mu)^2 ErlC(N^\lambda, \frac{\lambda}{\mu}) \sum_{i=1}^{k^\lambda - N^\lambda} i \left(\frac{\lambda}{N^\lambda \mu} \right)^i}{N^\lambda \left(\left(1 - \frac{a}{\mu} \right) - I^\lambda(\mu, \mu) \right)} \leq 0.$$

However, the left-hand side is non-negative, since $I^\lambda(\mu, \mu) - \left(1 - \frac{a}{\mu} \right) < 0$ for all large enough

λ from Claim 14 (i). Therefore, it must be that

$$\lim_{\lambda \rightarrow \infty} \frac{I^\lambda(\mu, \mu)^2 ErlC \left(N^\lambda, \frac{\lambda}{\mu} \right) \sum_{i=1}^{k^\lambda - N^\lambda} i \left(\frac{\lambda}{N^\lambda \mu} \right)^i}{N^\lambda \left(\left(1 - \frac{a}{\mu} \right) - I^\lambda(\mu, \mu) \right)} = 0.$$

■

A.8 Proofs from Section 1.6

A.8.1 Proof of Proposition 9

Under utility function (1.20), the FOC (1.6), when $k = N$, $p = 0$, $v = 1$ and $\mu_1 = \mu$, is given by

$$\mu c'(\mu) = \alpha I(\mu, \mu)^{\alpha-1} \left. \frac{\partial I(\mu_1, \mu)}{\partial \mu_1} \right|_{\mu_1=\mu} = \alpha I(\mu, \mu)^\alpha (1 - I(\mu, \mu)). \quad (\text{A.190})$$

recalling Corollary 4 (b). Note that the derivative of the right-hand side of (A.190) is

$$\alpha I(\mu, \mu)^{\alpha-1} (I(\mu, \mu))' (\alpha(1 - I(\mu, \mu)) - I(\mu, \mu)).$$

Since $(I(\mu, \mu))' > 0$ (either by directly calculating $\frac{dI(\mu, \mu)}{d\mu}$ using Corollary 4 (a) or by applying Lemma 21 (a) twice), the above display is strictly positive when $I(\mu, \mu) \in (0, \frac{\alpha}{\alpha+1})$ and strictly negative when $I(\mu, \mu) \in (\frac{\alpha}{\alpha+1}, 1)$. This implies that the right-hand side of (A.190) is maximized when $I(\mu, \mu) = \frac{\alpha}{\alpha+1}$, and the associated maximum value is $\left(\frac{\alpha}{\alpha+1} \right)^{\alpha+1}$.

Moreover, since the left-hand side of (A.190), $\mu c'(\mu)$, is a strictly increasing function of μ (recalling that c is strictly convex), the maximum candidate server equilibrium μ_{\max}^* , which satisfies (A.190), is attained when $\mu_{\max}^* c'(\mu_{\max}^*) = \left(\frac{\alpha}{\alpha+1} \right)^{\alpha+1}$, that is, μ_{\max}^* is the unique solution to

$$\mu c'(\mu) = \left(\frac{\alpha}{\alpha+1} \right)^{\alpha+1},$$

and satisfies

$$I(\mu_{\max}^{*?}, \mu_{\max}^{*?}) = \frac{\alpha}{\alpha + 1}. \quad (\text{A.191})$$

To complete the proof, it remains to be shown that $\mu_1 = \mu_{\max}^{*?}$ is a local maximum of $U(\mu_1, \mu_{\max}^{*?})$ for all $\alpha > 0$ and a global maximum when $\alpha \in (0, 1]$. For this, we investigate the second partial derivative of the generalized utility function $U(\mu_1, \mu)$ with respect to μ_1 . For ease of presentation, we denote $I(\mu_1, \mu)$, $\frac{\partial I(\mu_1, \mu)}{\partial \mu_1}$ and $\frac{\partial^2 I(\mu_1, \mu)}{\partial \mu_1^2}$ simply by I , I' and I'' .

Note that

$$(I^\alpha)'' = (\alpha I^{\alpha-1} I')' = \alpha(\alpha - 1) I^{\alpha-2} (I')^2 + \alpha I^{\alpha-1} I'' = \alpha I^{\alpha-2} \left\{ (\alpha - 1) (I')^2 + II'' \right\}. \quad (\text{A.192})$$

From Lemma 26 (b)(c),

$$I' = \frac{I(1 - I)}{\mu_1} \quad \text{and} \quad I'' = -\frac{2I^2(1 - I)}{\mu_1^2} = -\frac{2II'}{\mu_1}.$$

Substituting these expressions for I' and I'' into (A.192), we get

$$(I^\alpha)'' = \alpha I^{\alpha-2} I' \left((\alpha - 1) \frac{I(1 - I)}{\mu_1} - \frac{2I^2}{\mu_1} \right) = -\frac{\alpha(\alpha + 1) I^{\alpha-1} I'}{\mu_1} \left(I - \frac{\alpha - 1}{\alpha + 1} \right). \quad (\text{A.193})$$

From (A.191) and (A.193), it follows that

$$\begin{aligned} & \frac{\partial^2}{\partial \mu_1^2} \left(I(\mu_1, \mu_{\max}^{*?})^\alpha \right) \Big|_{\mu_1=\mu_{\max}^{*?}} \\ &= -\frac{\alpha(\alpha + 1) I(\mu_{\max}^{*?}, \mu_{\max}^{*?})^{\alpha-1} I'(\mu_{\max}^{*?}, \mu_{\max}^{*?})}{\mu_1} \left(I(\mu_{\max}^{*?}, \mu_{\max}^{*?}) - \frac{\alpha - 1}{\alpha + 1} \right) \\ &= -\frac{\alpha(\alpha + 1) I(\mu_{\max}^{*?}, \mu_{\max}^{*?})^{\alpha-1} I'(\mu_{\max}^{*?}, \mu_{\max}^{*?})}{\mu_1} \left(\frac{\alpha}{\alpha + 1} - \frac{\alpha - 1}{\alpha + 1} \right) < 0, \quad \forall \alpha > 0. \end{aligned}$$

Recalling that $c''(\mu) > 0$ for all $\mu > 0$ (since c is strictly convex),

$$\frac{\partial^2 U(\mu_1, \mu_{\max}^{*?})}{\partial \mu_1^2} \Big|_{\mu_1=\mu_{\max}^{*?}} = \frac{\partial^2}{\partial \mu_1^2} \left(I(\mu_1, \mu_{\max}^{*?})^\alpha \right) \Big|_{\mu_1=\mu_{\max}^{*?}} - c''(\mu_{\max}^{*?}) < 0,$$

which implies that $\mu_1 = \mu_{\max}^{*\?}$ is a local maximum of $U(\mu_1, \mu_{\max}^{*\?})$. In particular, when $\alpha \in (0, 1]$, (A.193) ≤ 0 because $\frac{\alpha-1}{\alpha+1} \leq 0$, implying that $\frac{\partial^2 U(\mu_1, \mu)}{\partial \mu_1^2} < 0$ for all $\mu_1, \mu > 0$, i.e., the utility function is concave and $\mu_{\max}^{*\?}$ is a global maximum, for $\alpha \in (0, 1]$. Therefore, when $\alpha \in (0, 1]$, $\mu^* > 0$ is an equilibrium if and only if it satisfies the FOC (A.190). ■

APPENDIX B

APPENDIX FOR CHAPTER 2

In this chapter, we provide proofs for the results stated in Chapter 2. The proofs of these results are in the order in which they appear in Chapter 2.

B.1 Proofs from Section 2.3

B.1.1 Preliminaries

We begin by specifying the system state in the 2-class $D/D/1+M$ queue, defined in Problem Instance 1. The system state at any time $t \geq 0$ is given by $Y(t) = (Q_1(t), Q_2(t), U(t)) \in \mathbb{Y}$, where $\mathbb{Y} = \mathbb{Z}_+^2 \times [0, 1]$. Specifically, for each $j \in \{1, 2\}$,

- $Q_j(t) \in \mathbb{Z}_+$ is the number of class j customers waiting in queue at time t .
- $U(t) \in [0, 1]$ is the remaining service time of the customer in service at time t .

If $U(t) = 0$ for all $t \in [t_1, t_2]$ for some $t_2 > t_1 \geq 0$, then it means that the server is idle over $[t_1, t_2]$. Compared to the state space for the more general model (see Section 2.2.1.2), some information is lost in this state descriptor. First, we do not include $\alpha_1(t)$ and $\alpha_2(t)$, because $\alpha_1(t) = \alpha_2(t) = t - \lfloor t \rfloor$ (since the inter-arrival times are deterministic). Second, $\nu_1(t)$ and $\nu_2(t)$ are not separately tracked (given the assumption that two classes have identical service time distributions). If $\nu(t)$ denotes the time t age-in-service measure for the customer currently in service, which disregards the specific class, then $U(t) = 1 - \nu(t)$. Finally, due to the memoryless property of the exponential distribution, there is no need to track the amount of time that has passed between each customer's arrival time up until that customer's potential abandonment time; as a result, $\eta_1(t)$ and $\eta_2(t)$ are no longer needed in the state descriptor.

For each admissible policy $\pi \in \Pi$ (see Definition 4), we use $\{Y(s; \pi, y) : s \geq 0\}$ to denote the system state process under policy π , when the initial state is $y \in \mathbb{Y}$. Fixing $\omega \in \Omega$, the realization of patience times for both classes of customers is fixed. Then, $\{Y(s; \pi, y, \omega) : s \geq 0\}$ represents a sample path. We may express or suppress the dependency on π , y and ω as appropriate given the context. When the system state process $\{Y(s; \pi) : s \geq 0\}$ admits a stationary distribution, we denote it by p_∞^π , and let $Y_\infty^\pi = (Q_{1,\infty}^\pi, Q_{2,\infty}^\pi, U_\infty^\pi)$ be a random vector distributed according to p_∞^π .

Finally, we define some special scheduling policies, which are useful for the proofs from Section 2.3. Let π_0 denote a “do-nothing” scheduling policy, wherein the server does not admit any job into service. Let $\pi_{1>2}$ denote the non-preemptive static priority scheduling policy that strictly prioritizes class 1 over 2 (and idles only if no customers are waiting), and $\pi_{2>1}$ denote the non-preemptive static priority policy that strictly prioritizes class 2 over 1 (and idles only if no customers are waiting).

B.1.2 Proof of Lemma 7

Without loss of generality, we prove the result by assuming $a_1 > a_2$ and $\theta_1 > \theta_2$. The result when $a_1 < a_2$ and $\theta_1 < \theta_2$ follows similarly. When $a_1 > a_2$ and $\theta_1 > \theta_2$, the $a\mu$ -rule is $\pi_{1>2}$ (see Definition 5). We prove the result by induction on T .

Base case: Show that $\pi_T^* = \pi_{1>2}$ for all $T \in [0, 2)$.

Note that the system is empty (with no customers waiting or being served) over $[0, 1)$, so $\mathcal{C}_T(\pi) = 0$ for any $T \in [0, 1)$ and for all $\pi \in \Pi$.

At time 1, a pair of class 1 and class 2 customers arrive and the server is idle. Since the class 1 customer is more likely to abandon ($\theta_1 > \theta_2$), and the class 1 abandonment cost is higher ($a_1 > a_2$), it is optimal to admit the class 1 customer into service at time 1. This is because any deterministic admissible policy that is different from $\pi_{1>2}$, namely, $\pi_{2>1}$ or π_0 , would incur additional cost. As a result, any randomization would also incur additional cost.

Moreover, by the non-preemption assumption (see Definition 4), the server is busy with the class 1 customer from time 1 through time 2; that is, $\pi_T^* = \pi_{1>2}$ for all $T \in [1, 2]$.

Induction step: Show that for every $S \in \{2, 3, \dots\}$, if $\pi_T^* = \pi_{1>2}$ for all $T \in [0, S]$, then $\pi_T^* = \pi_{1>2}$ for all $T \in [0, S + 1]$.

Assume the induction hypothesis that for a particular $S \in \{2, 3, \dots\}$, $\pi_T^* = \pi_{1>2}$ for all $T \in [0, S]$, meaning that the server only serves class 1 customers and never serves class 2 customers in $[0, S]$, recalling that the inter-arrival and service times are both deterministic and equal to one. Thus, at time S , the server finishes the class 1 customer at hand and becomes available. Meanwhile, a pair of class 1 and class 2 customers enter the system. Similar to the analysis in the base case, it is optimal to admit the HL class 1 customer into service at time S . By the non-preemption assumption, that customer occupies the server from time S through time $S + 1$; that is, $\pi_T^* = \pi_{1>2}$ for all $T \in [S, S + 1]$. Together with the induction hypothesis, we deduce that $\pi_T^* = \pi_{1>2}$ for all $T \in [0, S + 1]$, establishing the induction step.

Conclusion: Since both the base case and the induction step have been proved as true, we can conclude that $\pi_T^* = \pi_{1>2}$ for all $T \geq 0$. ■

B.1.3 Proof of Lemma 8

Without loss of generality, we prove the result by assuming $a_1 > a_2$ and $\theta_1 > \theta_2$. The result when $a_1 < a_2$ and $\theta_1 < \theta_2$ follows similarly. When $a_1 > a_2$ and $\theta_1 > \theta_2$, the $a\mu$ -rule is $\pi_{1>2}$, which is proved to be exactly optimal with respect to the finite-horizon expected total cost (from Lemma 7).

Suppose we can establish the following claim, whose proof is delayed to the end.

Claim 15. *Given any idling policy $\pi_I \in \Pi$, there exists a non-idling policy $\pi_N \in \Pi$ such that $\mathcal{R}_{a\mu}(T; \pi_N) \leq \mathcal{R}_{a\mu}(T; \pi_I)$, for all $T \geq 0$.*

Then, we can focus on non-idling policies in the remainder of the proof. The benefit of

non-idling policies is that the server always starts and finishes work at integer-valued time points and never idles. Fix a non-idling policy $\pi_N \in \Pi$. Note that the system is empty over $[0, 1)$, so $\mathcal{R}_{a\mu}(1; \pi_N) = 0$. For each $S \in \{2, 3, \dots, T\}$, by Definition 6,

$$\mathcal{R}_{a\mu}(S; \pi_N) - \mathcal{R}_{a\mu}(S-1; \pi_N) = [\mathcal{C}_S(\pi_N) - \mathcal{C}_{S-1}(\pi_N)] - [\mathcal{C}_S(\pi_{a\mu}) - \mathcal{C}_{S-1}(\pi_{a\mu})], \quad (\text{B.1})$$

where

$$\begin{aligned} \mathcal{C}_S(\pi_N) - \mathcal{C}_{S-1}(\pi_N) &= a_1 \cdot \mathbb{E}[R_1(S; \pi_N) - R_1(S-1; \pi_N)] \\ &\quad + a_2 \cdot \mathbb{E}[R_2(S; \pi_N) - R_2(S-1; \pi_N)], \\ \mathcal{C}_S(\pi_{a\mu}) - \mathcal{C}_{S-1}(\pi_{a\mu}) &= a_1 \cdot \mathbb{E}[R_1(S; \pi_{a\mu}) - R_1(S-1; \pi_{a\mu})] \\ &\quad + a_2 \cdot \mathbb{E}[R_2(S; \pi_{a\mu}) - R_2(S-1; \pi_{a\mu})]. \end{aligned}$$

We first compute the expected cost under $\pi_{a\mu}$. It is clear that, for each $S \in \{2, 3, \dots, T\}$,

$$\begin{aligned} \mathcal{C}_S(\pi_{a\mu}) - \mathcal{C}_{S-1}(\pi_{a\mu}) &= a_1 \cdot 0 + a_2 \cdot \left[\left(e^{-\theta_2(S-2)} - e^{-\theta_2(S-1)} \right) + \left(e^{-\theta_2(S-3)} - e^{-\theta_2(S-2)} \right) \right. \\ &\quad \left. + \dots + \left(e^{-\theta_2} - e^{-2\theta_2} \right) + \left(1 - e^{-\theta_2} \right) \right] = a_2 \left(1 - e^{-\theta_2(S-1)} \right). \end{aligned} \quad (\text{B.2})$$

Next, we compute the expected cost under π_N depending on its value. Note that non-idling $\pi_N \in \Pi$ can be $\pi_{1>2}$, $\pi_{2>1}$, or any randomization over the former two.

- Case 1: $\pi_N = \pi_{1>2}$ over $[S-1, S]$. Recall that $\pi_{1>2}$ incurs the least possible expected cost; that is, $\mathcal{C}_S(\pi_{a\mu}) - \mathcal{C}_{S-1}(\pi_{a\mu}) \leq \mathcal{C}_S(\pi_N) - \mathcal{C}_{S-1}(\pi_N)$ for any $\pi_N \in \Pi$. Hence, from (B.1),

$$\mathcal{R}_{a\mu}(S; \pi_N) - \mathcal{R}_{a\mu}(S-1; \pi_N) \geq 0. \quad (\text{B.3})$$

- Case 2: $\pi_N = \pi_{2>1}$ over $[S-1, S]$. For $S = 2$,

$$\mathcal{C}_S(\pi_N) - \mathcal{C}_{S-1}(\pi_N) = a_1 \left(1 - e^{-\theta_1}\right). \quad (\text{B.4})$$

For $S \in \{3, 4, \dots, T\}$, there is a lower bound for $\mathcal{C}_S(\pi_N) - \mathcal{C}_{S-1}(\pi_N)$ when $\pi_{1>2}$ is applied on $[0, S-1]$ (recalling that $\pi_{1>2}$ incurs the least possible expected cost over any time horizon):

$$\begin{aligned} \mathcal{C}_S(\pi_N) - \mathcal{C}_{S-1}(\pi_N) &\geq a_2 \cdot \left[\left(e^{-\theta_2(S-2)} - e^{-\theta_2(S-1)}\right) + \left(e^{-\theta_2(S-3)} - e^{-\theta_2(S-2)}\right) \right. \\ &\quad \left. + \dots + \left(e^{-\theta_2} - e^{-2\theta_2}\right) \right] + a_1 \cdot \left(1 - e^{-\theta_1}\right) = a_2 \left(e^{-\theta_2} - e^{-\theta_2(S-1)}\right) + a_1 \left(1 - e^{-\theta_1}\right). \end{aligned} \quad (\text{B.5})$$

Note that (B.4) can be equivalently written as $\mathcal{C}_S(\pi_N) - \mathcal{C}_{S-1}(\pi_N) = a_2 \left(e^{-\theta_2} - e^{-\theta_2(S-1)}\right) + a_1 \left(1 - e^{-\theta_1}\right)$ (given that $S = 2$). Hence, (B.4) and (B.5) can be combined as follows: for each $S \in \{2, 3, \dots, T\}$,

$$\mathcal{C}_S(\pi_N) - \mathcal{C}_{S-1}(\pi_N) \geq a_2 \left(e^{-\theta_2} - e^{-\theta_2(S-1)}\right) + a_1 \left(1 - e^{-\theta_1}\right). \quad (\text{B.6})$$

Substituting for $\mathcal{C}_S(\pi_N) - \mathcal{C}_{S-1}(\pi_N)$ and $\mathcal{C}_S(\pi_{a\mu}) - \mathcal{C}_{S-1}(\pi_{a\mu})$ using (B.6) and (B.2), respectively, into (B.1) implies that, for each $S \in \{2, 3, \dots, T\}$,

$$\begin{aligned} &\mathcal{R}_{a\mu}(S; \pi_N) - \mathcal{R}_{a\mu}(S-1; \pi_N) \\ &\geq a_2 \left(e^{-\theta_2} - e^{-\theta_2(S-1)}\right) + a_1 \left(1 - e^{-\theta_1}\right) - a_2 \left(1 - e^{-\theta_2(S-1)}\right) \\ &\geq a_1(1 - e^{-\theta_1}) - a_2(1 - e^{-\theta_2}) > 0. \end{aligned} \quad (\text{B.7})$$

Finally, summing over S , and using (B.3) and (B.7), we obtain that, for any real $T \geq 2$,

$$\begin{aligned}
\mathcal{R}_{a\mu}(T; \pi_N) &\geq \mathcal{R}_{a\mu}(\lfloor T \rfloor; \pi_N) \\
&= \sum_{S=2}^{\lfloor T \rfloor} (\mathcal{R}_{a\mu}(S; \pi_N) - \mathcal{R}_{a\mu}(S-1; \pi_N)) \\
&\geq 0 \cdot \mathbb{E}[T(\pi_{1>2})] + (a_1(1 - e^{-\theta_1}) - a_2(1 - e^{-\theta_2})) \cdot (\lfloor T \rfloor - 1 - \mathbb{E}[T(\pi_{1>2})]) \\
&\geq (a_1(1 - e^{-\theta_1}) - a_2(1 - e^{-\theta_2})) \cdot (T - 2 - \mathbb{E}[T(\pi_{1>2})]).
\end{aligned}$$

This, together with Claim 15, implies that, for any $\pi \in \Pi$ (may or may not be idling),

$$\mathcal{R}_{a\mu}(T; \pi) \geq (a_1(1 - e^{-\theta_1}) - a_2(1 - e^{-\theta_2})) \cdot (T - 2 - \mathbb{E}[T(\pi_{1>2})]). \quad (\text{B.8})$$

If Assumption 4 is not satisfied, then $T - \mathbb{E}[T(\pi_{1>2})] = \Theta(T)$. Then, from (B.8),

$$\mathcal{R}_{a\mu}(T; \pi) = \Omega(T).$$

■

To complete the proof, we verify Claim 15 as follows.

B.1.3.1 Proof of Claim 15

It suffices to show that $\mathcal{C}_T(\pi_N) \leq \mathcal{C}_T(\pi_I)$, where \mathcal{C}_T is given by (2.4). We prove the result by induction on T . Note that the system stays empty over $[0, 1)$, so we can ignore this period and start the induction from $T = 2$.

Base case: Show that there exists π_N over $[1, 2)$ such that $\mathcal{C}_2(\pi_N) \leq \mathcal{C}_2(\pi_I)$.

At time 1, a pair of class 1 and class 2 customers arrive into the system. Suppose the idling policy π_I starts to serve a customer (say, the class 1 customer if she is present) at time

$t \in (1, 2]$. Then, the expected cumulative abandonment cost at time 2 is given by

$$\mathcal{C}_2(\pi_I) = a_1 \left(1 - e^{-\theta_1(t-1)}\right) + a_2 \left(1 - e^{-\theta_2}\right).$$

Let π_N do exactly the same as π_I over $[1, 2)$ except that it does not idle in $[1, t)$; that is, π_N starts to serve the class 1 customer at time 1. This incurs a lower expected cost:

$$\mathcal{C}_2(\pi_N) = a_2 \left(1 - e^{-\theta_2}\right) < \mathcal{C}_2(\pi_I).$$

Induction step: Show that for every $S \in \{2, 3, \dots\}$, if there exists π_N over $[1, S)$ such that $\mathcal{C}_S(\pi_N) \leq \mathcal{C}_S(\pi_I)$, then there exists π_N on $[S, S+1)$ such that $\mathcal{C}_{S+1}(\pi_N) \leq \mathcal{C}_{S+1}(\pi_I)$.

Suppose the server under π_I becomes free at time $t \in [S, S+1)$, stays idle in $[t, t^+)$ for some $t^+ \in (t, S+1]$, and starts to work on a customer (say, the HL class 1 customer without loss of generality) at time t^+ . Then, the expected abandonment cost over $[S, S+1)$ is given by

$$\mathcal{C}_{S+1}(\pi_I) - \mathcal{C}_S(\pi_I) = C_b + a_1 \left(1 - e^{-\theta_1 t^+}\right) + a_2 \left(1 - e^{-\theta_2}\right),$$

where C_b represents the expected abandonment cost associated with the non-HL customers. Let π_N over $[S+1, S)$ do exactly the same as π_I except that it does not idle in $[t, t^+)$; that is, π_N starts to work on the class 1 customer at time t . This incurs a lower expected cost:

$$\mathcal{C}_{S+1}(\pi_N) - \mathcal{C}_S(\pi_N) = C_b + a_1 \left(1 - e^{-\theta_1 t}\right) + a_2 \left(1 - e^{-\theta_2}\right) < \mathcal{C}_{S+1}(\pi_I) - \mathcal{C}_S(\pi_I),$$

noting that $t < t^+$. Together with the induction hypothesis, we deduce that

$$\mathcal{C}_{S+1}(\pi_N) = (\mathcal{C}_{S+1}(\pi_N) - \mathcal{C}_S(\pi_N)) + \mathcal{C}_S(\pi_N) < (\mathcal{C}_{S+1}(\pi_I) - \mathcal{C}_S(\pi_I)) + \mathcal{C}_S(\pi_I) = \mathcal{C}_{S+1}(\pi_I),$$

establishing the induction step.

Conclusion: Since both the base case and the induction step have been proved as true, we can conclude the proof. ■

B.1.4 Proof of Theorem 6

The proof outline is below:

Step 1: We first construct two problem instances within Problem Instance 1 such that the optimal scheduling policy with respect to the finite-horizon expected total cost (which is the $a\mu$ -rule by Lemma 7) is $\pi_{1>2}$ in one instance and $\pi_{2>1}$ in the other instance.

Step 2: We then use divergence decomposition and the Bretagnolle-Huber inequality to establish a lower bound on the cumulative time of not applying the optimal scheduling policy.

Step 3: We finally associate the cumulative time of not applying the optimal policy to regret.

In what follows, we proceed with each step.

Step 1: Constructing two instances.

We consider the following two problem instances within Problem Instance 1:

(a) Instance η : $a_1^{(\eta)} = 2$, $a_2^{(\eta)} = 1$, $\theta_1^{(\eta)} = 3.5$, $\theta_2^{(\eta)} = 3$.

(b) Instance γ : $a_1^{(\gamma)} = 1$, $a_2^{(\gamma)} = 2$, $\theta_1^{(\gamma)} = 2.5$, $\theta_2^{(\gamma)} = 3$.

For the remainder of this proof, we superscript all model inputs (see Section 2.2.1.1) that are associated with these two instances by (η) or (γ) , e.g., $(g_1^a)^{(\eta)}$, $(g_1^s)^{(\eta)}$ and $(g_1^r)^{(\eta)}$.

The instances η and γ form a valid competing pair since the realizations (including inter-arrival, service, and patience times) on each sample path can potentially occur in both instances. This is because the inter-arrival and service times are deterministic and equal to one in both instances, and the patience times in both instances follow an exponential

distribution with the same support $[0, \infty)$. From Lemma 7, the optimal scheduling policy is $\pi_{a\mu}$ (see Definition 5), which is $\pi_{1>2}$ for instance η and $\pi_{2>1}$ for instance γ .

Fix an admissible scheduling policy $\pi \in \Pi$ and fix a time horizon T . We restrict the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\Omega^T, \mathcal{F}^T, \mathbb{P}^T)$, where Ω^T represents the possible realizations of sequence of inter-arrival, service, and patience times over $[0, T]$, and \mathcal{F}^T and \mathbb{P}^T are the restrictions of \mathcal{F} and \mathbb{P} to Ω^T . In the remainder of the proof, we suppress the dependency on T for ease of notation. Moreover, let \mathbb{P}_η (which induces \mathbb{E}_η) denote the probability measures induced by the interconnection of policy π and instance η over $[0, T]$ (respectively, π and γ).

Step 2: Lower bounding the cumulative time of not applying $\pi_{1>2}$ in instance η .

Let $T(\pi_{1>2})$ and $T(\pi_{2>1})$ be the cumulative times that $\pi_{1>2}$ and $\pi_{2>1}$ are applied over $[0, T]$, respectively, under policy π . Then, by Assumption 4, there exist some finite positive constants T_0 and Ψ such that the following hold for all $T > T_0$:

$$T - \mathbb{E}_\eta [T(\pi_{1>2})] \leq \Psi T^\alpha, \quad (\text{B.9})$$

$$T - \mathbb{E}_\gamma [T(\pi_{2>1})] \leq \Psi T^\alpha. \quad (\text{B.10})$$

Define event

$$\mathcal{A}_T := \left\{ T(\pi_{1>2}) < \frac{T}{2} \right\}. \quad (\text{B.11})$$

We evaluate the probability of this event for each instance separately. By Markov's inequality, together with (B.9) and (B.10), it follows that, for all $T > T_0$,

$$\mathbb{P}_\eta(\mathcal{A}_T) = \mathbb{P}_\eta \left(T(\pi_{1>2}) < \frac{T}{2} \right) = \mathbb{P}_\eta \left(T - T(\pi_{1>2}) \geq \frac{T}{2} \right) \leq \frac{\mathbb{E}_\eta [T - T(\pi_{1>2})]}{\frac{T}{2}} \leq \frac{2\Psi}{T^{1-\alpha}}, \quad (\text{B.12})$$

$$\mathbb{P}_\gamma(\mathcal{A}_T^c) = \mathbb{P}_\gamma \left(T(\pi_{1>2}) \geq \frac{T}{2} \right) \leq \frac{\mathbb{E}_\gamma [T(\pi_{1>2})]}{\frac{T}{2}} \leq \frac{\mathbb{E}_\gamma [T - T(\pi_{2>1})]}{\frac{T}{2}} \leq \frac{2\Psi}{T^{1-\alpha}}. \quad (\text{B.13})$$

Then, using the Bretagnolle-Huber inequality,

$$\mathbb{P}_\eta(\mathcal{A}_T) + \mathbb{P}_\gamma(\mathcal{A}_T^c) \geq \frac{1}{2} \exp \left(-\mathbb{E}_\eta \left[\log \left(\frac{\mathcal{L}(\eta | \{Y(t) : t \in [0, T]\})}{\mathcal{L}(\gamma | \{Y(t) : t \in [0, T]\})} \right) \right] \right),$$

where $\mathcal{L}(\eta | \{Y(t) : t \in [0, T]\})$ represents the likelihood function in instance η under $\{Y(t) : t \in [0, T]\}$. The above display, together with (B.12) and (B.13), implies that, for all $T > T_0$,

$$\begin{aligned} \mathbb{E}_\eta \left[\log \left(\frac{\mathcal{L}(\eta | \{Y(t) : t \in [0, T]\})}{\mathcal{L}(\gamma | \{Y(t) : t \in [0, T]\})} \right) \right] &\geq \log \left(\frac{1}{2(\mathbb{P}_\eta(\mathcal{A}_T) + \mathbb{P}_\gamma(\mathcal{A}_T^c))} \right) \\ &\geq \log \left(\frac{1}{2 \left(\frac{2\Psi}{T^{1-\alpha}} + \frac{2\Psi}{T^{1-\alpha}} \right)} \right) \\ &= (1 - \alpha) \log T - \log 8\Psi. \end{aligned} \tag{B.14}$$

The likelihood function can be expressed using the realized inter-arrival, service, and patience times. Recall that the inter-arrival and service times are deterministic and equal to one, and the patience times for class 2 are exponential distributed with identical means, in both η and γ . The likelihoods in η and γ only differ on the patience time realizations for class 1. Let $w_{1,i}$ represent the time spent in queue for the i -th class 1 customer, for $i \in \{1, 2, \dots, E_1(T)\}$. Let $\mathcal{M}_{T,1}$ and $\mathcal{M}_{T,2}$ respectively represent the set of customers who have abandoned the system by time T and who have not (i.e., either completed the service and departed, or still been in the system). Note that the patience times are right-censored data, it follows that

$$\begin{aligned} &\mathbb{E}_\eta \left[\log \left(\frac{\mathcal{L}(\eta | \{Y(t) : t \in [0, T]\})}{\mathcal{L}(\gamma | \{Y(t) : t \in [0, T]\})} \right) \right] \\ &= \mathbb{E}_\eta \left[\log \left(\prod_{i \in \mathcal{M}_{T,1}} \frac{(g_1^r)^{(\eta)}(w_{1,i})}{(g_1^r)^{(\gamma)}(w_{1,i})} \right) \right] + \mathbb{E}_\eta \left[\log \left(\prod_{i \in \mathcal{M}_{T,2}} \frac{\int_{w_{1,i}}^\infty (g_1^r)^{(\eta)}(dw)}{\int_{w_{1,i}}^\infty (g_1^r)^{(\gamma)}(dw)} \right) \right] \end{aligned}$$

$$\begin{aligned}
&= \mathbb{E}_\eta \left[\sum_{i \in \mathcal{M}_{T,1}} \log \left(\frac{(g_1^r)^{(\eta)}(w_{1,i})}{(g_1^r)^{(\gamma)}(w_{1,i})} \right) \right] + \mathbb{E}_\eta \left[\log \left(\prod_{i \in \mathcal{M}_{T,2}} e^{-\left(\theta_1^{(\eta)} - \theta_1^{(\gamma)}\right) w_{1,i}} \right) \right] \\
&= \mathbb{E}_\eta[R_1(T; \pi)] \cdot D_{KL} \left((g_1^r)^{(\eta)} \parallel (g_1^r)^{(\gamma)} \right) + \mathbb{E}_\eta \left[\log \left(e^{-\left(\theta_1^{(\eta)} - \theta_1^{(\gamma)}\right) \sum_{i \in \mathcal{M}_{T,2}} w_{1,i}} \right) \right] \\
&= \mathbb{E}_\eta[R_1(T; \pi)] \cdot D_{KL} \left((g_1^r)^{(\eta)} \parallel (g_1^r)^{(\gamma)} \right) + \mathbb{E}_\eta \left[-\left(\theta_1^{(\eta)} - \theta_1^{(\gamma)}\right) \sum_{i \in \mathcal{M}_{T,2}} w_{1,i} \right] \\
&\leq \mathbb{E}_\eta[R_1(T; \pi)] \cdot D_{KL} \left((g_1^r)^{(\eta)} \parallel (g_1^r)^{(\gamma)} \right) \\
&= \mathbb{E}_\eta[R_1(T; \pi)] \cdot \left(\log \theta_1^{(\eta)} - \log \theta_1^{(\gamma)} + \frac{\theta_1^{(\gamma)}}{\theta_1^{(\eta)}} - 1 \right), \tag{B.15}
\end{aligned}$$

where $D_{KL} \left((g_1^r)^{(\eta)} \parallel (g_1^r)^{(\gamma)} \right)$ represents the KL-divergence between $(g_1^r)^{(\eta)}$ and $(g_1^r)^{(\gamma)}$, the inequality follows from the fact that $\theta_1^{(\eta)} > \theta_1^{(\gamma)}$, and the last equality follows from the KL-divergence between two exponential distributions.

Combining (B.14) and (B.15) implies that, for all $T > T_0$,

$$\mathbb{E}_\eta[R_1(T; \pi)] \cdot \left(\log \theta_1^{(\eta)} - \log \theta_1^{(\gamma)} + \frac{\theta_1^{(\gamma)}}{\theta_1^{(\eta)}} - 1 \right) \geq (1 - \alpha) \log T - \log 8\Psi.$$

which can be equivalently written as, for all $T > T_0$,

$$\mathbb{E}_\eta[R_1(T; \pi)] \geq \left(\log \theta_1^{(\eta)} - \log \theta_1^{(\gamma)} + \frac{\theta_1^{(\gamma)}}{\theta_1^{(\eta)}} - 1 \right)^{-1} \left((1 - \alpha) \log T - \log 8\Psi \right). \tag{B.16}$$

The next step is to use (B.16) to establish a lower bound on the expected cumulative time of not applying the optimal policy $\pi_{1>2}$ in instance η , namely, $T - \mathbb{E}_\eta[T(\pi_{1>2})]$. For this, it suffices to use $T - \mathbb{E}_\eta[T(\pi_{1>2})]$ to bound $\mathbb{E}_\eta[R_1(T; \pi)]$ from above in some way. Since an upper bound on the largest possible $\mathbb{E}_\eta[R_1(T; \pi)]$ naturally serves as an upper bound for $\mathbb{E}_\eta[R_1(T; \pi)]$ under any $\pi \in \Pi$, we, henceforth, aim to bound the largest possible $\mathbb{E}_\eta[R_1(T; \pi)]$, namely, in the worst-case scenario. The worst-case scenario is that we always

apply the “do-nothing” policy π_0 whenever $\pi_{1>2}$ is not used. In this case, class 1 customers enter service if and only if $\pi_{1>2}$ is applied. We leverage this equivalence structure to establish a connection between numbers and time.

By conservation of mass (from Equation (17) in Puha and Ward (2022), recalling from Section 2.2.1.1 that the full system dynamics were not provided in the main body of the paper but can be found in Puha and Ward (2022)):

$$R_1(t) = X_1(0) + E_1(t) - X_1(t) - D_1(t), \quad \forall t \geq 0.$$

Taking expectation on both sides, and recalling that $X_1(0) = 0$ and $E_1(t) = \lfloor t \rfloor$ (by definition of Problem Instance 1), the above implies

$$\mathbb{E}_\eta[R_1(t)] = \lfloor t \rfloor - \mathbb{E}_\eta[X_1(t)] - \mathbb{E}_\eta[D_1(t)], \quad \forall t \geq 0.$$

Using this, we have

$$\begin{aligned} T - \mathbb{E}_\eta[R_1(T; \pi)] &\geq \lfloor T \rfloor - \mathbb{E}_\eta[R_1(T; \pi)] = \mathbb{E}_\eta[X_1(T; \pi)] + \mathbb{E}_\eta[D_1(T; \pi)] \\ &\stackrel{(1)}{\geq} \mathbb{E}_\eta[B_1(T; \pi)] + \mathbb{E}_\eta[D_1(T; \pi)] \\ &\stackrel{(2)}{\geq} \mathbb{E}_\eta[T(\pi_{1>2})], \end{aligned}$$

where (1) follows from (2.3), and (2) follows by recalling that customers enter service (and become either departures if service is completed, or the last one with server if service is not completed) if and only if $\pi_{1>2}$ is applied; moreover, each service departure consumes one unit of time applying $\pi_{1>2}$, and the customer in service, if any, consumes a fractional unit of time (i.e., $(0, 1]$) applying $\pi_{1>2}$. Hence, the above display implies that

$$\mathbb{E}_\eta[R_1(T; \pi)] \leq T - \mathbb{E}_\eta[T(\pi_{1>2})]. \tag{B.17}$$

Combining (B.16) and (B.17) implies that, for all $T > T_0$,

$$T - \mathbb{E}_\eta[T(\pi_{1>2})] \geq \left(\log \theta_1^{(\eta)} - \log \theta_1^{(\gamma)} + \frac{\theta_1^{(\gamma)}}{\theta_1^{(\eta)}} - 1 \right)^{-1} \left((1 - \alpha) \log T - \log 8\Psi \right). \quad (\text{B.18})$$

Step 3: Translating the expected cumulative time of not applying the optimal policy to regret.

The objective in this final step is to translate the lower bound for the expected cumulative time of not applying the optimal policy $\pi_{1>2}$ (given in (B.18)) into a lower bound on regret in instance η . We do this by leveraging Equation (B.8) in the proof of Lemma 8, which states that

$$\mathcal{R}_{a\mu}(T; \pi) \geq \left(a_1^{(\eta)} \left(1 - e^{-\theta_1^{(\eta)}} \right) - a_2^{(\eta)} \left(1 - e^{-\theta_2^{(\eta)}} \right) \right) \cdot \left(T - 2 - \mathbb{E}_\eta[T(\pi_{1>2})] \right).$$

Substituting the lower bound for $T - \mathbb{E}_\eta[T(\pi_{a\mu})]$ given in (B.18) into the above display implies that, for all $T > T_0$,

$$\begin{aligned} \mathcal{R}_{a\mu}(T; \pi) &\geq \left(a_1^{(\eta)} \left(1 - e^{-\theta_1^{(\eta)}} \right) - a_2^{(\eta)} \left(1 - e^{-\theta_2^{(\eta)}} \right) \right) \left(\log \theta_1^{(\eta)} - \log \theta_1^{(\gamma)} + \frac{\theta_1^{(\gamma)}}{\theta_1^{(\eta)}} - 1 \right)^{-1} \\ &\quad \cdot \left((1 - \alpha) \log T - \log 8\Psi - 2 \right) \\ &= \left(1 + e^{-3} - 2e^{-3.5} \right) \left(\log \frac{7}{5} - \frac{2}{7} \right)^{-1} \left((1 - \alpha) \log T - \log 8\Psi - 2 \right), \end{aligned}$$

where the constant in front of $\log T$ is strictly positive when $\alpha \in (0, 1)$.

Therefore, under any admissible policy $\pi \in \Pi$, if Assumption 4 is satisfied, then

$$\mathcal{R}_{a\mu}(T; \pi) = \Omega(\log T).$$

■

B.1.5 Proof of Proposition 10 (i)

Let $Y(n; \pi) = (Q_1(n; \pi), Q_2(n; \pi), U(n; \pi))$, $n \in \mathbb{Z}_+$, be the embedded discrete-time Markov chain (DTMC) defined by observing the continuous-time state process at each integer-valued time n , under policy $\pi \in \Pi$. Then, it is sufficient to show that the DTMC $Y(n; \pi)$ is positive recurrent under any $\pi \in \Pi$ (Meyn and Tweedie (1993b,c)). We prove this by the Foster's criterion.

Define the Lyapunov function as $V(y) = q_1^2 + q_2^2 + u^2$, for any state variable $y = (q_1, q_2, u) \in \mathbb{Y}$. Define $M := 2 + 28 \cdot \max\left\{e^{\theta_1-2}/\theta_1^2, e^{\theta_2-2}/\theta_2^2\right\}$, and a compact set $\mathcal{K} := \{(q_1, q_2, u) \in \mathbb{Y} : q_1 + q_2 \leq M\}$. Let $\varepsilon = 1$. We want to show that the following two conditions hold under any $\pi \in \Pi$:

- (i) $\mathbb{E}[V(Y(n+1; \pi)) \mid Y(n; \pi) = (q_1, q_2, u)] < \infty$ for all $(q_1, q_2, u) \in \mathcal{K}$;
- (ii) $\mathbb{E}[V(Y(n+1; \pi)) - V(Y(n; \pi)) \mid Y(n; \pi) = (q_1, q_2, u)] \leq -\varepsilon$ for all $(q_1, q_2, u) \in \mathbb{Y} \setminus \mathcal{K}$.

We first verify condition (i). Since there is a pair of class 1 and class 2 customers entering the system at time $n+1$, the queue length of each class can increase by at most one from time n to time $n+1$, which happens if no customer enters service or abandons the queue. Thus, for all $(q_1, q_2, u) \in \mathcal{K}$,

$$\begin{aligned} & \mathbb{E}[V(Y(n+1; \pi)) \mid Y(n; \pi) = (q_1, q_2, u)] \\ &= \mathbb{E}[Q_1(n+1; \pi)^2 + Q_2(n+1; \pi)^2 + U(n+1; \pi)^2 \mid (Q_1(n; \pi), Q_2(n; \pi), U(n; \pi)) = (q_1, q_2, u)] \\ &\leq (q_1 + 1)^2 + (q_2 + 1)^2 + 1 < 2(M+1)^2 + 1 < \infty. \end{aligned}$$

We next verify condition (ii). Note that

$$\begin{aligned} & \mathbb{E}[V(Y(n+1; \pi)) - V(Y(n; \pi)) \mid Y(n; \pi) = (q_1, q_2, u)] \\ &= \mathbb{E}[Q_1(n+1; \pi)^2 + Q_2(n+1; \pi)^2 + U(n+1; \pi)^2 - Q_1(n; \pi)^2 - Q_2(n; \pi)^2 - U(n; \pi)^2] \end{aligned}$$

$$\begin{aligned}
& | (Q_1(n; \pi), Q_2(n; \pi), U(n; \pi)) = (q_1, q_2, u)] \\
& \leq \mathbb{E}[Q_1(n+1; \pi)^2 + Q_2(n+1; \pi)^2 | (Q_1(n; \pi), Q_2(n; \pi), U(n; \pi)) = (q_1, q_2, u)] - (q_1^2 + q_2^2) + 1.
\end{aligned} \tag{B.19}$$

It is not difficult to see that the upper bound in the above display is maximized when the policy employed during the interval $[n, n+1]$ is the “do-nothing” policy π_0 (under which the server does not take any customers from the queue into service, although she continues working on the current customer, if any, due to the non-preemption assumption). In what follows, we study the state transition from time n to time $n+1$ when π_0 is applied during $[n, n+1]$. Without loss of generality, we focus on the state transition of Q_1 , and the state transition of Q_2 follows similarly.

- Case 1: The event $Q_1(n+1) = q_1 + 1$ happens (given that $Q_1(n) = q_1$) if and only if the patience times of all the q_1 customers in the queue are at least 1, meaning that no class 1 customer in the queue abandons the system during this interval. This happens with probability $(e^{-\theta_1})^{q_1}$.
- Case 2: The event $Q_1(n+1) = q_1$ happens (given that $Q_1(n) = q_1$) if and only if one customer in the queue has a patience time less than 1, and the rest of the $q_1 - 1$ customers in the queue have patience times that are at least 1, meaning that there is exactly one class 1 customer in the queue abandoning the queue during this interval. This happens with probability $q_1(e^{-\theta_1})^{q_1-1}(1 - e^{-\theta_1})$.
- Case 3: The event $Q_1(n+1) \leq q_1 - 1$ happens (given that $Q_1(n) = q_1$) with probability $1 - (e^{-\theta_1})^{q_1} - q_1(e^{-\theta_1})^{q_1-1}(1 - e^{-\theta_1})$.

Thus, if π_0 is applied during $[n, n+1]$, then, for $q_1 \geq 1$,

$$\mathbb{E}[Q_1(n+1)^2 | Q_1(n; \pi) = q_1] \leq (q_1 + 1)^2 \cdot (e^{-\theta_1})^{q_1} + q_1^2 \cdot q_1(e^{-\theta_1})^{q_1-1}(1 - e^{-\theta_1})$$

$$\begin{aligned}
& + (q_1 - 1)^2 \cdot \left(1 - \left(e^{-\theta_1} \right)^{q_1} - q_1 (e^{-\theta_1})^{q_1-1} (1 - e^{-\theta_1}) \right) \\
& = (q_1^2 - 2q_1 + 1) + (5q_1 - 2q_1^2) \cdot e^{-\theta_1 q_1} + (2q_1^2 - q_1) e^{-\theta_1(q_1-1)} \\
& \leq (q_1^2 - 2q_1 + 1) + 7q_1^2 \cdot e^{-\theta_1(q_1-1)}. \tag{B.20}
\end{aligned}$$

Similarly, if π_0 is applied during $[n, n+1]$, then, for $q_2 \geq 1$,

$$\mathbb{E}[Q_2(n+1)^2 \mid Q_2(n; \pi) = q_2] \leq (q_2^2 - 2q_2 + 1) + 7q_2^2 \cdot e^{-\theta_2(q_2-1)}. \tag{B.21}$$

Combining (B.19), (B.20) and (B.21) yields

$$\begin{aligned}
& \mathbb{E}[V(Y(n+1; \pi)) - V(Y(n; \pi)) \mid Y(n; \pi) = (q_1, q_2, u)] \\
& \leq (q_1^2 - 2q_1 + 1) + 7q_1^2 \cdot e^{-\theta_1(q_1-1)} + (q_2^2 - 2q_2 + 1) + 7q_2^2 \cdot e^{-\theta_2(q_2-1)} - (q_1^2 + q_2^2) + 1 \\
& = -2(q_1 + q_2) + 3 + 7q_1^2 \cdot e^{-\theta_1(q_1-1)} + 7q_2^2 \cdot e^{-\theta_2(q_2-1)}. \tag{B.22}
\end{aligned}$$

It is straightforward that (B.22) $\leq -\varepsilon$ if

$$q_1 + q_2 > \frac{1}{2} \left(4 + 7q_1^2 \cdot e^{-\theta_1(q_1-1)} + 7q_2^2 \cdot e^{-\theta_2(q_2-1)} \right),$$

which is true for all $(q_1, q_2, u) \in \mathbb{Y} \setminus \mathcal{K}$. To see this, one can show that $7q_1^2 \cdot e^{-\theta_1(q_1-1)} \leq 28e^{\theta_1-2}/\theta_1^2$ for $q_1 \in [0, \infty)$, and $7q_2^2 \cdot e^{-\theta_2(q_2-1)} \leq 28e^{\theta_2-2}/\theta_2^2$ for $q_2 \in [0, \infty)$. Hence, for all $(q_1, q_2, u) \in \mathbb{Y} \setminus \mathcal{K}$,

$$\mathbb{E}[V(Y(n+1; \pi)) - V(Y(n; \pi)) \mid Y(n; \pi) = (q_1, q_2, u)] \leq -\varepsilon.$$

In conclusion, conditions (i) and (ii) are both satisfied. Hence, we deduce from the Foster's theorem that the embedded DTMC is positive recurrent. This, in turn, implies that the 2-class $D/D/1+M$ queue under any $\pi \in \Pi$ is positive recurrent, and thus, its associated

state process $\{Y(s; \pi) : s \geq 0\}$ admits a stationary distribution. ■

B.1.6 Proof of Proposition 10 (ii)

The proof outline is below:

Step 1: We first show that a shift-coupling time is finite in expectation. This necessitates the use of three results, Lemmas 54–56, all of which are approved at the end of this subsection.

Step 2: We then leverage the finite shift-coupling time proved in Step 1 to bound the distance between the expected time-average cumulative number of abandonments for two systems with distinct initial states that are operated under the same scheduling policy.

In what follows, we proceed with each step.

Step 1: Establishing a finite shift-coupling time.

In a 2-class $D/D/1+M$ queue, let $T((q_1, q_2, u); \pi)$ be the first hitting time from any system state $(q_1, q_2, u) \in \mathbb{Y}$ to state $(0, 0, 0)$ under policy $\pi \in \Pi$; that is,

$$T((q_1, q_2, u); \pi) := \inf \{s \geq 0 : Y(s; \pi, (q_1, q_2, u)) = (0, 0, 0)\},$$

which is a random variable due to the randomness in the patience time realizations. Let

$$\bar{T}((q_1, q_2, u); \pi) := \mathbb{E}[T((q_1, q_2, u); \pi)],$$

be the expected first hitting time from state $(q_1, q_2, u) \in \mathbb{Y}$ to state $(0, 0, 0)$, where the expectation is taken with respect to the random patience times. Fixing $\omega \in \Omega$, the associated first hitting time to zero, namely, $T((q_1, q_2, u); \pi, \omega)$, becomes a constant.

We adopt the shifting-coupling method (Thorisson (2000)), where two queues are coupled at potentially different times. In particular, we consider operating two 2-class $D/D/1+M$ queues, one starting from fixed initial state $y \in \mathbb{Y}$, and one starting from an initial state distributed according to stationary distribution p_∞^π (which exists by Proposition 10 (i)). A

pair of shift-coupling times is the pair of first times at which both system states hit $(0, 0, 0)$. Without loss of generality, suppose that the initial state associated with the first queue is dominated by that associated with the second queue; that is, $y \leq Y_\infty^\pi = (Q_{1,\infty}^\pi, Q_{2,\infty}^\pi, U_\infty^\pi)$. Since both two systems have continuous paths, evidently the first hitting time to zero associated with the second queue, namely, $T((Q_{1,\infty}^\pi, Q_{2,\infty}^\pi, U_\infty^\pi); \pi)$, stochastically dominates that associated with the first queue, namely, $T(y; \pi)$. In what follows, our goal is to establish an upper bound for $\mathbb{E} [\bar{T}((Q_{1,\infty}^\pi, Q_{2,\infty}^\pi, U_\infty^\pi); \pi)]$, which denotes the expected time required for the system state, starting from an initial state chosen according to a stationary distribution p_∞^π , to first hit $(0, 0, 0)$, and the expectation is taken with respect to the random vector $(Q_{1,\infty}^\pi, Q_{2,\infty}^\pi, U_\infty^\pi)$. To do this, we use the hitting time to zero of two FCFS $D/M/1$ queues, having service times equal to the patience times in the 2-class $D/D/1+M$ queue, to provide an upper bound for the first hitting time to $(0, 0, 0)$ in the 2-class $D/D/1+M$ queue, as detailed in the next result. Let $T^{D/M(\theta)/1}(q)$ be the first hitting time from state $q \in \mathbb{Z}_+$ to state 0 in the $D/M/1$ queue with service rate θ , and $\bar{T}^{D/M(\theta)/1}(q)$ as its expectation with respect to the random service times.

Lemma 54. *For any $\pi \in \Pi$, and for any fixed initial state $(q_1, q_2, u) \in \mathbb{Y}$,*

$$\bar{T}((q_1, q_2, u); \pi) \leq 1 + \bar{T}^{D/M(\theta_1)/1}(q_1) + \bar{T}^{D/M(\theta_2)/1}(q_2).$$

Let $(q_{1,\infty}^\pi, q_{2,\infty}^\pi, u_\infty^\pi) \in \mathbb{Y}$ be a random draw from p_∞^π . Then, by Lemma 54,

$$\bar{T}((q_{1,\infty}^\pi, q_{2,\infty}^\pi, u_\infty^\pi); \pi) \leq 1 + \bar{T}^{D/M(\theta_1)/1}(q_{1,\infty}^\pi) + \bar{T}^{D/M(\theta_2)/1}(q_{2,\infty}^\pi).$$

Integrating both sides of the above display with respect to p_∞^π yields

$$\mathbb{E} [\bar{T}((Q_{1,\infty}^\pi, Q_{2,\infty}^\pi, U_\infty^\pi); \pi)] \leq 1 + \mathbb{E} [\bar{T}^{D/M(\theta_1)/1}(Q_{1,\infty}^\pi)] + \mathbb{E} [\bar{T}^{D/M(\theta_2)/1}(Q_{2,\infty}^\pi)]. \quad (\text{B.23})$$

Suppose we can establish the following stochastic order result. Let $p_\infty^{D/M(\theta)/1}$ be the stationary distribution of the FCFS $D/M/1$ queue with service rate θ . (The state descriptor of the FCFS $D/M/1$ queue counts the total jobs in the system.) Let $Q_\infty^{D/M(\theta)/1}$ be a random variable distributed according to $p_\infty^{D/M(\theta)/1}$.

Lemma 55. *For any $\pi \in \Pi$,*

$$Q_{j,\infty}^\pi \leq_{st} Q_\infty^{D/M(\theta_j)/1}, \quad \forall j \in \{1, 2\}.$$

Then, applying the (deterministic) increasing function $\bar{T}^{D/M(\theta_j)/1}$, $j \in \{1, 2\}$, to the stochastic order in Lemma 55 yields

$$\mathbb{E} \left[\bar{T}^{D/M(\theta_j)/1} \left(Q_{j,\infty}^\pi \right) \right] \leq \mathbb{E} \left[\bar{T}^{D/M(\theta_j)/1} \left(Q_\infty^{D/M(\theta_j)/1} \right) \right], \quad \forall j \in \{1, 2\}.$$

Using this, (B.23) implies

$$\begin{aligned} \mathbb{E} \left[\bar{T}((Q_{1,\infty}^\pi, Q_{2,\infty}^\pi, U_\infty^\pi); \pi) \right] &\leq 1 + \mathbb{E} \left[\bar{T}^{D/M(\theta_1)/1} \left(Q_\infty^{D/M(\theta_1)/1} \right) \right] \\ &\quad + \mathbb{E} \left[\bar{T}^{D/M(\theta_2)/1} \left(Q_\infty^{D/M(\theta_2)/1} \right) \right]. \end{aligned} \quad (\text{B.24})$$

Given that the FCFS $D/M/1$ queue has a closed-form expression for the stationary distribution, it is possible to explicitly bound the two expectations on the right-hand side of (B.24), as detailed in the next lemma.

Lemma 56. *Assume that $\theta > 2.2$. The expected first hitting time, starting from any fixed initial state $q \in \mathbb{Z}_+$, to zero, satisfies*

$$\bar{T}^{D/M(\theta)/1}(q) \leq 1 + \frac{(e\theta)^q}{e^\theta - e\theta} \in (1, \infty).$$

Moreover, the expected first hitting time, starting from $Q_\infty^{D/M(\theta)/1}$, to zero, satisfies

$$\mathbb{E} \left[\bar{T}^{D/M(\theta)/1} \left(Q_\infty^{D/M(\theta)/1} \right) \right] \leq 1 + \frac{(1-\delta)e\theta}{(e^\theta - e\theta)(1-e\theta\delta)} \in (1, \infty).$$

Using Lemma 56, (B.24) implies that

$$\begin{aligned} \mathbb{E} \left[\bar{T}((Q_{1,\infty}^\pi, Q_{2,\infty}^\pi, U_\infty^\pi); \pi) \right] &\leq 1 + 1 + \frac{(1-\delta_1)e\theta_1}{(e^{\theta_1} - e\theta_1)(1-e\theta_1\delta_1)} + 1 + \frac{(1-\delta_2)e\theta_2}{(e^{\theta_2} - e\theta_2)(1-e\theta_2\delta_2)} \\ &= 3 + \frac{(1-\delta_1)e\theta_1}{(e^{\theta_1} - e\theta_1)(1-e\theta_1\delta_1)} + \frac{(1-\delta_2)e\theta_2}{(e^{\theta_2} - e\theta_2)(1-e\theta_2\delta_2)} \in (3, \infty). \end{aligned} \tag{B.25}$$

Step 2: Bounding the distance between the expected time-average cumulative number of abandonments.

Now, we leverage the finite shift-coupling time proved in (B.25) in Step 1 to establish a lower bound on the distance between the expected time-average cumulative number of abandonments. Note that, for each $j \in \{1, 2\}$,

$$\begin{aligned} &\left| \mathbb{E} \left[\frac{R_j(t; \pi, y)}{t} \right] - \mathbb{E} \left[\frac{R_j(t; \pi, y')}{t} \right] \right| \\ &\stackrel{(*)}{=} \left| \mathbb{E} \left[\frac{1}{t} \int_0^t \theta_j Q_j(s; \pi, y) ds \right] - \mathbb{E} \left[\frac{1}{t} \int_0^t \theta_j Q_j(s; \pi, y') ds \right] \right| \\ &= \theta_j \cdot \left| \mathbb{E} \left[\frac{1}{t} \int_0^t Q_j(s; \pi, y) ds \right] - \mathbb{E} \left[\frac{1}{t} \int_0^t Q_j(s; \pi, y') ds \right] \right| \\ &\stackrel{(**)}{=} \theta_j \cdot \left| \mathbb{E} \left[\frac{1}{t} \int_0^t Q_j(s; \pi, y) ds \right] - \mathbb{E} \left[\frac{1}{t} \int_0^t Q_j(s; \pi, Y_\infty^\pi) ds \right] \right. \\ &\quad \left. + \mathbb{E} \left[\frac{1}{t} \int_0^t Q_j(s; \pi, Y_\infty^\pi) ds \right] - \mathbb{E} \left[\frac{1}{t} \int_0^t Q_j(s; \pi, y') ds \right] \right| \\ &\leq \theta_j \cdot \left| \mathbb{E} \left[\frac{1}{t} \int_0^t Q_j(s; \pi, y) ds \right] - \mathbb{E} \left[\frac{1}{t} \int_0^t Q_j(s; \pi, Y_\infty^\pi) ds \right] \right| \\ &\quad + \theta_j \cdot \left| \mathbb{E} \left[\frac{1}{t} \int_0^t Q_j(s; \pi, Y_\infty^\pi) ds \right] - \mathbb{E} \left[\frac{1}{t} \int_0^t Q_j(s; \pi, y') ds \right] \right|, \end{aligned} \tag{B.26}$$

where $(*)$ follows by the property of exponential distributions, and the added and subtracted expectations in $(**)$ are taken with respect to both the random patience times and the random vector Y_∞^π . In what follows, we derive an upper bound for the first absolute value in (B.26), and an upper bound for the second absolute value in (B.26) follows similarly.

Recall from Step 1 that two systems operating under π , one starting from fixed initial state y , and one starting from an initial state chosen according to p_∞^π , are coupled at a pair of first hitting time to $(0, 0, 0)$, namely, $T(y; \pi)$ and $T(Y_\infty^\pi; \pi)$. By definition of shift-coupling, for each $j \in \{1, 2\}$,

$$\mathbb{P}\left\{Q_j\left(s + T(y; \pi); \pi, y\right) = Q_j\left(s + T(Y_\infty^\pi; \pi); \pi, Y_\infty^\pi\right)\right\} = 1, \quad \forall s \geq 0. \quad (\text{B.27})$$

Then, for each $j \in \{1, 2\}$,

$$\begin{aligned} & \left| \mathbb{E} \left[\frac{1}{t} \int_0^t Q_j(s; \pi, y) ds - \frac{1}{t} \int_0^t Q_j(s; \pi, Y_\infty^\pi) ds \right] \right| \\ &= \frac{1}{t} \left| \mathbb{E} \left[\int_0^{T(y; \pi)} Q_j(s; \pi, y) ds + \int_{T(y; \pi)}^{T(y; \pi) + t - T(Y_\infty^\pi; \pi)} Q_j(s; \pi, y) ds + \int_{T(y; \pi) + t - T(Y_\infty^\pi; \pi)}^t Q_j(s; \pi, y) ds \right. \right. \\ & \quad \left. \left. - \int_0^{T(Y_\infty^\pi; \pi)} Q_j(s; \pi, Y_\infty^\pi) ds - \int_{T(Y_\infty^\pi; \pi)}^t Q_j(s; \pi, Y_\infty^\pi) ds \right] \right| \\ &\leq \frac{1}{t} \left| \mathbb{E} \left[\int_0^{T(y; \pi)} Q_j(s; \pi, y) ds + \int_{T(y; \pi) + t - T(Y_\infty^\pi; \pi)}^t Q_j(s; \pi, y) ds - \int_0^{T(Y_\infty^\pi; \pi)} Q_j(s; \pi, Y_\infty^\pi) ds \right] \right| \\ & \quad + \frac{1}{t} \left| \mathbb{E} \left[\int_{T(y; \pi)}^{T(y; \pi) + t - T(Y_\infty^\pi; \pi)} Q_j(s; \pi, y) ds - \int_{T(Y_\infty^\pi; \pi)}^t Q_j(s; \pi, Y_\infty^\pi) ds \right] \right| \\ &\stackrel{(\dagger)}{=} \frac{1}{t} \left| \mathbb{E} \left[\int_0^{T(y; \pi)} Q_j(s; \pi, y) ds + \int_{T(y; \pi) + t - T(Y_\infty^\pi; \pi)}^t Q_j(s; \pi, y) ds - \int_0^{T(Y_\infty^\pi; \pi)} Q_j(s; \pi, Y_\infty^\pi) ds \right] \right| \\ &\leq \frac{1}{t} \cdot \left\{ \mathbb{E} \left[\int_0^{T(y; \pi)} Q_j(s; \pi, y) ds \right] + \mathbb{E} \left[\int_{T(y; \pi) + t - T(Y_\infty^\pi; \pi)}^t Q_j(s; \pi, y) ds \right] \right. \\ & \quad \left. + \mathbb{E} \left[\int_0^{T(Y_\infty^\pi; \pi)} Q_j(s; \pi, Y_\infty^\pi) ds \right] \right\}, \end{aligned} \quad (\text{B.28})$$

where (\dagger) follows from (B.27), because two random variables being equal almost surely implies being equal in expectation. Furthermore, we note that

$$\begin{aligned} \mathbb{E} \left[\int_0^{T(y; \pi)} Q_j(s; \pi, y) ds \right] &= \int_{\tau=0}^{\infty} \mathbb{E} \left[\int_0^{T(y; \pi)} Q_j(s; \pi, y) ds \mid T(y; \pi) = \tau \right] \cdot \mathbb{P} \{ T(y; \pi) = \tau \} d\tau \\ &\leq \int_{\tau=0}^{\infty} \left(\sup_{s \geq 0} \mathbb{E} [Q_j(s; \pi, y)] \right) \cdot \tau \cdot \mathbb{P} \{ T(y; \pi) = \tau \} d\tau \\ &= \left(\sup_{s \geq 0} \mathbb{E} [Q_j(s; \pi, y)] \right) \cdot \bar{T}(y; \pi). \end{aligned} \quad (\text{B.29})$$

Similarly,

$$\begin{aligned} &\mathbb{E} \left[\int_{T(y; \pi) + t - T(Y_{\infty}^{\pi}; \pi)}^t Q_j(s; \pi, y) ds \right] \\ &= \int_{\tau=0}^T \mathbb{E} \left[\int_{T(y; \pi) + t - T(Y_{\infty}^{\pi}; \pi)}^t Q_j(s; \pi, y) ds \mid T(Y_{\infty}^{\pi}; \pi) - T(y; \pi) = \tau \right] \\ &\quad \cdot \mathbb{P} \{ T(Y_{\infty}^{\pi}; \pi) - T(y; \pi) = \tau \} d\tau \\ &\leq \int_{\tau=0}^{\infty} \left(\sup_{s \geq 0} \mathbb{E} [Q_j(s; \pi, y)] \right) \cdot \tau \cdot \mathbb{P} \{ T(Y_{\infty}^{\pi}; \pi) - T(y; \pi) = \tau \} d\tau \\ &= \left(\sup_{s \geq 0} \mathbb{E} [Q_j(s; \pi, y)] \right) \cdot (\mathbb{E} [\bar{T}(Y_{\infty}^{\pi}; \pi)] - \bar{T}(y; \pi)), \end{aligned} \quad (\text{B.30})$$

and

$$\begin{aligned} &\mathbb{E} \left[\int_0^{T(Y_{\infty}^{\pi}; \pi)} Q_j(s; \pi, Y_{\infty}^{\pi}) ds \right] \\ &= \int_{\tau=0}^{\infty} \mathbb{E} \left[\int_0^{T(Y_{\infty}^{\pi}; \pi)} Q_j(s; \pi, Y_{\infty}^{\pi}) ds \mid T(Y_{\infty}^{\pi}; \pi) = \tau \right] \cdot \mathbb{P} \{ T(Y_{\infty}^{\pi}; \pi) = \tau \} d\tau \\ &\leq \int_{\tau=0}^{\infty} \left(\sup_{s \geq 0} \mathbb{E} [Q_j(s; \pi, Y_{\infty}^{\pi})] \right) \cdot \tau \cdot \mathbb{P} \{ T(Y_{\infty}^{\pi}; \pi) = \tau \} d\tau \\ &= \left(\sup_{s \geq 0} \mathbb{E} [Q_j(s; \pi, Y_{\infty}^{\pi})] \right) \cdot \mathbb{E} [\bar{T}(Y_{\infty}^{\pi}; \pi)]. \end{aligned} \quad (\text{B.31})$$

Substitution into (B.28) using (B.29)-(B.31) yields, for each $j \in \{1, 2\}$,

$$\begin{aligned}
& \left| \mathbb{E} \left[\frac{1}{t} \int_0^t Q_j(s; \pi, y) ds - \frac{1}{t} \int_0^t Q_j(s; \pi, Y_\infty^\pi) ds \right] \right| \\
& \leq \frac{1}{t} \cdot \left\{ \left(\sup_{s \geq 0} \mathbb{E} [Q_j(s; \pi, y)] \right) \cdot \left(\bar{T}(y; \pi) + \mathbb{E} [\bar{T}(Y_\infty^\pi; \pi)] - \bar{T}(y; \pi) \right) \right. \\
& \quad \left. + \left(\sup_{s \geq 0} \mathbb{E} [Q_j(s; \pi, Y_\infty^\pi)] \right) \cdot \mathbb{E} [\bar{T}(Y_\infty^\pi; \pi)] \right\} \\
& \leq \frac{2}{t} \cdot \max \left\{ \sup_{s \geq 0} \mathbb{E} [Q_j(s; \pi, y)], \sup_{s \geq 0} \mathbb{E} [Q_j(s; \pi, Y_\infty^\pi)] \right\} \cdot \mathbb{E} [\bar{T}(Y_\infty^\pi; \pi)]. \tag{B.32}
\end{aligned}$$

Next, we derive closed-form expressions for $\sup_{s \geq 0} \mathbb{E} [Q_j(s; \pi, y)]$ and $\sup_{s \geq 0} \mathbb{E} [Q_j(s; \pi, Y_\infty^\pi)]$.

For any $(q_1, q_2, b) \in \mathbb{Y}$, note that, for each $j \in \{1, 2\}$,

$$\mathbb{E} [Q_j(s; \pi, (q_1, q_2, b))] \leq \mathbb{E} [Q_j(s; \pi_0, (q_1, q_2, b))] \leq q_j + \int_0^s e^{-\theta_j(s-u)} d\mathbb{E}[E_j(u)], \forall s \geq 0,$$

where the second inequality follows from Claim 17 in the proof of Lemma 9, that applies to $G/G/\infty$ queues, noting that the system operating under π_0 is equivalent to two independent $D/M/\infty$ queues, having service times equal to the patience times in the 2-class $D/D/1+M$ queue. Note that $\mathbb{E}[E_j(u)]$ is a step function with unit jumps happening at time $1, 2, 3, \dots$ (recalling from Problem Instance 1 that the inter-arrival times are deterministic and equal to one), which satisfies $\mathbb{E}[E_j(u)] \leq u$ for all $u \geq 0$. Thus, the above display implies that, for each $j \in \{1, 2\}$,

$$\mathbb{E} [Q_j(s; \pi, (q_1, q_2, b))] \leq q_j + \int_0^s e^{-\theta_j(s-u)} du = q_j + \frac{1}{\theta_j} (1 - e^{-\theta_j s}) \leq q_j + \frac{1}{\theta_j}, \forall s \geq 0.$$

This, together with the assumption that $y \leq Y_\infty^\pi$, implies that, for each $j \in \{1, 2\}$,

$$\max \left\{ \sup_{s \geq 0} \mathbb{E} [Q_j(s; \pi, y)], \sup_{s \geq 0} \mathbb{E} [Q_j(s; \pi, Y_\infty^\pi)] \right\} \leq \mathbb{E} [Q_{j,\infty}^\pi] + \frac{1}{\theta_j} \leq \mathbb{E} [Q_{j,\infty}^{\pi_0}] + \frac{1}{\theta_j} = \frac{2}{\theta_j},$$

where $\mathbb{E}[Q_{j,\infty}^{\pi_0}] = \frac{1}{\theta_j}$ is the steady-state queue length in the $D/M/\infty$ queue with service rate θ_j .

Substituting the above into (B.32) yields, for each $j \in \{1, 2\}$,

$$\left| \mathbb{E} \left[\frac{1}{t} \int_0^t Q_j(s; \pi, y) ds - \frac{1}{t} \int_0^t Q_j(s; \pi, Y_\infty^\pi) ds \right] \right| \leq \frac{4}{\theta_j} \cdot \mathbb{E} [\bar{T}(Y_\infty^\pi; \pi)] \cdot \frac{1}{t}.$$

Using the upper bound for $\mathbb{E} [\bar{T}(Y_\infty^\pi; \pi)]$ given in (B.25), the above implies that

$$\begin{aligned} & \left| \mathbb{E} \left[\frac{1}{t} \int_0^t Q_j(s; \pi, y) ds - \frac{1}{t} \int_0^t Q_j(s; \pi, Y_\infty^\pi) ds \right] \right| \\ & \leq \frac{4}{\theta_j} \left(3 + \frac{(1 - \delta_1)e\theta_1}{(e^{\theta_1} - e\theta_1)(1 - e\theta_1\delta_1)} + \frac{(1 - \delta_2)e\theta_2}{(e^{\theta_2} - e\theta_2)(1 - e\theta_2\delta_2)} \right) \frac{1}{t}. \end{aligned} \quad (\text{B.33})$$

Similarly,

$$\begin{aligned} & \left| \mathbb{E} \left[\frac{1}{t} \int_0^t Q_j(s; \pi, Y_\infty^\pi) ds \right] - \mathbb{E} \left[\frac{1}{t} \int_0^t Q_j(s; \pi, y') ds \right] \right| \\ & \leq \frac{4}{\theta_j} \left(3 + \frac{(1 - \delta_1)e\theta_1}{(e^{\theta_1} - e\theta_1)(1 - e\theta_1\delta_1)} + \frac{(1 - \delta_2)e\theta_2}{(e^{\theta_2} - e\theta_2)(1 - e\theta_2\delta_2)} \right) \frac{1}{t}. \end{aligned} \quad (\text{B.34})$$

Substituting (B.33) and (B.34) into (B.26), we conclude that, for each $j \in \{1, 2\}$,

$$\begin{aligned} & \left| \mathbb{E} \left[\frac{R_j(t; \pi, y)}{t} \right] - \mathbb{E} \left[\frac{R_j(t; \pi, y')}{t} \right] \right| \\ & \leq 8 \left(3 + \frac{(1 - \delta_1)e\theta_1}{(e^{\theta_1} - e\theta_1)(1 - e\theta_1\delta_1)} + \frac{(1 - \delta_2)e\theta_2}{(e^{\theta_2} - e\theta_2)(1 - e\theta_2\delta_2)} \right) \frac{1}{t}, \end{aligned}$$

which establishes Proposition 10 (ii). ■

To complete the proof, we provide proofs of Lemmas 54-56 as follows. In particular, Lemmas 54 and 55 are proved via sample-path stochastic comparison arguments, and Lemma 56 presents a standalone result for the FCFS $D/M/1$ queue.

B.1.6.1 Proof of Lemma 54

We prove the result for each $\omega \in \Omega$ (i.e., a realization of patience times for both classes of customers in the 2-class $D/D/1+M$ queue). Fix a $\omega \in \Omega$.

For any $\pi \in \Pi$, and for any $(q_1, q_2, u) \in \mathbb{Y}$, we first note that

$$T((q_1, q_2, u); \pi, \omega) \leq T((q_1, q_2, u); \pi_0, \omega). \quad (\text{B.35})$$

Under the “do-nothing” policy π_0 , the initial remaining job must be completed by time $u \in [0, 1]$ due to the non-preemption assumption (see Definition 4), and the class 1 and class 2 queues behave equivalently as two independent $D/M/\infty$ queues having service times equal to the patience times realized on ω , and so, having service rates θ_1 and θ_2 . Let $T^{D/M(\theta)/\infty}(q)$ be the first hitting time from state $q \in \mathbb{Z}_+$ to state 0 in the $D/M/\infty$ queue with service rate θ . (The state descriptor of the $D/M/\infty$ queue counts the total jobs in the system.) Then,

$$\begin{aligned} T((q_1, q_2, u); \pi_0, \omega) &\leq 1 + \max \left\{ T^{D/M(\theta_1)/\infty}(q_1; \omega), T^{D/M(\theta_2)/\infty}(q_2; \omega) \right\} \\ &\leq 1 + T^{D/M(\theta_1)/\infty}(q_1; \omega) + T^{D/M(\theta_2)/\infty}(q_2; \omega). \end{aligned} \quad (\text{B.36})$$

Additionally, the $D/M/\infty$ queue has smaller queue length than the FCFS $D/M/1$ queue (both having service times equal to the patience times realized on ω), and thus the first hitting time to zero in the $D/M/\infty$ queue is shorter than that in the FCFS $D/M/1$ queue; that is,

$$T^{D/M(\theta_j)/\infty}(q_j; \omega) \leq T^{D/M(\theta_j)/1}(q_j; \omega), \quad \forall j \in \{1, 2\}. \quad (\text{B.37})$$

Substituting (B.37) into (B.36) implies that

$$T((q_1, q_2, u); \pi_0, \omega) \leq 1 + T^{D/M(\theta_1)/1}(q_1; \omega) + T^{D/M(\theta_2)/1}(q_2; \omega). \quad (\text{B.38})$$

Finally, combination of (B.35) and (B.38) yields

$$T((q_1, q_2, u); \pi, \omega) \leq 1 + T^{D/M(\theta_1)/1}(q_1; \omega) + T^{D/M(\theta_2)/1}(q_2; \omega).$$

Therefore, for any $\pi \in \Pi$, and any $(q_1, q_2, u) \in \mathbb{Y}$,

$$\bar{T}((q_1, q_2, u); \pi) \leq 1 + \bar{T}^{D/M(\theta_1)/1}(q_1) + \bar{T}^{D/M(\theta_2)/1}(q_2).$$

■

B.1.6.2 Proof of Lemma 55

We prove the statement for class $j = 1$, and the result for class $j = 2$ follows similarly.

By definition of stochastic order, it is sufficient to show that, for any $\pi \in \Pi$,

$$\mathbb{P}\left\{Q_{1,\infty}^{\pi} > x\right\} \leq \mathbb{P}\left\{Q_{\infty}^{D/M(\theta_1)/1} > x\right\}, \quad \forall x \in (0, \infty). \quad (\text{B.39})$$

Denote $\{(Q_{1,\infty}(s; \pi), Q_{2,\infty}(s; \pi), U_{\infty}(s; \pi)) : s \geq 0\}$ as a stationary process associated with the 2-class $D/D/1+M$ queue, initialized at time 0 according to a (joint) stationary distribution p_{∞}^{π} , which exists by Proposition 10 (i). Then, $(Q_{1,\infty}(s; \pi), Q_{2,\infty}(s; \pi), U_{\infty}(s; \pi)) \stackrel{d}{=} (Q_{1,\infty}^{\pi}, Q_{2,\infty}^{\pi}, U_{\infty}^{\pi})$ for all $s \geq 0$. Since equal joint distributions implies equal marginal distributions, it follows that $Q_{1,\infty}(s; \pi) \stackrel{d}{=} Q_{1,\infty}^{\pi}$ for all $s \geq 0$. Therefore, for all $s \geq 0$,

$$\mathbb{P}\left\{Q_{1,\infty}(s; \pi) > x\right\} = \mathbb{P}\left\{Q_{1,\infty}^{\pi} > x\right\}, \quad \forall x \in (0, \infty). \quad (\text{B.40})$$

For any $s \geq 0$ and for any $x \in (0, \infty)$, it follows from the law of total probability that

$$\mathbb{P}\left\{Q_{1,\infty}(s; \pi) > x\right\} = \iint_{(q_1, q_2, u) \in \mathbb{Y}} \mathbb{P}\left\{Q_{1,\infty}(s; \pi, (q_1, q_2, u)) > x\right\} \cdot p_{\infty}^{\pi}((dq_1, dq_2, du)).$$

Given that the initial state is fixed at $(Q_{1,\infty}(0; \pi), Q_{2,\infty}(0; \pi), U_\infty(0; \pi)) = (q_1, q_2, u)$, we can rewrite $Q_{1,\infty}(s; \pi, (q_1, q_2, u))$ as $Q_1(s; \pi, (q_1, q_2, u))$ in the above display and obtain

$$\mathbb{P}\{Q_{1,\infty}(s; \pi) > x\} = \iiint_{(q_1, q_2, u) \in \mathbb{Y}} \mathbb{P}\{Q_1(s; \pi, (q_1, q_2, u)) > x\} \cdot p_\infty^\pi((dq_1, dq_2, du)). \quad (\text{B.41})$$

Similar to (B.35)-(B.37) and the surrounding texts in the proof of Lemma 54, we can use the FCFS $D/M/1$ queue to provide an upper bound on the right-hand side of (B.41). Fixing a sample path $\omega \in \Omega$, for $j \in \{1, 2\}$, let $\{Q^{D/M(\theta_j)/\infty}(s; q, \omega) : s \geq 0\}$ and $\{Q^{D/M(\theta_j)/1}(s; q, \omega) : s \geq 0\}$ denote the state processes of the $D/M/\infty$ queue and the FCFS $D/M/1$ queue, respectively, initialized from $q \in \mathbb{Z}_+$ and having service times equal to the patience times of class j customers realized on ω , and so, having service rates θ_j . Then, for any $(q_1, q_2, u) \in \mathbb{Y}$, any $s \geq 0$, and for each $\omega \in \Omega$,

$$Q_1(s; \pi, (q_1, q_2, u), \omega) \leq Q_1(s; \pi_0, (q_1, q_2, u), \omega) = Q^{D/M(\theta_1)/\infty}(s; q_1, \omega) \leq Q^{D/M(\theta_1)/1}(s; q_1, \omega).$$

Therefore, for any $(q_1, q_2, u) \in \mathbb{Y}$ and any $s \geq 0$,

$$Q_1(s; \pi, (q_1, q_2, u)) \leq_{st} Q^{D/M(\theta_1)/1}(s; q_1),$$

which, by definition of stochastic order, means that, for any $(q_1, q_2, u) \in \mathbb{Y}$,

$$\mathbb{P}\{Q_1(s; \pi, (q_1, q_2, u)) > x\} \leq \mathbb{P}\{Q^{D/M(\theta_1)/1}(s; q_1) > x\}, \quad \forall x \in (0, \infty). \quad (\text{B.42})$$

Combining (B.41) and (B.42) implies that, for any $s \geq 0$ and any $x \in (0, \infty)$,

$$\mathbb{P}\{Q_{1,\infty}(s; \pi) > x\} \leq \iiint_{(q_1, q_2, u) \in \mathbb{Y}} \mathbb{P}\{Q^{D/M(\theta_1)/1}(s; q_1) > x\} \cdot p_\infty^\pi((dq_1, dq_2, du)),$$

which, by replacing its left-hand side with $\mathbb{P}\left\{Q_{1,\infty}^\pi > x\right\}$ (recalling (B.40)), can be equivalently written as

$$\mathbb{P}\left\{Q_{1,\infty}^\pi > x\right\} \leq \iiint_{(q_1,q_2,u) \in \mathbb{Y}} \mathbb{P}\left\{Q^{D/M(\theta_1)/1}(s; q_1) > x\right\} \cdot p_\infty^\pi((dq_1, dq_2, du)). \quad (\text{B.43})$$

Letting $s \rightarrow \infty$ on both sides of (B.43) yields

$$\begin{aligned} \mathbb{P}\left\{Q_{1,\infty}^\pi > x\right\} &\leq \lim_{s \rightarrow \infty} \iiint_{(q_1,q_2,u) \in \mathbb{Y}} \mathbb{P}\left\{Q^{D/M(\theta_1)/1}(s; q_1) > x\right\} \cdot p_\infty^\pi((dq_1, dq_2, du)) \\ &= \iiint_{(q_1,q_2,u) \in \mathbb{Y}} \lim_{s \rightarrow \infty} \mathbb{P}\left\{Q^{D/M(\theta_1)/1}(s; q_1) > x\right\} \cdot p_\infty^\pi((dq_1, dq_2, du)), \end{aligned} \quad (\text{B.44})$$

where the equality follows by the dominated convergence theorem, given that $\mathbb{P}\{Q^{D/M(\theta_1)/1}(s; q_1) > x\} \leq 1$ and $p_\infty^\pi(A) \leq 1$ for any Borel measurable $A \subseteq \mathbb{Y}$.

Finally, we utilize the following weak convergence result, as established in equation (17) in Jansson (1966). For any fixed initial state $q \in \mathbb{Z}_+$,

$$Q^{D/M(\theta_1)/1}(s; q) \Rightarrow Q_\infty^{D/M(\theta_1)/1}, \quad \text{as } s \rightarrow \infty.$$

recalling that $Q_\infty^{D/M(\theta_1)/1}$ represents a random variable distributed according to the stationary distribution $p_\infty^{D/M(\theta_1)/1}$. Then, it follows that

$$\lim_{s \rightarrow \infty} \mathbb{P}\left\{Q^{D/M(\theta_1)/1}(s; q_1) > x\right\} = \mathbb{P}\left\{Q_\infty^{D/M(\theta_1)/1} > x\right\}, \quad \forall x \in (0, \infty).$$

Substitution into (B.44) implies that, for any $x \in (0, \infty)$,

$$\mathbb{P}\left\{Q_{1,\infty}^\pi > x\right\} \leq \iiint_{(q_1,q_2,u) \in \mathbb{Y}} \lim_{s \rightarrow \infty} \mathbb{P}\left\{Q^{D/M(\theta_1)/1}(s; q_1) > x\right\} \cdot p_\infty^\pi((dq_1, dq_2, du))$$

$$\begin{aligned}
&= \iiint_{(q_1, q_2, u) \in \mathbb{Y}} \mathbb{P}\left\{Q_{\infty}^{D/M(\theta_1)/1} > x\right\} \cdot p_{\infty}^{\pi}((dq_1, dq_2, du)) \\
&= \mathbb{P}\left\{Q_{\infty}^{D/M(\theta_1)/1} > x\right\} \cdot \left(\iiint_{(q_1, q_2, u) \in \mathbb{Y}} p_{\infty}^{\pi}((dq_1, dq_2, du)) \right) \\
&= \mathbb{P}\left\{Q_{\infty}^{D/M(\theta_1)/1} > x\right\},
\end{aligned}$$

which establishes (B.39). \blacksquare

B.1.6.3 Proof of Lemma 56

Suppose we can establish the following claim whose proof is postponed to the end of this subsection.

Claim 16. *If $\theta > 2.2$, then $\theta\delta < 1/e$, where δ is the smallest solution to $\delta = e^{-\theta(1-\delta)}$.*

Then, by Claim 16, it suffices to prove the result for a generic FCFS $D/M/1$ queue with service rate θ that satisfies $\theta\delta < 1/e$, where δ is the smallest solution to $\delta = e^{-\theta(1-\delta)}$.

We first derive an upper bound on $\bar{T}^{D/M(\theta)/1}(q)$ for any fixed initial state $q \in \mathbb{Z}_+$. Then, we extend the bound to allow for random initial state $X_{\infty}^{D/M(\theta)/1}$ that is distributed according to the stationary distribution $p_{\infty}^{D/M(\theta)/1}$.

Step 1: An upper bound for any fixed initial state.

When $q = 0$, $\bar{T}^{D/M(\theta)/1}(q) = 0$ trivially. We assume $q \geq 1$, henceforth, without loss of generality. Note that

$$\begin{aligned}
\bar{T}^{D/M(\theta)/1}(q) &= \int_{s=0}^{\infty} \mathbb{P}\left\{T^{D/M(\theta)/1}(q) > s\right\} ds \\
&= \int_{s=0}^1 \mathbb{P}\left\{T^{D/M(\theta)/1}(q) > s\right\} ds + \int_{s=1}^2 \mathbb{P}\left\{T^{D/M(\theta)/1}(q) > s\right\} ds + \dots \\
&\stackrel{(*)}{\leq} 1 \cdot \mathbb{P}\left\{T^{D/M(\theta)/1}(q) > 0\right\} + 1 \cdot \mathbb{P}\left\{T^{D/M(\theta)/1}(q) > 1\right\} + \dots \\
&= \sum_{n=0}^{\infty} \mathbb{P}\left\{T^{D/M(\theta)/1}(q) > n\right\}, \tag{B.45}
\end{aligned}$$

where $(*)$ follows by noting that $\mathbb{P}\{T^{D/M(\theta)/1}(q) > s_1\} \geq \mathbb{P}\{T^{D/M(\theta)/1}(q) > s_2\}$ for any $s_1 \leq s_2$. Then, it suffices to bound $\mathbb{P}\{T^{D/M(\theta)/1}(q) > n\}$ for each integer $n \in \mathbb{Z}_+$.

For $n = 0$, $\mathbb{P}\{T^{D/M(\theta)/1}(q) > 0\} \leq 1$ trivially. For $n = 1, 2, \dots$, the event $\{T^{D/M(\theta)/1}(q) > n\}$ happens if and only if the system has not been emptied before time n . Let Z_1, Z_2, \dots, Z_q denote the service times of the initial q old jobs, and let Z^1, Z^2, \dots denote the service times of the newly arriving jobs. Since jobs are served in a FCFS sequential fashion, it follows that

$$\begin{aligned} & \mathbb{P}\left\{T^{D/M(\theta)/1}(q) > n\right\} \\ &= \mathbb{P}\left\{Z_1 + Z_2 + \dots + Z_q + Z^1 > 1, \text{ and } Z_1 + Z_2 + \dots + Z_q + Z^1 + Z^2 > 2, \text{ and} \right. \\ &\quad \left. Z_1 + Z_2 + \dots + Z_q + Z^1 + Z^2 + Z^3 > 3, \dots, Z_1 + Z_2 + \dots + Z_q + Z^1 + Z^2 + \dots + Z^n > n\right\} \\ &\leq \mathbb{P}\left\{Z_1 + Z_2 + \dots + Z_q + Z^1 + Z^2 + \dots + Z^n > n\right\}. \end{aligned}$$

Since Z_1, Z_2, \dots, Z_q and Z^1, Z^2, \dots are all exponentially distributed with rate θ , the sum $Z_1 + Z_2 + \dots + Z_q + Z^1 + Z^2 + \dots + Z^n$ follows an Erlang distribution with rate parameter θ and shape parameter $q + n$ (positive integer). Thus, for $n = 1, 2, \dots$,

$$\begin{aligned} & \mathbb{P}\left\{Z_1 + Z_2 + \dots + Z_q + Z^1 + Z^2 + \dots + Z^n > n\right\} \stackrel{(1)}{=} 1 - e^{-\theta n} \sum_{i=q+n}^{\infty} \frac{(\theta n)^i}{i!} \\ &= e^{-\theta n} \sum_{i=0}^{q+n-1} \frac{(\theta n)^i}{i!} \\ &\stackrel{(2)}{\leq} \frac{(e\theta n)^{q+n-1} e^{-\theta n}}{(q+n-1)^{q+n-1}} \\ &= e^{-(\theta-1)n+q-1} \left(\frac{\theta n}{q+n-1}\right)^{q+n-1} \\ &\stackrel{(3)}{\leq} e^{-(\theta-1)n+q-1} \left(\frac{\theta(q+n-1)}{q+n-1}\right)^{q+n-1} \\ &= e^{-(\theta-1)n+q-1} \cdot \theta^{q+n-1}, \end{aligned}$$

where (1) follows from the c.d.f. of the Erlang distribution, (2) follows from the Chernoff

bound for a Poisson random variable $X \sim \text{Pois}(\lambda)$, namely, $\mathbb{P}(X \leq v) = e^{-\lambda} \sum_{i=0}^v \frac{\lambda^i}{i!} \leq \frac{(e\lambda)^v e^{-\lambda}}{v^v}$, $\forall v \in \mathbb{Z}_+$, and (3) follows by noting that $\theta n \leq \theta(q+n-1)$ for $q \geq 1$. Hence, for $n = 1, 2, \dots$,

$$\mathbb{P}\left\{T^{D/M(\theta)/1}(q) > n\right\} \leq e^{-(\theta-1)n+q-1} \cdot \theta^{q+n-1}. \quad (\text{B.46})$$

Then, from (B.45), it follows that

$$\begin{aligned} \bar{T}^{D/M(\theta)/1}(q) &\leq \sum_{n=0}^{\infty} \mathbb{P}\{T^{D/M(\theta)/1}(q) > n\} \\ &\leq 1 + \sum_{n=1}^{\infty} \left(e^{-(\theta-1)n+q-1} \cdot \theta^{q+n-1} \right) \\ &= 1 + (e\theta)^{q-1} \cdot \sum_{n=1}^{\infty} \left(e^{-(\theta-1)} \theta \right)^n \\ &= 1 + \frac{(e\theta)^q}{e^\theta - e\theta}, \quad \forall q = 1, 2, \dots. \end{aligned} \quad (\text{B.47})$$

where the last equality follows by noting that $e^{-(\theta-1)}\theta \leq 1$ for all $\theta > 1$, and thus the infinite summation converges.

Step 2: An upper bound for an initial state distributed according to the stationary distribution.

The stationary distribution of a $D/M/1$ queue with arrival rate 1 and service rate $\theta > 1$ is given by $p_\infty^{D/M(\theta)/1}(0) = 0$ and

$$p_\infty^{D/M(\theta)/1}(q) = (1-\delta)\delta^{q-1}, \quad \forall q = 1, 2, \dots, \quad (\text{B.48})$$

where $\delta \in (0, 1)$ is the smallest root to the equation $\delta = e^{-\theta(1-\delta)}$.

Using the stationary distribution in (B.48), together with (B.47), it follows that

$$\begin{aligned}
\mathbb{E} \left[\bar{T}^{D/M(\theta)/1} (Q_\infty^{D/M(\theta)/1}) \right] &= \sum_{q=0}^{\infty} p_\infty^{D/M(\theta)/1}(q) \cdot \bar{T}^{D/M(\theta)/1}(q) \\
&= \sum_{q=1}^{\infty} (1-\delta)\delta^{q-1} \cdot \bar{T}^{D/M(\theta)/1}(q) \\
&\leq \sum_{q=1}^{\infty} (1-\delta)\delta^{q-1} \left(1 + \frac{(e\theta)^q}{e^\theta - e\theta} \right) \\
&= (1-\delta) \left[\sum_{q=1}^{\infty} \delta^{q-1} + \sum_{q=1}^{\infty} \delta^{q-1} \frac{(e\theta)^q}{e^\theta - e\theta} \right] \\
&= (1-\delta) \left[\sum_{q=1}^{\infty} \delta^{q-1} + \frac{1}{\delta(e^\theta - e\theta)} \sum_{q=1}^{\infty} (e\theta\delta)^q \right] \\
&\stackrel{(\dagger)}{=} (1-\delta) \left[\frac{1}{1-\delta} + \frac{e\theta}{(e^\theta - e\theta)(1-e\theta\delta)} \right] \\
&= 1 + \frac{(1-\delta)e\theta}{(e^\theta - e\theta)(1-e\theta\delta)} \in (1, \infty),
\end{aligned}$$

where (\dagger) follows by noting that the infinite summation $\sum_{q=1}^{\infty} (e\theta\delta)^q$ converges, given the assumption that $\theta\delta < 1/e$.

To complete the proof, we verify Claim 16 below.

Proof of Claim 16.

Let $\xi(\theta) = \theta\delta$, where δ is the smallest solution to $\delta = e^{-\theta(1-\delta)}$. We sometimes express the dependence of δ on θ for clarity. Differentiating both sides of

$$\delta(\theta) = e^{-\theta(1-\delta(\theta))} \tag{B.49}$$

with respect to θ yields

$$\delta'(\theta) = e^{-\theta(1-\delta(\theta))} \cdot (-1 + \delta(\theta) + \theta\delta'(\theta)).$$

This implies that

$$\delta'(\theta) = \frac{(1-\delta)e^{-\theta(1-\delta)}}{\theta e^{-\theta(1-\delta)} - 1} = \frac{(1-\delta)e^{-\theta(1-\delta)}}{\theta\delta - 1},$$

where the second equality follows from (B.49). Differentiating $\xi(\theta)$ with respect to θ :

$$\begin{aligned} \xi'(\theta) &= \delta + \theta \cdot \delta'(\theta) \\ &= \delta + \theta \cdot \frac{(1-\delta)e^{-\theta(1-\delta)}}{\theta\delta - 1} \\ &= e^{-\theta(1-\delta)} + \theta \cdot \frac{(1-\delta)e^{-\theta(1-\delta)}}{\theta\delta - 1} \\ &= e^{-\theta(1-\delta)} \cdot \frac{\theta - 1}{\theta\delta - 1} \\ &= e^{-\theta(1-\delta)} \cdot \frac{\theta - 1}{\xi(\theta) - 1}, \end{aligned} \tag{B.50}$$

where the third equality follows from (B.49). We first note that $\xi(0) = 0$, $\xi(1) = 1$, and $\lim_{\theta \rightarrow \infty} \xi(\theta) = 0$ (because $\lim_{\theta \rightarrow \infty} \delta(\theta) = 0$, and thus $\lim_{\theta \rightarrow \infty} \theta \cdot e^{-\theta(1-\delta(\theta))} = \lim_{\theta \rightarrow \infty} \frac{1}{1-\delta(\theta)} [(1-\delta(\theta))\theta \cdot e^{-\theta(1-\delta(\theta))}] = 0$, noting that $(1-\delta(\theta))\theta \rightarrow \infty$ as $\theta \rightarrow \infty$ and exponential decay dominates linear growth).

- When $\theta \in (0, 1)$, suppose that there exists some $\theta_0 \in (0, 1)$ with $\xi(\theta_0) > 1$, then (B.50) implies that $\xi'(\theta_0) < 0$. This means that $\xi(\theta)$ would never exceed 1 on $(0, 1)$, i.e., $\xi(\theta) \leq 1$ for all $\theta \in (0, 1)$. Then, from (B.50), $\xi'(\theta) > 0$ for all $\theta \in (0, 1)$, implying that $\xi(\theta)$ is increasing in $\theta \in (0, 1)$.
- When $\theta \in (1, \infty)$, suppose that there exists some $\theta_m \in (1, \infty)$ with $\xi(\theta_m) > 1$, then (B.50) implies that $\xi'(\theta_m) > 0$, meaning that $\xi(\theta)$ keeps increasing for all $\theta \geq \theta_m$. This

contradicts with the fact that $\lim_{\theta \rightarrow \infty} \xi(\theta) = 0$. Therefore, $\xi(\theta) \leq 1$ for all $\theta \in (1, \infty)$. Then, from (B.50), $\xi'(\theta) < 0$ for all $\theta \in (1, \infty)$, implying that $\xi(\theta)$ is decreasing in $\theta \in (1, \infty)$.

Hence, we conclude that $\xi(\theta)$ increases in $\theta \in (0, 1)$ from $\xi(0) = 0$ to $\xi(1) = 1$, and decreases in $\theta \in (1, \infty)$ from $\xi(1) = 1$ to $\lim_{\xi \rightarrow \infty} \xi(\theta) = 0$. Note that $\xi(2.2) = 0.34 < 1/e = 0.37$, so $\xi(\theta) < 1/e$ for all $\theta > 2.2$. \blacksquare

B.2 Proofs from Section 2.4

B.2.1 Proof of Theorem 7

The proof is given in Section 2.5 in the main body, and uses the results proved below in Section B.3. \blacksquare

B.3 Proofs from Section 2.5

B.3.1 Proof of Proposition 11

By definition in (2.8), a trivial upper bound on the regret accumulated in the learning phase is the expected total abandonment cost over $[0, \tau]$ under policy $\pi_{LTS}(\tau)$, ignoring the expected total abandonment cost over $[0, \tau]$ incurred by the benchmark policy; that is,

$$\mathcal{R}_{a\mu}^{Learn}(T; \pi_{LTS}(\tau)) \leq \mathbb{E} \left[\sum_{j=1}^J a_j R_j(\tau; \pi_{LTS}(\tau)) \right]. \quad (\text{B.51})$$

Note that, for each class $j \in [J]$, the cumulative number of class j abandonments over $[0, \tau]$ is upper bounded by the cumulative number of class j arrivals over $[0, \tau]$ plus the number of

initial class j customers; that is,

$$R_j(\tau) \leq E_j(\tau) + X_j(0).$$

This, together with (B.51), imply that

$$\mathcal{R}_{a\mu}^{Learn}(T; \pi_{LTS}(\tau)) \leq \sum_{j=1}^J a_j \left(\mathbb{E}[E_j(\tau)] + X_j(0) \right). \quad (\text{B.52})$$

From Marshall (1973), it follows that

$$\mathbb{E}[E_j(t)] \leq \lambda_j t + U_j, \quad \forall j \in [J], \quad (\text{B.53})$$

where $U_j := \sup_{x \geq 0} \frac{G_j^a(x) - \lambda_j \int_{u=0}^x G_j^a(u) du}{G_j^a(x)} \in [0, \infty)$, $j \in [J]$. In particular, $U_j < \infty$ for all $j \in [J]$ is due to the light-tailed assumption of G_j^a , namely, $\int_{x=0}^{\infty} e^{\Upsilon^a x} dG_j^a(x) < \infty$ for some sufficiently small constant $\Upsilon^a > 0$ for all $j \in [J]$ (recalling Section 2.2.1.1).

Finally, combining (B.52) and (B.53) implies that

$$\mathcal{R}_{a\mu}^{Learn}(T; \pi_{LTS}(\tau)) \leq \left(\sum_{j=1}^J a_j \lambda_j \right) \tau + \sum_{j=1}^J a_j (X_j(0) + U_j).$$

■

B.3.2 Proof of Lemma 9

We prove the first part of the statement in (i) below, and the second part in (ii) below.

(i): Define π_0 as a “do-nothing” scheduling policy, wherein the server continuously completes any initial remaining job in service, if exists, but does not admit any jobs waiting in queues into service.

Fix a sample path $\omega \in \Omega$ (i.e., a realization of inter-arrival, service, and patience times

for each class). For any $\pi \in \Pi$, any initial state $Y(0) = (\alpha(0), X(0), \nu(0), \eta(0)) \in \mathbb{Y}$, and for all $t \geq 0$, we first note that

$$Q_j(t; \pi, Y(0), \omega) \leq_{st} Q_j(t; \pi_0, Y(0), \omega), \quad \forall j \in [J],$$

which implies that

$$\mathbb{E}[Q_j(t; \pi, Y(0))] \leq \mathbb{E}[Q_j(t; \pi_0, Y(0))], \quad \forall j \in [J]. \quad (\text{B.54})$$

Suppose we can establish the below claim, whose proof is postponed to the end of this subsection.

Claim 17. Define $e_j(t) := \mathbb{E}[E_j(t)]$, $t \geq 0$, for each $j \in [J]$. Given initial state $Y(0) = (\alpha(0), X(0), \nu(0), \eta(0)) \in \mathbb{Y}$, the following holds for all $t \geq 0$:

$$\mathbb{E}[Q_j(t; \pi_0, Y(0))] \leq X_j(0) + \int_0^t \bar{G}_j^r(t-u) de_j(u), \quad \forall j \in [J].$$

From Claim 17, for each $j \in [J]$,

$$\begin{aligned} & \mathbb{E}[Q_j(t; \pi_0, Y(0))] \\ & \leq X_j(0) + \int_0^t \bar{G}_j^r(t-u) de_j(u) \\ & = X_j(0) + \bar{G}_j^r(t-u)e_j(u) \Big|_{u=0}^t - \int_0^t e_j(u) \frac{d\bar{G}_j^r(t-u)}{du} du \\ & = X_j(0) + \bar{G}_j^r(0)e_j(t) - \bar{G}_j^r(t)e_j(0) - \int_0^t e_j(u) g_j^r(t-u) du \\ & = X_j(0) + e_j(t) - \int_0^t e_j(u) g_j^r(t-u) du \\ & = X_j(0) + e_j(t) - e_j(t) \int_0^t g_j^r(u) du + e_j(t) \int_0^t g_j^r(u) du - \int_0^t e_j(u) g_j^r(t-u) du \\ & = X_j(0) + e_j(t) \bar{G}_j^r(t) + \int_0^t (e_j(t) - e_j(t-u)) g_j^r(u) du. \end{aligned} \quad (\text{B.55})$$

By the key renewal theorem, $\frac{e_j(t)}{t} \rightarrow \lambda_j$, as $t \rightarrow \infty$. Moreover, since $\int_0^\infty \bar{G}_j^r(u)du = 1/\theta_j$, and

$$\int_0^\infty \bar{G}_j^r(u)du = \lim_{t \rightarrow \infty} \int_0^t \bar{G}_j^r(u)du = \lim_{t \rightarrow \infty} u\bar{G}_j^r(u)\Big|_0^t + \lim_{t \rightarrow \infty} \int_0^t ug_j^r(u)du = \lim_{t \rightarrow \infty} t\bar{G}_j^r(t) + \frac{1}{\theta_j}.$$

We conclude that $\lim_{t \rightarrow \infty} t\bar{G}_j^r(t) = 0$, and, therefore,

$$\sup_{t \geq 0} e_j(t)\bar{G}_j^r(t) = \sup_{t \geq 0} \frac{e_j(t)}{t} \cdot t\bar{G}_j^r(t) < \infty. \quad (\text{B.56})$$

Additionally, by the Blackwell renewal theorem, $e_j(t) - e_j(t-u) \rightarrow u\lambda_j$ for any $u > 0$, as $t \rightarrow \infty$. This implies that

$$\sup_{t \geq 0} \int_0^t (e_j(t) - e_j(t-u)) g_j^r(u)du < \infty. \quad (\text{B.57})$$

Substituting the bounds given in (B.56) and (B.57) into (B.55) implies $\mathbb{E}[Q_j(t; \pi_0, Y(0))] < \infty$ for each $j \in [J]$, and for any initial state $Y(0) \in \mathbb{Y}$. Then, from (B.54), for any $\pi \in \Pi$,

$$\mathbb{E}[Q_j(t; \pi, Y(0))] \leq \mathbb{E}[Q_j(t; \pi_0, Y(0))] < \infty, \quad \forall j \in [J].$$

Moreover, since $\mathbb{E}[B_j(t; \pi, Y(0))] \leq N$ for each $j \in [J]$, it follows that, for any $\pi \in \Pi$, any initial state $Y(0) \in \mathbb{Y}$, and for all $t \geq 0$,

$$\mathbb{E}[X_j(t; \pi, Y(0))] = \mathbb{E}[Q_j(t; \pi, Y(0))] + \mathbb{E}[B_j(t; \pi, Y(0))] < \infty, \quad \forall j \in [J].$$

Equivalently, there exists some finite positive constant \check{X} such that $\mathbb{E}[X_j(t; \pi)] \leq \check{X}$ for all $t \geq 0$, $j \in [J]$, and $\pi \in \Pi$.

(ii): By definition in (2.9), the regret accumulated in the exploitation phase is given by

$$\begin{aligned}
& \mathcal{R}_{a\mu}^{\text{Exploit}}(T; \pi_{LTS}(\tau) \mid \hat{\pi}_{a\mu}(\tau) = \pi_{a\mu}) \\
&= \mathbb{E} \left[\sum_{j=1}^J a_j \left((R_j(T; \pi_{LTS}(\tau)) - R_j(\tau; \pi_{LTS}(\tau))) - (R_j(T; \pi_{a\mu}) - R_j(\tau; \pi_{a\mu})) \right) \mid \hat{\pi}_{a\mu}(\tau) = \pi_{a\mu} \right] \\
&= \mathbb{E} \left[\sum_{j=1}^J a_j \left(R_j(T - \tau; \pi_{a\mu}, Y(\tau; \pi_{LTS}(\tau))) - R_j(T - \tau; \pi_{a\mu}, Y(\tau; \pi_{a\mu})) \right) \right], \tag{B.58}
\end{aligned}$$

where the last equality follows by regarding the state at time τ as an initial state by the Markov property. We discuss two cases below based on which condition in Assumption 3 (ii) is satisfied.

Case 1: If $\left| \mathbb{E} \left[\frac{R_j(t; \pi_{a\mu}, y)}{t} \right] - \mathbb{E} \left[\frac{R_j(t; \pi_{a\mu}, y')}{t} \right] \right| \leq \frac{\kappa}{t}$ for any two distinct initial states $y, y' \in \mathbb{Y}$, for each $j \in [J]$, and for all $t > 0$, then an upper bound can be directly obtained. Specifically, from (B.58),

$$\begin{aligned}
& \mathcal{R}_{a\mu}^{\text{Exploit}}(T; \pi_{LTS}(\tau) \mid \hat{\pi}_{a\mu}(\tau) = \pi_{a\mu}) \\
&= \mathbb{E} \left[\sum_{j=1}^J a_j \left(R_j(T - \tau; \pi_{a\mu}, Y(\tau; \pi_{LTS}(\tau))) - R_j(T - \tau; \pi_{a\mu}, Y(\tau; \pi_{a\mu})) \right) \right] \\
&\leq (T - \tau) \sum_{j=1}^J a_j \cdot \left| \mathbb{E} \left[\frac{R_j(T - \tau; \pi_{a\mu}, Y(\tau; \pi_{LTS}(\tau)))}{T - \tau} \right] - \mathbb{E} \left[\frac{R_j(T - \tau; \pi_{a\mu}, Y(\tau; \pi_{a\mu}))}{T - \tau} \right] \right| \\
&\leq (T - \tau) \sum_{j=1}^J a_j \cdot \frac{\kappa}{T - \tau} = \kappa \left(\sum_{j=1}^J a_j \right).
\end{aligned}$$

Case 2: If $\left| \mathbb{E} \left[\frac{D_j(t; \pi_{a\mu}, y)}{t} \right] - \mathbb{E} \left[\frac{D_j(t; \pi_{a\mu}, y')}{t} \right] \right| \leq \frac{\kappa}{t}$ for any two distinct initial states $y, y' \in \mathbb{Y}$, for each $j \in [J]$, and for all $t > 0$, then we can use conservation of mass (from Equation (17) in Puha and Ward (2022), recalling from Section 2.2.1.1 that the full system dynamics were not provided in the main body of the paper but can be found in Puha

and Ward (2022)):

$$R_j(t) = X_j(0) + E_j(t) - X_j(t) - D_j(t), \quad \forall j \in [J], \quad \forall t \geq 0, \quad (\text{B.59})$$

to express regret in terms of the cumulative number of service completions and the system size, and bound each term separately.

From (B.58) and (B.59),

$$\begin{aligned} & \mathcal{R}_{a\mu}^{\text{Exploit}}(T; \pi_{LTS}(\tau) \mid \hat{\pi}_{a\mu}(\tau) = \pi_{a\mu}) \\ &= \mathbb{E} \left[\sum_{j=1}^J a_j \left(R_j(T - \tau; \pi_{a\mu}, Y(\tau; \pi_{LTS}(\tau))) - R_j(T - \tau; \pi_{a\mu}, Y(\tau; \pi_{a\mu})) \right) \right] \\ &= \mathbb{E} \left[\sum_{j=1}^J a_j \left(X_j(\tau; \pi_{LTS}(\tau)) + E_j(T - \tau) - X_j(T - \tau; \pi_{a\mu}, Y(\tau; \pi_{LTS}(\tau))) \right. \right. \\ &\quad \left. \left. - D_j(T - \tau; \pi_{a\mu}, Y(\tau; \pi_{LTS}(\tau))) \right) \right] - \mathbb{E} \left[\sum_{j=1}^J a_j \left(X_j(\tau; \pi_{a\mu}) + E_j(T - \tau) \right. \right. \\ &\quad \left. \left. - X_j(T - \tau; \pi_{a\mu}, Y(\tau; \pi_{a\mu})) - D_j(T - \tau; \pi_{a\mu}, Y(\tau; \pi_{a\mu})) \right) \right] \\ &= \mathbb{E} \left[\sum_{j=1}^J a_j \left(\left(X_j(\tau; \pi_{LTS}(\tau)) - X_j(\tau; \pi_{a\mu}) \right) \right. \right. \\ &\quad \left. \left. - \left(X_j(T - \tau; \pi_{a\mu}, Y(\tau; \pi_{LTS}(\tau))) - X_j(T - \tau; \pi_{a\mu}, Y(\tau; \pi_{a\mu})) \right) \right) \right] \\ &\quad + \mathbb{E} \left[\sum_{j=1}^J a_j \left(D_j(T - \tau; \pi_{a\mu}, Y(\tau; \pi_{a\mu})) - D_j(T - \tau; \pi_{a\mu}, Y(\tau; \pi_{LTS}(\tau))) \right) \right] \\ &\leq \left| \mathbb{E} \left[\sum_{j=1}^J a_j \left(X_j(\tau; \pi_{LTS}(\tau)) - X_j(\tau; \pi_{a\mu}) \right) \right] \right| \\ &\quad + \left| \mathbb{E} \left[\sum_{j=1}^J a_j \left(X_j(T - \tau; \pi_{a\mu}, Y(\tau; \pi_{LTS}(\tau))) - X_j(T - \tau; \pi_{a\mu}, Y(\tau; \pi_{a\mu})) \right) \right] \right| \end{aligned}$$

$$\begin{aligned}
& + \left| \mathbb{E} \left[\sum_{j=1}^J a_j \left(D_j(T-\tau; \pi_{a\mu}, Y(\tau; \pi_{a\mu})) - D_j(T-\tau; \pi_{a\mu}, Y(\tau; \pi_{LTS}(\tau))) \right) \right] \right| \\
& \leq \sum_{j=1}^J a_j \cdot \max \left\{ \mathbb{E} [X_j(\tau; \pi_{LTS}(\tau))], \mathbb{E} [X_j(\tau; \pi_{a\mu})] \right\} \\
& \quad + \sum_{j=1}^J a_j \cdot \max \left\{ \mathbb{E} [X_j(T-\tau; \pi_{a\mu}, Y(\tau; \pi_{LTS}(\tau)))], \mathbb{E} [X_j(T-\tau; \pi_{a\mu}, Y(\tau; \pi_{a\mu}))] \right\} \\
& \quad + (T-\tau) \sum_{j=1}^J a_j \cdot \left| \mathbb{E} \left[\frac{D_j(T-\tau; \pi_{a\mu}, Y(\tau; \pi_{a\mu}))}{T-\tau} \right] - \mathbb{E} \left[\frac{D_j(T-\tau; \pi_{a\mu}, Y(\tau; \pi_{LTS}(\tau)))}{T-\tau} \right] \right| \\
& \leq 2 \sum_{j=1}^J a_j \cdot \check{X} + (T-\tau) \sum_{j=1}^J a_j \cdot \frac{\kappa}{T-\tau} = (2\check{X} + \kappa) \left(\sum_{j=1}^J a_j \right).
\end{aligned}$$

Combining the above two cases, we conclude that, regardless of which condition in Assumption 3 (ii) is satisfied,

$$\mathcal{R}_{a\mu}^{Exploit}(T; \pi_{LTS}(\tau) \mid \hat{\pi}_{a\mu}(\tau) = \pi_{a\mu}) \leq (2\check{X} + \kappa) \left(\sum_{j=1}^J a_j \right).$$

To complete the proof, we verify Claim 17 below.

B.3.2.1 Proof of Claim 17

We begin by noting that, under the “do-nothing” policy π_0 , the system operates like J independent $GI/GI/\infty$ queues, having service times equal to the patience times in the multiclass $GI/GI/N+GI$ queue. We employ a similar proof technique to the one used in establishing transient Little’s law (see, e.g., Equation (11) in Bertsimas and Mourtzinou (1997) for a $GI/GI/\infty$ queue that starts empty).

Fix a sample path $\omega \in \Omega$, and an initial state $Y(0) = (\alpha(0), X(0), \nu(0), \eta(0)) \in \mathbb{Y}$. For each $j \in [J]$, let $Q_j^1(t; \pi_0, Y(0), \omega)$ denote the number of class j initial customers that are still waiting in queue at time $t \geq 0$, under policy π_0 . Let $Q_j^2(t; \pi_0, \omega)$ denote the number of

class j customers (who arrived after time 0 and before time t) that are still waiting in queue at time $t \geq 0$, under policy π_0 . (We remove the dependence of $Q_j^2(t; \pi_0, \omega)$ on $Y(0)$ because the abandonment behavior of customers who arrived after time 0 does not depend on the initial customers under policy π_0 .) Then,

$$Q_j(t; \pi_0, Y(0), \omega) = Q_j^1(t; \pi_0, Y(0), \omega) + Q_j^2(t; \pi_0, \omega), \quad \forall t \geq 0, \quad \forall j \in [J]. \quad (\text{B.60})$$

It is clear from (2.3) that

$$Q_j^1(t; \pi_0, Y(0), \omega) \leq Q_j(t) \leq X_j(0), \quad \forall t \geq 0, \quad \forall j \in [J]. \quad (\text{B.61})$$

The next step is to bound $Q_j^2(t; \pi_0, \omega)$ for any $t \geq 0$ and for each $j \in [J]$. For any $t \geq 0$ and for each $j \in [J]$, define a deterministic function $f_j(u; t, \pi_0, \omega) : [0, t] \mapsto \mathbb{Z}_+$ that encodes the number of class j customers who arrived at $(u - du, u]$ and still wait in queue at time t under policy π_0 on sample path ω . Then, for each $j \in [J]$, and for all $t \geq 0$,

$$Q_j^2(t; \pi_0, \omega) = \int_0^t f_j(u; t, \pi_0, \omega) du.$$

Let $\{F_j(u; t, \pi_0) : 0 \leq u \leq t\}$ be the stochastic process corresponding to $f_j(u; t, \pi_0, \omega)$, then the above display implies

$$\mathbb{E} [Q_j^2(t; \pi_0)] = \mathbb{E} \left[\int_0^t F_j(u; t) du \right] \stackrel{(1)}{=} \int_0^t \mathbb{E} [F_j(u; t)] du \stackrel{(2)}{=} \int_0^t \bar{G}_j^r(t - u) de_j(u), \quad (\text{B.62})$$

where (1) follows from Fubini's theorem, and $\bar{G}_j^r(t - u)$ in (2) represents the probability of still waiting in queue at time t if entering the system at time $(u - du, u]$.

Using the bounds given in (B.61) and (B.62), together with (B.60), it follows that

$$\mathbb{E} [Q_j(t; \pi_0, Y(0))] = \mathbb{E} [Q_j^1(t; \pi_0, Y(0))] + \mathbb{E} [Q_j^2(t; \pi_0)] \leq X_j(0) + \int_0^t \bar{G}_j^r(t - u) de_j(u), \quad j \in [J].$$

■

B.3.3 Proof of Lemma 10

For each $j \in [J]$ and for any $z \geq 0$, given a sample size $D_j(\tau)$, it follows that

$$\begin{aligned}
& \mathbb{P} \{ |a_j \hat{\mu}_j(\tau) - a_j \mu_j| \geq z \} \\
&= \mathbb{P} \{ a_j \hat{\mu}_j(\tau) \geq a_j \mu_j + z \text{ or } a_j \hat{\mu}_j(\tau) \leq a_j \mu_j - z \} \\
&= \mathbb{P} \{ a_j \hat{\mu}_j(\tau) \geq a_j \mu_j + z \} + \mathbb{P} \{ a_j \hat{\mu}_j(\tau) \leq a_j \mu_j - z \} \\
&= \mathbb{P} \left\{ \hat{\mu}_j(\tau) \geq \mu_j + \frac{z}{a_j} \right\} + \mathbb{P} \left\{ \hat{\mu}_j(\tau) \leq \mu_j - \frac{z}{a_j} \right\} \\
&= \mathbb{P} \left\{ \frac{D_j(\tau)}{\sum_{i=1}^{D_j(\tau)} v_{j,i}} \geq \mu_j + \frac{z}{a_j} \right\} + \mathbb{P} \left\{ \frac{D_j(\tau)}{\sum_{i=1}^{D_j(\tau)} v_{j,i}} \leq \mu_j - \frac{z}{a_j} \right\}, \tag{B.63}
\end{aligned}$$

where the last equality follows from (2.7).

Next, we discuss two cases depending on the value of z .

- If $0 \leq z < a_j \mu_j$, then $\mu_j - \frac{z}{a_j} > 0$ in the second probability in (B.63). Hence, (B.63) can be rewritten as

$$\begin{aligned}
& \mathbb{P} \{ |a_j \hat{\mu}_j(\tau) - a_j \mu_j| \geq z \} \\
&= \mathbb{P} \left\{ \left(\frac{D_j(\tau)}{\sum_{i=1}^{D_j(\tau)} v_{j,i}} \right)^{-1} \leq \left(\mu_j + \frac{z}{a_j} \right)^{-1} \right\} + \mathbb{P} \left\{ \left(\frac{D_j(\tau)}{\sum_{i=1}^{D_j(\tau)} v_{j,i}} \right)^{-1} \geq \left(\mu_j - \frac{z}{a_j} \right)^{-1} \right\} \\
&= \mathbb{P} \left\{ \frac{\sum_{i=1}^{D_j(\tau)} v_{j,i}}{D_j(\tau)} - \frac{1}{\mu_j} \leq -\frac{\frac{z}{a_j}}{\mu_j \left(\mu_j + \frac{z}{a_j} \right)} \right\} + \mathbb{P} \left\{ \frac{\sum_{i=1}^{D_j(\tau)} v_{j,i}}{D_j(\tau)} - \frac{1}{\mu_j} \geq \frac{\frac{z}{a_j}}{\mu_j \left(\mu_j - \frac{z}{a_j} \right)} \right\} \\
&= \mathbb{P} \left\{ \frac{\sum_{i=1}^{D_j(\tau)} v_{j,i}}{D_j(\tau)} - \frac{1}{\mu_j} \geq \frac{\frac{z}{a_j}}{\mu_j \left(\mu_j + \frac{z}{a_j} \right)} \right\} + \mathbb{P} \left\{ \frac{\sum_{i=1}^{D_j(\tau)} v_{j,i}}{D_j(\tau)} - \frac{1}{\mu_j} \geq \frac{\frac{z}{a_j}}{\mu_j \left(\mu_j - \frac{z}{a_j} \right)} \right\} \\
&\leq 2 \cdot \mathbb{P} \left\{ \frac{\sum_{i=1}^{D_j(\tau)} v_{j,i}}{D_j(\tau)} - \frac{1}{\mu_j} \geq \frac{\frac{z}{a_j}}{\mu_j \left(\mu_j + \frac{z}{a_j} \right)} \right\}. \tag{B.64}
\end{aligned}$$

where the last inequality follows by noting that $\frac{\frac{z}{a_j}}{\mu_j \left(\mu_j - \frac{z}{a_j} \right)} \geq \frac{\frac{z}{a_j}}{\mu_j \left(\mu_j + \frac{z}{a_j} \right)} > 0$.

- If $z \geq a_j \mu_j$, then $\mu_j - \frac{z}{a_j} \leq 0$ and thus the second probability in (B.63) becomes zero.

Then, (B.63) can be written as

$$\begin{aligned}
\mathbb{P} \{ |a_j \hat{\mu}_j(\tau) - a_j \mu_j| \geq z \} &= \mathbb{P} \left\{ \frac{D_j(\tau)}{\sum_{i=1}^{D_j(\tau)} v_{j,i}} \geq \mu_j + \frac{z}{a_j} \right\} \\
&= \mathbb{P} \left\{ \left(\frac{D_j(\tau)}{\sum_{i=1}^{D_j(\tau)} v_{j,i}} \right)^{-1} \leq \left(\mu_j + \frac{z}{a_j} \right)^{-1} \right\} \\
&= \mathbb{P} \left\{ \frac{\sum_{i=1}^{D_j(\tau)} v_{j,i}}{D_j(\tau)} - \frac{1}{\mu_j} \leq -\frac{\frac{z}{a_j}}{\mu_j \left(\mu_j + \frac{z}{a_j} \right)} \right\} \\
&= \mathbb{P} \left\{ \frac{\sum_{i=1}^{D_j(\tau)} v_{j,i}}{D_j(\tau)} - \frac{1}{\mu_j} \geq \frac{\frac{z}{a_j}}{\mu_j \left(\mu_j + \frac{z}{a_j} \right)} \right\}. \tag{B.65}
\end{aligned}$$

Combining (B.64) and (B.65) corresponding to the two cases above, we have

$$\mathbb{P} \{ |a_j \hat{\mu}_j(\tau) - a_j \mu_j| \geq z \} \leq 2 \cdot \mathbb{P} \left\{ \frac{\sum_{i=1}^{D_j(\tau)} v_{j,i}}{D_j(\tau)} - \frac{1}{\mu_j} \geq \frac{\frac{z}{a_j}}{\mu_j \left(\mu_j + \frac{z}{a_j} \right)} \right\}. \tag{B.66}$$

Recalling the condition in Section 2.2.1.1 that $\log \left(\int_0^{H_j^s} e^{l(x - \frac{1}{\mu_j})} dG_j^s(x) \right) \leq l^2 (\sigma_j^s)^2 / 2$ for all $l \leq \Upsilon^s$, the service time random variables, $v_{j,i}$, for each i -th class j customer, all have mean $\frac{1}{\mu_j}$ and are sub-exponential with variance parameter $(\sigma_j^s)^2$ and scale parameter $\frac{1}{\Upsilon^s}$. By Bernstein's inequality that applies for a normalized sum of independent sub-exponential random variables X_1, \dots, X_n , each of which is sub-exponential with variance parameter σ^2 and scale parameter ν (e.g., Section 5.1.2 in ?, Corollary 2.8.3 in ?): for any $t > 0$,

$$\mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[X_i] \geq t \right\} \leq \exp \left(-\frac{n}{2} \min \left\{ \frac{t^2}{\sigma^2}, \frac{t}{\nu} \right\} \right).$$

Using this, it follows that

$$\mathbb{P} \left\{ \frac{\sum_{i=1}^{D_j(\tau)} v_{j,i}}{D_j(\tau)} - \frac{1}{\mu_j} \geq t \right\} \leq \exp \left(-\frac{D_j(\tau)}{2} \min \left\{ \frac{t^2}{(\sigma_j^s)^2}, \Upsilon^s t \right\} \right). \quad (\text{B.67})$$

Setting $t = \frac{\frac{z}{a_j}}{\mu_j \left(\mu_j + \frac{z}{a_j} \right)}$ in (B.67) yields

$$\begin{aligned} & \mathbb{P} \left\{ \frac{\sum_{i=1}^{D_j(\tau)} v_{j,i}}{D_j(\tau)} - \frac{1}{\mu_j} \geq \frac{\frac{z}{a_j}}{\mu_j \left(\mu_j + \frac{z}{a_j} \right)} \right\} \\ & \leq \exp \left(-\frac{D_j(\tau)}{2} \min \left\{ \frac{1}{(\sigma_j^s)^2} \left(\frac{\frac{z}{a_j}}{\mu_j \left(\mu_j + \frac{z}{a_j} \right)} \right)^2, \Upsilon^s \frac{\frac{z}{a_j}}{\mu_j \left(\mu_j + \frac{z}{a_j} \right)} \right\} \right) \\ & = \exp \left(-\frac{D_j(\tau)}{2} \min \left\{ \frac{1}{(\sigma_j^s)^2} \left(\frac{a_j \mu_j^2}{z} + \mu_j \right)^{-2}, \Upsilon^s \left(\frac{a_j \mu_j^2}{z} + \mu_j \right)^{-1} \right\} \right). \end{aligned}$$

Substituting the above into (B.66) implies that, for any given $D_j(\tau)$,

$$\begin{aligned} & \mathbb{P} \{ |a_j \hat{\mu}_j(\tau) - a_j \mu_j| \geq z \} \\ & \leq 2 \cdot \exp \left(-\frac{D_j(\tau)}{2} \min \left\{ \frac{1}{(\sigma_j^s)^2} \left(\frac{a_j \mu_j^2}{z} + \mu_j \right)^{-2}, \Upsilon^s \left(\frac{a_j \mu_j^2}{z} + \mu_j \right)^{-1} \right\} \right). \end{aligned}$$

Finally, conditional on $D_j(\tau) \geq \epsilon \tau$, the above display implies that, for any $\epsilon > 0$ and $z \geq 0$,

$$\begin{aligned} & \mathbb{P} \{ |a_j \hat{\mu}_j(\tau) - a_j \mu_j| \geq z \mid D_j(\tau) \geq \epsilon \tau \} \\ & \leq 2 \cdot \exp \left(-\frac{\epsilon \tau}{2} \min \left\{ \frac{1}{(\sigma_j^s)^2} \left(\frac{a_j \mu_j^2}{z} + \mu_j \right)^{-2}, \Upsilon^s \left(\frac{a_j \mu_j^2}{z} + \mu_j \right)^{-1} \right\} \right). \end{aligned}$$

■

B.3.4 Proof of Lemma 11

Note that

$$\begin{aligned}
& \mathbb{P}\{\hat{\pi}_{a\mu}(\tau) \neq \pi_{a\mu}\} \\
&= \mathbb{P}\{\hat{\pi}_{a\mu}(\tau) \neq \pi_{a\mu} \text{ and } D_j(\tau) \geq \epsilon\tau \text{ for all } j \in [J]\} + \mathbb{P}\{\hat{\pi}_{a\mu}(\tau) \neq \pi_{a\mu} \text{ and } D_j(\tau) < \epsilon\tau \text{ for some } j \in [J]\} \\
&= \mathbb{P}\{\hat{\pi}_{a\mu}(\tau) \neq \pi_{a\mu} \mid D_j(\tau) \geq \epsilon\tau \text{ for all } j \in [J]\} \cdot \mathbb{P}\{D_j(\tau) \geq \epsilon\tau \text{ for all } j \in [J]\} \\
&\quad + \mathbb{P}\{\hat{\pi}_{a\mu}(\tau) \neq \pi_{a\mu} \mid D_j(\tau) < \epsilon\tau \text{ for some } j \in [J]\} \cdot \mathbb{P}\{D_j(\tau) < \epsilon\tau \text{ for some } j \in [J]\} \\
&\leq \mathbb{P}\{\hat{\pi}_{a\mu}(\tau) \neq \pi_{a\mu} \mid D_j(\tau) \geq \epsilon\tau \text{ for all } j \in [J]\} \cdot 1 + 1 \cdot \mathbb{P}\{D_j(\tau) < \epsilon\tau \text{ for some } j \in [J]\} \\
&\leq \mathbb{P}\{\hat{\pi}_{a\mu}(\tau) \neq \pi_{a\mu} \mid D_j(\tau) \geq \epsilon\tau \text{ for all } j \in [J]\} + \sum_{j \in [J]} \mathbb{P}\{D_j(\tau) < \epsilon\tau\}, \tag{B.68}
\end{aligned}$$

where the last inequality follows by noting that the probability of event A or event B is upper bounded by the probability of event A plus the probability of event B.

Below we establish upper bounds for the probabilities in the two terms in (B.68) separately.

Step 1: Bounding $\mathbb{P}\{\hat{\pi}_{a\mu}(\tau) \neq \pi_{a\mu} \mid D_j(\tau) \geq \epsilon\tau \text{ for all } j \in [J]\}$.

Note that

$$\begin{aligned}
& \mathbb{P}\{\hat{\pi}_{a\mu}(\tau) = \pi_{a\mu} \mid D_j(\tau) \geq \epsilon\tau \text{ for all } j \in [J]\} \\
&\stackrel{(1)}{\geq} \mathbb{P}\left\{ \left| a_j \hat{\mu}_j(\tau) - a_j \mu_j \right| < \frac{\Delta}{2} \text{ for all } j \in [J] \mid D_j(\tau) \geq \epsilon\tau \text{ for all } j \in [J] \right\} \\
&\stackrel{(2)}{=} \prod_{j \in [J]} \mathbb{P}\left\{ \left| a_j \hat{\mu}_j(\tau) - a_j \mu_j \right| < \frac{\Delta}{2} \mid D_j(\tau) \geq \epsilon\tau \right\} \\
&= \prod_{j \in [J]} \left(1 - \mathbb{P}\left\{ \left| a_j \hat{\mu}_j(\tau) - a_j \mu_j \right| \geq \frac{\Delta}{2} \mid D_j(\tau) \geq \epsilon\tau \right\} \right) \\
&\stackrel{(3)}{\geq} \prod_{j \in [J]} \left(1 - 2 \cdot \exp\left(-\frac{\epsilon\tau}{2} \min\left\{ \frac{1}{(\sigma_j^s)^2} \left(\frac{2a_j \mu_j^2}{\Delta} + \mu_j \right)^{-2}, \Upsilon^s \left(\frac{2a_j \mu_j^2}{\Delta} + \mu_j \right)^{-1} \right\} \right) \right)
\end{aligned}$$

$$\begin{aligned}
&\geq \left(1 - 2 \cdot \max_{j \in [J]} \left\{ \exp \left(-\frac{\epsilon\tau}{2} \min \left\{ \frac{1}{(\sigma_j^s)^2} \left(\frac{2a_j\mu_j^2}{\Delta} + \mu_j \right)^{-2}, \Upsilon^s \left(\frac{2a_j\mu_j^2}{\Delta} + \mu_j \right)^{-1} \right\} \right) \right\} \right)^J \\
&= \left(1 - 2 \cdot \exp \left(- \min_{j \in [J]} \left\{ \frac{\epsilon\tau}{2} \min \left\{ \frac{1}{(\sigma_j^s)^2} \left(\frac{2a_j\mu_j^2}{\Delta} + \mu_j \right)^{-2}, \Upsilon^s \left(\frac{2a_j\mu_j^2}{\Delta} + \mu_j \right)^{-1} \right\} \right\} \right) \right)^J,
\end{aligned}$$

where (1) follows from Assumption 2, (2) follows because service time sequences for different classes are mutually independent, and (3) follows from Lemma 10. Additionally, note that, there exists some finite positive constant τ_1 such that the exponential term in the above is smaller than $1/2$ for all $\tau \geq \tau_1$. Then, by Bernoulli's inequality, the above display further implies that

$$\begin{aligned}
&\mathbb{P}\{\hat{\pi}_{a\mu}(\tau) = \pi_{a\mu} \mid D_j(\tau) \geq \epsilon\tau \text{ for all } j \in [J]\} \geq 1 - 2J \\
&\quad \cdot \exp \left(- \min_{j \in [J]} \left\{ \frac{\epsilon\tau}{2} \min \left\{ \frac{1}{(\sigma_j^s)^2} \left(\frac{2a_j\mu_j^2}{\Delta} + \mu_j \right)^{-2}, \Upsilon^s \left(\frac{2a_j\mu_j^2}{\Delta} + \mu_j \right)^{-1} \right\} \right\} \right), \quad \forall \tau \geq \tau_1.
\end{aligned}$$

Let

$$\zeta := \min_{j \in [J]} \left\{ \frac{\epsilon}{2} \min \left\{ \frac{1}{(\sigma_j^s)^2} \left(\frac{2a_j\mu_j^2}{\Delta} + \mu_j \right)^{-2}, \Upsilon^s \left(\frac{2a_j\mu_j^2}{\Delta} + \mu_j \right)^{-1} \right\} \right\}, \quad (\text{B.69})$$

then, $\mathbb{P}\{\hat{\pi}_{a\mu}(\tau) = \pi_{a\mu} \mid D_j(\tau) \geq \epsilon\tau \text{ for all } j \in [J]\} \geq 1 - 2J \cdot \exp(-\zeta\tau)$ for all $\tau \geq \tau_1$, which implies

$$\mathbb{P}\{\hat{\pi}_{a\mu}(\tau) \neq \pi_{a\mu} \mid D_j(\tau) \geq \epsilon\tau \text{ for all } j \in [J]\} \leq 2J \cdot \exp(-\zeta\tau), \quad \forall \tau \geq \tau_1. \quad (\text{B.70})$$

Step 2: Bounding $\mathbb{P}\{D_j(\tau) < \epsilon\tau\}$ for each $j \in [J]$.

In order to establish an upper bound for $\mathbb{P}\{D_j(\tau) < \epsilon\tau\}$ for each $j \in [J]$, it is equivalent to establishing a lower bound for $\mathbb{P}\{D_j(\tau) \geq \epsilon\tau\}$ for each $j \in [J]$.

For each $j \in [J]$, let $\varpi_j \geq 0$ be the shortest remaining service time among all the servers at time $(j-1)\frac{\tau}{J}$; in other words, $(j-1)\frac{\tau}{J} + \varpi_j$ is the earliest point after (and including) time $(j-1)\frac{\tau}{J}$ when a server becomes available. Then, it is clear that, for each $j \in [J]$,

$$\mathbb{P}\left\{\varpi_j \leq \frac{\tau}{2J}\right\} \geq \min_{j \in [J]} G_j^s\left(\frac{\tau}{2J}\right), \quad (\text{B.71})$$

where $\frac{\tau}{2J}$ is chosen arbitrarily and can be replaced by any value in $(0, \frac{\tau}{J})$.

Note that, for each $j \in [J]$,

$$\begin{aligned} \mathbb{P}\left\{D_j(\tau) \geq \epsilon\tau\right\} &\geq \mathbb{P}\left\{D_j\left(j\frac{\tau}{J}\right) - D_j\left((j-1)\frac{\tau}{J} + \varpi_j\right) \geq \epsilon\tau \text{ and } \varpi_j < \frac{\tau}{2J}\right\} \\ &= \mathbb{P}\left\{D_j\left(j\frac{\tau}{J}\right) - D_j\left((j-1)\frac{\tau}{J} + \varpi_j\right) \geq \epsilon\tau \mid \varpi_j < \frac{\tau}{2J}\right\} \cdot \mathbb{P}\left\{\varpi_j < \frac{\tau}{2J}\right\}. \end{aligned} \quad (\text{B.72})$$

A lower bound for the second probability in (B.72) is given by (B.71). The remainder of the proof is dedicated to establishing a lower bound for the first probability in (B.72).

Recalling our LTS policy (see Definition 9), in each cycle $j \in [J]$, only class j customers are admitted into service (under the dedicated policy π_j^d). Thus, we can use the service departure number in a FCFS $GI/GI/1+GI$ queue to provide a lower bound for $D_j(j\frac{\tau}{J}) - D_j((j-1)\frac{\tau}{J} + \varpi_j)$. For each $j \in [J]$, consider a FCFS $GI/GI/1+GI$ queue with inter-arrival, service, and patience time distributions G_j^a , G_j^s , and G_j^r , initialized from $q \in \mathbb{Z}_+$. Associated with this queue, let $D_j^{GI/GI/1+GI}(t; q)$ track the cumulative number of service departures over $[0, t]$, and $B_j^{GI/GI/1+GI}(t; q)$ track the cumulative server busy time over $[0, t]$. Define $\{v_{j,i} : i \geq 1\}$ and $\{w_{j,i} : i \geq 1\}$ as random variables for service and patience times, respectively. Let S_j be a renewal process such that $S_j(t) := \{n \geq 0 : \sum_{i=1}^n v_{j,i} \leq t\}$.

Then,

$$\begin{aligned}
& \mathbb{P} \left\{ D_j \left(j \frac{\tau}{J} \right) - D_j \left((j-1) \frac{\tau}{J} + \varpi_j \right) \geq \epsilon \tau \mid \varpi_j < \frac{\tau}{2J} \right\} \\
& \stackrel{(1)}{\geq} \mathbb{P} \left\{ D_j^{GI/GI/1+GI} \left(\frac{\tau}{J} - \varpi_j; Q_j((j-1) \frac{\tau}{J} + \varpi_j) \right) \geq \epsilon \tau \mid \varpi_j < \frac{\tau}{2J} \right\} \\
& \stackrel{(2)}{\geq} \mathbb{P} \left\{ D_j^{GI/GI/1+GI} \left(\frac{\tau}{2J}; 0 \right) \geq \epsilon \tau \right\} \\
& = \mathbb{P} \left\{ S_j \left(B_j^{GI/GI/1+GI} \left(\frac{\tau}{2J}; 0 \right) \right) \geq \epsilon \tau \right\}, \tag{B.73}
\end{aligned}$$

where (1) follows by noting that the service departure number over $[(j-1) \frac{\tau}{J} + \varpi_j, j \frac{\tau}{J}]$ stochastically dominates that in the FCFS $GI/GI/1+GI$ queue, having identical inter-arrival, service, and patience time distributions, and initialized from $Q_j((j-1) \frac{\tau}{J} + \varpi_j)$, because the original system has many servers, and (2) follows by noting that $D_j^{GI/GI/1+GI}(t; q)$ is a non-decreasing function of t and q .

Suppose we can establish the following claim, whose proof is delayed to the end of this subsection.

Claim 18. *There exists some finite positive constant τ_2 and $\psi \in (0, 1)$ such that $B_j^{GI/GI/1+GI}(\tau; 0) \geq \psi \tau$, for all $\tau \geq \tau_2$.*

Then, by Claim 18,

$$\mathbb{P} \left\{ S_j \left(B_j^{GI/GI/1+GI} \left(\frac{\tau}{2J}; 0 \right) \right) \geq \epsilon \tau \right\} \geq \mathbb{P} \left\{ S_j \left(\psi \frac{\tau}{2J} \right) \geq \epsilon \tau \right\}, \quad \forall \tau \geq \tau_2. \tag{B.74}$$

Substituting the above into (B.73) implies that

$$\mathbb{P} \left\{ D_j \left(j \frac{\tau}{J} \right) - D_j \left((j-1) \frac{\tau}{J} + \varpi_j \right) \geq \epsilon \tau \mid \varpi_j < \frac{\tau}{2J} \right\} \geq \mathbb{P} \left\{ S_j \left(\psi \frac{\tau}{2J} \right) \geq \epsilon \tau \right\}, \quad \forall \tau \geq \tau_2. \tag{B.75}$$

Next, we use a large deviations bound for the renewal process S_j to establish a lower

bound for $\mathbb{P}\left\{S_j\left(\psi \frac{\tau}{2J}\right) \geq \epsilon\tau\right\}$ in (B.75). Similar to (62) in Bell and Williams (2001) (in particular, one uses (184) in Appendix A in Bell and Williams (2001) with $\zeta^r(1) = 0$, $\delta^r = 0$, and $\chi^r t^r + 1$ in place of $\chi^r t$ in the right members there), for any $0 < \tilde{\epsilon} < \mu_j$, we have

$$\mathbb{P}\left\{S_j\left(\psi \frac{\tau}{2J}\right) < (\mu_j - \tilde{\epsilon}) \cdot \psi \frac{\tau}{2J}\right\} \leq \exp\left(-\left((\mu_j - \tilde{\epsilon}) \cdot \psi \frac{\tau}{2J} + 1\right) \Lambda_j^{s,*}\left(\frac{1}{\mu_j} \frac{1}{1 - \frac{\tilde{\epsilon}}{3\mu_j}}\right)\right), \quad \forall \tau \geq 0,$$

where $\Lambda_j^{s,*}(x) := \sup_{p \in \mathbb{R}} (px - \Lambda_j^s(p))$ is the Legendre-Fenchel transform of moment generating function of G_j^s and takes values in $[0, \infty]$. Let $\epsilon = (\mu_j - \tilde{\epsilon}) \frac{\psi}{2J}$ (which satisfies $0 < \epsilon < \mu_j \frac{\psi}{2J}$, given that $0 < \tilde{\epsilon} < \mu_j$), the above inequality implies

$$\begin{aligned} \mathbb{P}\left\{S_j\left(\psi \frac{\tau}{2J}\right) \geq \epsilon\tau\right\} &> 1 - \exp\left(-(\epsilon\tau + 1) \Lambda_j^{s,*}\left(\frac{1}{\mu_j} \frac{1}{1 - \frac{\mu_j - \frac{2J\epsilon}{\psi}}{3\mu_j}}\right)\right) \\ &= 1 - \exp\left(-(\epsilon\tau + 1) \Lambda_j^{s,*}\left(\frac{3}{2} \left(\mu_j + \frac{J\epsilon}{\psi}\right)^{-1}\right)\right) \\ &\geq 1 - \exp\left(-\epsilon \cdot \Lambda_j^{s,*}\left(\frac{3}{2} \left(\mu_j + \frac{J\epsilon}{\psi}\right)^{-1}\right) \tau\right), \quad \forall \tau \geq 0. \end{aligned} \quad (\text{B.76})$$

Substituting (B.76) into (B.75) implies, for any $0 < \epsilon < \mu_j \frac{\psi}{2J}$, and for all $\tau \geq \tau_2$,

$$\begin{aligned} \mathbb{P}\left\{D_j\left(j \frac{\tau}{J}\right) - D_j\left((j-1) \frac{\tau}{J} + \varpi_j\right) \geq \epsilon\tau \mid \varpi_j < \frac{\tau}{2J}\right\} \\ > 1 - \exp\left(-\epsilon \cdot \Lambda_j^{s,*}\left(\frac{3}{2} \left(\mu_j + \frac{J\epsilon}{\psi}\right)^{-1}\right) \tau\right). \end{aligned} \quad (\text{B.77})$$

Now, substituting the lower bounds in (B.71) and (B.77) into (B.72) yields, for any $0 < \epsilon < \mu_j \frac{\psi}{2J}$, and for all $\tau \geq \tau_2$,

$$\mathbb{P}\left\{D_j(\tau) \geq \epsilon\tau\right\} > \left(1 - \exp\left(-\epsilon \cdot \Lambda_j^{s,*}\left(\frac{3}{2} \left(\mu_j + \frac{J\epsilon}{\psi}\right)^{-1}\right) \tau\right)\right) \cdot \left(\min_{j \in [J]} G_j^s\left(\frac{\tau}{2J}\right)\right),$$

implying that

$$\begin{aligned}
& \mathbb{P} \{ D_j(\tau) < \epsilon \tau \} \\
& \leq 1 - \left(1 - \exp \left(-\epsilon \cdot \Lambda_j^{s,*} \left(\frac{3}{2} \left(\mu_j + \frac{J\epsilon}{\psi} \right)^{-1} \right) \tau \right) \right) \cdot \left(\min_{j \in [J]} G_j^s \left(\frac{\tau}{2J} \right) \right) \\
& = \left(1 - \min_{j \in [J]} G_j^s \left(\frac{\tau}{2J} \right) \right) + \left(\min_{j \in [J]} G_j^s \left(\frac{\tau}{2J} \right) \right) \cdot \exp \left(-\epsilon \cdot \Lambda_j^{s,*} \left(\frac{3}{2} \left(\mu_j + \frac{J\epsilon}{\psi} \right)^{-1} \right) \tau \right) \\
& \stackrel{(*)}{\leq} \left(1 - \min_{j \in [J]} G_j^s \left(\frac{\tau}{2J} \right) \right) + \exp \left(-\epsilon \cdot \Lambda_j^{s,*} \left(\frac{3}{2} \left(\mu_j + \frac{J\epsilon}{\psi} \right)^{-1} \right) \tau \right) \\
& = \max_{j \in [J]} \bar{G}_j^s \left(\frac{\tau}{2J} \right) + \exp \left(-\epsilon \cdot \Lambda_j^{s,*} \left(\frac{3}{2} \left(\mu_j + \frac{J\epsilon}{\psi} \right)^{-1} \right) \tau \right), \tag{B.78}
\end{aligned}$$

where $(*)$ follows by noting that $G_j^s(\cdot) \leq 1$ for all $j \in [J]$. Recalling that $v_{j,1}$, for each $j \in [J]$, is a random variable distributed according to G_j^s , we note that

$$\begin{aligned}
\bar{G}_j^s \left(\frac{\tau}{2J} \right) &= \mathbb{P} \left\{ v_{j,1} > \frac{\tau}{2J} \right\} = \mathbb{P} \left\{ \exp (\Upsilon^s v_{j,1}) > \exp \left(\Upsilon^s \frac{\tau}{2J} \right) \right\} \\
&\leq \frac{\mathbb{E} [\exp (\Upsilon^s v_{j,1})]}{\exp (\Upsilon^s \frac{\tau}{2J})} = \exp \left(\Lambda_j^s (\Upsilon^s) \right) \exp \left(-\frac{\Upsilon^s}{2J} \tau \right),
\end{aligned}$$

where the inequality follows by Markov's inequality, and $\exp (\Lambda_j^s (\Upsilon^s))$ is finite, recalling the finite exponential moment condition in Section 2.2.1.1. Substituting the upper bound in the above into (B.78) yields, for each $j \in [J]$, for any $0 < \epsilon < \mu_j \frac{\psi}{2J}$,

$$\begin{aligned}
\mathbb{P} \{ D_j(\tau) < \epsilon \tau \} &\leq \max_{j \in [J]} \exp \left(\Lambda_j^s (\Upsilon^s) \right) \exp \left(-\frac{\Upsilon^s}{2J} \tau \right) \\
&\quad + \exp \left(-\epsilon \cdot \Lambda_j^{s,*} \left(\frac{3}{2} \left(\mu_j + \frac{J\epsilon}{\psi} \right)^{-1} \right) \tau \right), \quad \forall \tau \geq \tau_2. \tag{B.79}
\end{aligned}$$

Step 3: Combining the results derived from Steps 1 and 2 to establish an upper bound for (B.68).

Substituting the bounds in (B.70) and (B.79) into (B.68) implies that, for any $0 < \epsilon <$

$\min_{j \in [J]} \mu_j \frac{\psi}{2J}$, and for all $\tau \geq \max\{\tau_1, \tau_2\}$,

$$\begin{aligned}
& \mathbb{P}\{\hat{\pi}_{a\mu}(\tau) \neq \pi_{a\mu}\} \\
& \leq 2J \cdot \exp(-\zeta\tau) + \sum_{j \in [J]} \left\{ \max_{j \in [J]} \exp(\Lambda_j^s(\Upsilon^s)) \exp\left(-\frac{\Upsilon^s}{2J}\tau\right) + \exp\left(-\epsilon \cdot \Lambda_j^{s,*}\left(\frac{3}{2}\left(\mu_j + \frac{J\epsilon}{\psi}\right)^{-1}\right)\tau\right) \right\} \\
& = 2J \cdot \exp(-\zeta\tau) + \max_{j \in [J]} \exp(\Lambda_j^s(\Upsilon^s)) J \cdot \exp\left(-\frac{\Upsilon^s}{2J}\tau\right) + \sum_{j \in [J]} \exp\left(-\epsilon \cdot \Lambda_j^{s,*}\left(\frac{3}{2}\left(\mu_j + \frac{J\epsilon}{\psi}\right)^{-1}\right)\tau\right) \\
& \leq \left(3 + \max_{j \in [J]} \exp(\Lambda_j^s(\Upsilon^s))\right) J \cdot \exp\left(-\min\left\{\zeta, \frac{\Upsilon^s}{2J}, \epsilon \cdot \min_{j \in [J]} \Lambda_j^{s,*}\left(\frac{3}{2}\left(\mu_j + \frac{J\epsilon}{\psi}\right)^{-1}\right)\right\} \cdot \tau\right) \\
& = \left(3 + \max_{j \in [J]} \exp(\Lambda_j^s(\Upsilon^s))\right) J \cdot e^{-\ell\tau},
\end{aligned}$$

where

$$\begin{aligned}
\ell &:= \min \left\{ \zeta, \frac{\Upsilon^s}{2J}, \epsilon \cdot \min_{j \in [J]} \Lambda_j^{s,*}\left(\frac{3}{2}\left(\mu_j + \frac{J\epsilon}{\psi}\right)^{-1}\right) \right\} \\
&= \min \left\{ \min_{j \in [J]} \left\{ \frac{\epsilon}{2} \min \left\{ \frac{1}{(\sigma_j^s)^2} \left(\frac{2a_j \mu_j^2}{\Delta} + \mu_j \right)^{-2}, \Upsilon^s \left(\frac{2a_j \mu_j^2}{\Delta} + \mu_j \right)^{-1} \right\}, \frac{\Upsilon^s}{2J}, \epsilon \cdot \min_{j \in [J]} \Lambda_j^{s,*}\left(\frac{3}{2}\left(\mu_j + \frac{J\epsilon}{\psi}\right)^{-1}\right) \right\} \right\} \\
&= \min \left\{ \frac{\epsilon}{2} \cdot \min_{j \in [J]} \frac{1}{(\sigma_j^s)^2} \left(\frac{2a_j \mu_j^2}{\Delta} + \mu_j \right)^{-2}, \frac{\epsilon}{2} \Upsilon^s \cdot \min_{j \in [J]} \left(\frac{2a_j \mu_j^2}{\Delta} + \mu_j \right)^{-1}, \frac{\Upsilon^s}{2J}, \epsilon \cdot \min_{j \in [J]} \Lambda_j^{s,*}\left(\frac{3}{2}\left(\mu_j + \frac{J\epsilon}{\psi}\right)^{-1}\right) \right\}.
\end{aligned}$$

Since this holds for any $0 < \epsilon < \min_{j \in [J]} \mu_j \frac{\psi}{2J}$, we arbitrarily choose $\epsilon = \min_{j \in [J]} \mu_j \frac{\psi}{3J}$, then

$$\begin{aligned}
\ell &= \min \left\{ \frac{\psi}{6J} \left(\min_{j \in [J]} \mu_j \right) \cdot \min_{j \in [J]} \frac{1}{(\sigma_j^s)^2} \left(\frac{2a_j \mu_j^2}{\Delta} + \mu_j \right)^{-2}, \frac{\psi}{6J} \left(\min_{j \in [J]} \mu_j \right) \Upsilon^s \cdot \min_{j \in [J]} \left(\frac{2a_j \mu_j^2}{\Delta} + \mu_j \right)^{-1}, \frac{\Upsilon^s}{2J}, \right. \\
&\quad \left. \frac{\psi}{3J} \left(\min_{j \in [J]} \mu_j \right) \cdot \min_{j \in [J]} \Lambda_j^{s,*}\left(\frac{3}{2}\left(\mu_j + \frac{1}{3} \min_{j \in [J]} \mu_j\right)^{-1}\right) \right\}.
\end{aligned}$$

To complete the proof, we verify Claim 18 below.

B.3.4.1 Proof of Claim 18.

We start by defining an auxiliary process that tracks the length of time a class $j \in [J]$ customer of infinite patience arriving at time $t \geq 0$ has to wait for service in the FCFS $GI/GI/1+GI$ queue having inter-arrival, service, and patience time distributions G_j^a , G_j^s , and G_j^r , and initialized from zero:

$$V_j^{GI/GI/1+GI}(t; 0) := \sum_{i=1}^{E_j(t)} v_{j,i} \mathbb{1} \left\{ V_j^{GI/GI/1+GI}(\alpha_{j,i}-; 0) < w_{j,i} \right\} - B_j^{GI/GI/1+GI}(t; 0), \quad (\text{B.80})$$

where $\alpha_{j,i}$, $v_{j,i}$, and $w_{j,i}$ represent the inter-arrival, service, and patience times of the i -th class j customer. This process defined in (B.80) is often called the offered waiting time process.

By Lemma 2 in Baccelli et al. (1984), when $\frac{\lambda_j}{\mu_j} \bar{G}_j^r(\infty) < 1$ (which is true because $1/\theta_j$ is assumed to be finite), $\{V_j^{GI/GI/1+GI}(t; 0) : t \geq 0\}$ is ergodic and thus has a limiting distribution with finite expectation (see Remark 1 right after Equation (3.11) in Baccelli et al. (1984)). Thus,

$$\lim_{t \rightarrow \infty} \frac{1}{t} V_j^{GI/GI/1+GI}(t; 0) = 0. \quad (\text{B.81})$$

Additionally, note that

$$\begin{aligned} & \frac{1}{t} \sum_{i=1}^{E_j(t)} v_{j,i} \mathbb{1} \left\{ V_j^{GI/GI/1+GI}(\alpha_{j,i}-; 0) < w_{j,i} \right\} \\ &= \frac{E_j(t)}{t} \cdot \frac{1}{E_j(t)} \sum_{i=1}^{E_j(t)} v_{j,i} \mathbb{1} \left\{ V_j^{GI/GI/1+GI}(\alpha_{j,i}-; 0) < w_{j,i} \right\}, \end{aligned} \quad (\text{B.82})$$

where

$$\lim_{t \rightarrow \infty} \frac{E_j(t)}{t} = \lambda_j, \quad (\text{B.83})$$

and

$$\begin{aligned} \lim_{t \rightarrow \infty} \frac{1}{E_j(t)} \sum_{i=1}^{E_j(t)} v_{j,i} \mathbb{1} \left\{ V_j^{GI/GI/1+GI}(\alpha_{j,i}-; 0) < w_{j,i} \right\} &\stackrel{(*)}{=} \mathbb{E}[v_{j,1}] \cdot \mathbb{P} \left\{ V_{j,\infty}^{GI/GI/1+GI} < w_{j,1} \right\} \\ &= \frac{1}{\mu_j} \cdot \mathbb{P} \left\{ V_{j,\infty}^{GI/GI/1+GI} < w_{j,1} \right\} > 0, \end{aligned} \quad (\text{B.84})$$

where $(*)$ follows from the fact that $v_{j,i}$, $w_{j,i}$, and $V_j^{GI/GI/1+GI}(\alpha_{j,i}-; 0)$ are mutually independent. To see this, let $\mathcal{F}_{j,i-1} \subset \mathcal{F}$ be a filtration that includes all the inter-arrival, service, and patience time information of the first $i-1$ customers as well as the inter-arrival time of the i -th customer. Then, $V_j^{GI/GI/1+GI}(\alpha_{j,i}-; 0)$ is $\mathcal{F}_{j,i-1}$ -measurable, yet $v_{j,i}$ and $w_{j,i}$ are independent of $\mathcal{F}_{j,i-1}$.

Substituting (B.83) and (B.84) into (B.82) yields

$$\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{i=1}^{E_j(t)} v_{j,i} \mathbb{1} \left\{ V_j^{GI/GI/1+GI}(\alpha_{j,i}-; 0) < w_{j,i} \right\} = \frac{\lambda_j}{\mu_j} \cdot \mathbb{P} \left\{ V_{j,\infty}^{GI/GI/1+GI} < w_{j,1} \right\} > 0. \quad (\text{B.85})$$

Finally, dividing both sides of (B.80) by t , letting $t \rightarrow \infty$, and using (B.81) and (B.85), we conclude that

$$\lim_{t \rightarrow \infty} \frac{1}{t} B_j^{GI/GI/1+GI}(t; 0) = \frac{\lambda_j}{\mu_j} \cdot \mathbb{P} \left\{ V_{j,\infty}^{GI/GI/1+GI} < w_{j,1} \right\} =: 2\psi,$$

with $2\psi \in (0, 1]$ from (B.85) and the fact that $B_j^{GI/GI/1+GI}(t; 0)/t \leq 1$. Therefore, for any

$\varepsilon > 0$, there always exists some finite time (depending on ε) such that $\frac{1}{t}B_j^{GI/GI/1+GI}(t; 0) \in [2\psi - \varepsilon, 2\psi + \varepsilon]$ for all time onwards. Setting $\varepsilon = \psi$, there exists some finite positive constant τ_2 such that $\frac{1}{t}B_j^{GI/GI/1+GI}(t; 0) \in [\psi, 3\psi]$ for all $t \geq \tau_2$. This implies that

$$\frac{1}{t}B_j^{GI/GI/1+GI}(t; 0) \geq \psi, \quad \forall t \geq \tau_2,$$

which establishes the statement. \blacksquare

B.3.5 Proof of Proposition 12

Substituting the bounds in Lemma 9 and Lemma 11 into (2.11):

$$\begin{aligned} & \mathcal{R}_{a\mu}^{Exploit}(T; \pi_{LTS}(\tau)) \\ & \leq (2\check{X} + \kappa) \left(\sum_{j=1}^J a_j \right) + \left(3 + \max_{j \in [J]} \exp(\Lambda_j^s(\Upsilon^s)) \right) J e^{-\ell\tau} \cdot \left(\left(\sum_{j=1}^J a_j \lambda_j \right) T + \sum_{j=1}^J a_j (X_j(0) + U_j) \right) \\ & =: C_0 + C_1 e^{-\ell\tau} + C_2 T e^{-\ell\tau}, \end{aligned}$$

where $C_0 = (2\check{X} + \kappa) \left(\sum_{j=1}^J a_j \right)$, $C_1 = \left(3 + \max_{j \in [J]} \exp(\Lambda_j^s(\Upsilon^s)) \right) J \left(\sum_{j=1}^J a_j (X_j(0) + U_j) \right)$, and $C_2 = \left(3 + \max_{j \in [J]} \exp(\Lambda_j^s(\Upsilon^s)) \right) J \left(\sum_{j=1}^J a_j \lambda_j \right)$. \blacksquare

B.4 Proofs from Section 2.6

B.4.1 Justification of the Non-Idling Assumption in the Multiclass $M/M/N+M$ Queue

We justify our restriction to non-idling scheduling policies by the following result.

Lemma 57. *In the multiclass $M/M/N+M$ queue, when $\sum_{j=1}^J \frac{\lambda_j}{\mu_j} > 1$, there exists a non-idling scheduling policy that achieves the optimality of (2.6).*

B.4.1.1 Proof of Lemma 57

In the multiclass $M/M/N+M$ queue, the cumulative number of service completions and abandonments for each class $j \in [J]$ can be expressed by $D_j(t) = \tilde{D}_j \left(\int_0^t \mu_j B_j(s) ds \right)$ and $R_j(t) = \tilde{R}_j \left(\int_0^t \theta_j Q_j(s) ds \right)$, where $\{\tilde{D}_j\}_{j \in [J]}$ and $\{\tilde{R}_j\}_{j \in [J]}$ are mutually independent standard Poisson processes with unit rate. By conservation of mass (from Equation (17) in Puha and Ward (2022), recalling from Section 2.2.1.1 that the full system dynamics were not provided in the main body of the paper but can be found in Puha and Ward (2022)), for each $j \in [J]$, and for all $t \geq 0$,

$$X_j(t) = X_j(0) + E_j(t) - D_j(t) - R_j(t),$$

which, together with (2.3), can be equivalently written as

$$Q_j(t) + \tilde{R}_j \left(\int_0^t \theta_j Q_j(s) ds \right) = X_j(0) + E_j(t) - B_j(t) - \tilde{D}_j \left(\int_0^t \mu_j B_j(s) ds \right).$$

Taking expectation on both sides of the above display yields

$$\mathbb{E} \left[Q_j(t) + \int_0^t \theta_j Q_j(s) ds \right] = X_j(0) + \lambda_j t - \mathbb{E} \left[B_j(t) + \int_0^t \mu_j B_j(s) ds \right].$$

The long-run average version of the above display is given by

$$\begin{aligned} & \limsup_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \left[Q_j(T) + \int_0^T \theta_j Q_j(s) ds \right] \\ &= \limsup_{T \rightarrow \infty} \frac{1}{T} (X_j(0) + \lambda_j T) - \limsup_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \left[B_j(T) + \int_0^T \mu_j B_j(s) ds \right], \end{aligned}$$

which implies that

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \left[\int_0^T \theta_j Q_j(s) ds \right] = \lambda_j - \limsup_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \left[\int_0^T \mu_j B_j(s) ds \right], \quad \forall j \in [J], \quad (\text{B.86})$$

noting that $\mathbb{E}[Q_j(s)] \leq \mathbb{E}[X_j(s)] \leq \check{X}$ for all $s \geq 0$ (recalling Lemma 9), $B_j(s) \leq N < \infty$ for all $s \geq 0$, and $X_j(0) < \infty$.

Moreover, the associated cost in (2.6) under any policy $\pi \in \Pi$ in the $M/M/N+M$ queue is

$$\mathcal{C}(\pi) = \limsup_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \left[\sum_{j=1}^J a_j \tilde{R}_j \left(\int_0^T \theta_j Q_j(s; \pi) ds \right) \right] = \limsup_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \left[\sum_{j=1}^J a_j \theta_j \int_0^T Q_j(s; \pi) ds \right].$$

Using (B.86), the above display can be rewritten as

$$\mathcal{C}(\pi) = \sum_{j=1}^J a_j \cdot \left(\lambda_j - \limsup_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \left[\int_0^T \mu_j B_j(s; \pi) ds \right] \right). \quad (\text{B.87})$$

The system cost in (B.87) under an idling policy can be further reduced by increasing the busy server number for classes that are under-supplied, given that $\sum_{j=1}^J \frac{\lambda_j}{\mu_j} > 1$. Hence, an optimal solution to (B.87) must be non-idling. ■

B.5 Proofs from Section 2.7

The proofs in this section heavily rely on the results in Puha and Ward (2022). The reader is advised to have a copy of that paper on-hand.

B.5.1 Preliminaries

We superscript all quantities that depend on N by N , and further use an overbar to denote the quantities scaled by N ; in particular, $\bar{X}^N = X^N/N$, $\bar{\nu}^N = \nu^N/N$, and $\bar{\eta}^N = \eta^N/N$. We

define \bar{V}^N exactly as in Theorem 2 in Puha and Ward (2022). Note that \bar{V}^N includes components $\bar{X}^N, \bar{\nu}^N, \bar{\eta}^N, \bar{Q}^N = (\bar{Q}_1^N, \dots, \bar{Q}_J^N), \bar{R}^N = (\bar{R}_1^N, \dots, \bar{R}_J^N), \bar{D}^N = (\bar{D}_1^N, \dots, \bar{D}_J^N)$.

B.5.2 Proof of Proposition 13

The proof for statements (i) and (ii) for the multiclass $GI/GI/N+M$ queue can be directly found in Proposition 5.1 and Theorem 5.1 in Atar et al. (2014), respectively. The rest of the proof is devoted to the multiclass $GI/M/N+GI$ setting.

(i): Under Assumption 5 and the conditions on the model inputs (specified in Section 2.2.1.1), one can check that Assumptions 1-5 in Puha and Ward (2022) hold. Thus, we can appeal to Theorem 1 in Puha and Ward (2022) (which allows the policy to depend on N , as is the case here with π^N for the N -server queue).

By Theorem 1 in Puha and Ward (2022), all limit points of $\left\{ \left(\bar{X}^N, \bar{\nu}^N, \bar{\eta}^N \right) \right\}_{N \in \mathbb{N}}$ are fluid model solutions almost surely for the arrival function $\bar{E} = (\bar{E}_1, \dots, \bar{E}_J)$, where

$$\bar{E}_j(t) = \lambda_j t, \quad \forall t \geq 0, \quad \forall j \in [J]. \quad (\text{B.88})$$

A fluid model solution is defined in Definition 4 in Puha and Ward (2022). In particular, from Lemma 10 in the proof of Theorem 1 in Puha and Ward (2022), we have that, any distributional limit point \bar{V} of $\{\bar{V}^N\}_{N \in \mathbb{N}}$ has components $\bar{E}, \bar{X}, \bar{B}, \bar{Q}, \bar{R}, \bar{D}$ that satisfy, for all $t \geq 0$ and $j \in [J]$,

$$\bar{X}_j(t) = \bar{X}_j(0) + \bar{E}_j(t) - \bar{R}_j(t) - \bar{D}_j(t), \quad (\text{B.89})$$

and

$$\bar{D}_j(t) = \int_0^t \mu_j \bar{B}_j(u) du. \quad (\text{B.90})$$

Since \bar{V} is a distributional limit point of $\{\bar{V}^N\}_{N \in \mathbb{N}}$, $\sum_{j=1}^J a_j \bar{R}_j(T)$ is a distributional limit point of $\mathcal{C}_T^N(\pi^N)/N$. Then, Fatou's lemma implies the sequence of costs, $\{\mathcal{C}_T^N(\pi^N)\}_{N \in \mathbb{N}}$ satisfies

$$\liminf_{N \rightarrow \infty} \frac{\mathcal{C}_T^N(\pi^N)}{N} = \liminf_{N \rightarrow \infty} \mathbb{E} \left[\frac{1}{N} \sum_{j=1}^J a_j R_j^N(T; \pi^N) \right] \geq \sum_{j=1}^J a_j \bar{R}_j(T).$$

Dividing by T and letting $T \rightarrow \infty$ in the above yields

$$\lim_{T \rightarrow \infty} \liminf_{N \rightarrow \infty} \frac{\mathcal{C}_T^N(\pi^N)}{NT} \geq \sum_{j=1}^J a_j \left(\lim_{T \rightarrow \infty} \frac{\bar{R}_j(T)}{T} \right). \quad (\text{B.91})$$

Suppose we can establish the following claim, whose proof is postponed to the end.

Claim 19. *For each $j \in [J]$, $\lim_{t \rightarrow \infty} \frac{\bar{X}_j(t)}{t} = 0$.*

Then, from Claim 19, dividing by t in (B.89) and taking the limit as $t \rightarrow \infty$ yields

$$\lambda_j = \lim_{t \rightarrow \infty} \frac{\bar{R}_j(t)}{t} + \lim_{t \rightarrow \infty} \frac{\bar{D}_j(t)}{t}, \quad \forall j \in [J]. \quad (\text{B.92})$$

From Proposition 1 in Long et al. (2020),

$$\lim_{t \rightarrow \infty} \bar{B}_j(t) = b_j, \quad \forall j \in [J],$$

for some $b \in \mathbb{B}$. This, together with (B.90), imply that

$$\lim_{t \rightarrow \infty} \frac{\bar{D}_j(t)}{t} = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \mu_j \bar{B}_j(t) = b_j \mu_j, \quad \forall j \in [J],$$

for some $b \in \mathbb{B}$. Substituting the above into (B.92) implies

$$\lim_{t \rightarrow \infty} \frac{\bar{R}_j(t)}{t} = \lambda_j - b_j \mu_j, \quad \forall j \in [J], \quad (\text{B.93})$$

for some $b \in \mathbb{B}$. Hence, from (B.91),

$$\lim_{T \rightarrow \infty} \liminf_{N \rightarrow \infty} \frac{\mathcal{C}_T^N(\pi^N)}{NT} \geq \sum_{j=1}^J a_j (\lambda_j - b_j \mu_j),$$

for some $b \in \mathbb{B}$.

Finally, from the definition of b^\star given in (2.15),

$$\sum_{j=1}^J a_j (\lambda_j - b_j^\star \mu_j) \leq \sum_{j=1}^J a_j (\lambda_j - b_j \mu_j),$$

for any $b \in \mathbb{B}$. Therefore,

$$\lim_{T \rightarrow \infty} \liminf_{N \rightarrow \infty} \frac{\mathcal{C}_T^N(\pi^N)}{NT} \geq \sum_{j=1}^J a_j (\lambda_j - b_j^\star \mu_j).$$

(ii): Section 4.4 in Puha and Ward (2022) shows how to apply Theorem 1 in Puha and Ward (2022) to conclude that all limit points of $\{\bar{X}^N, \bar{\nu}^N, \bar{\eta}^N\}_{N \in \mathbb{N}}$ are static priority fluid model solutions almost surely for the arrival function \bar{E} (given in (B.88)), where a static priority fluid model solution satisfies the additional equations (72)-(73) in Puha and Ward (2022) based on Definition 4 in Puha and Ward (2022).

Similar to part (i), $\sum_{j=1}^J a_j \bar{R}_j(T; \pi_{a\mu})$ is a distributional limit point of $\mathcal{C}_T^N(\pi_{au})/N$. Since $R_j^N(T; \pi_{a\mu})/N \leq (X_j(0) + E_j^N(T))/N$ and $\mathbb{E}[E_j^N(T)/N] \leq \lambda_j T + U_j$, for each $j \in [J]$ (from the second inequality in Proposition 11, recalling that U_j is defined in Proposition 11), the dominated convergence theorem implies the sequence of costs, $\{\mathcal{C}_T^N(\pi_{a\mu})\}_{N \in \mathbb{N}}$ satisfies

$$\lim_{N \rightarrow \infty} \frac{\mathcal{C}_T^N(\pi_{a\mu})}{N} = \lim_{N \rightarrow \infty} \mathbb{E} \left[\frac{1}{N} \sum_{j=1}^J a_j R_j^N(T; \pi_{a\mu}) \right] = \sum_{j=1}^J a_j \bar{R}_j(T; \pi_{a\mu}),$$

Then, to complete the proof, this is sufficient to show

$$\lim_{T \rightarrow \infty} \frac{\bar{R}_j(T; \pi_{a\mu})}{T} = \lambda_j - b_j^* \mu_j. \quad (\text{B.94})$$

From Claim 19, dividing by t in (B.89) and taking the limit as $t \rightarrow \infty$ yields

$$\lambda_j = \lim_{t \rightarrow \infty} \frac{\bar{R}_j(t; \pi_{a\mu})}{t} + \lim_{t \rightarrow \infty} \frac{\bar{D}_j(t; \pi_{a\mu})}{t}, \quad \forall j \in [J]. \quad (\text{B.95})$$

From Proposition 2 in Long et al. (2020),

$$\lim_{t \rightarrow \infty} \bar{B}_j(t; \pi_{a\mu}) = b_j^*, \quad \forall j \in [J],$$

where b^* is defined in (2.15). This, together with (B.90), imply that

$$\lim_{t \rightarrow \infty} \frac{\bar{D}_j(t; \pi_{a\mu})}{t} = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \mu_j \bar{B}_j(t; \pi_{a\mu}) dt = b_j^* \mu_j, \quad \forall j \in [J].$$

Substituting the above into (B.95) implies

$$\lim_{t \rightarrow \infty} \frac{\bar{R}_j(t; \pi_{a\mu})}{t} = \lambda_j - b_j^* \mu_j, \quad \forall j \in [J],$$

which establishes (B.94).

To complete the proof, we prove Claim 19 as follows.

B.5.2.1 Proof of Claim 19

Recall that \bar{V} is a distributional limit point of $\{\bar{V}^N\}_{N \in \mathbb{N}}$, so each component satisfies Lemma 10 in Puha and Ward (2022). The proof often appeals to Lemma 10 in Puha and Ward (2022).

By Lemma 10 in Puha and Ward (2022), $\bar{\eta}$ satisfies Equation (47) in Puha and Ward

(2022); that is, for any bounded continuous function f , for each $j \in [J]$, and for all $t \geq 0$,

$$\int_0^{H_j^r} f(x) \bar{\eta}_j(t)(dx) = \int_0^{H_j^r} f(x+t) \frac{1 - G_j^r(x+t)}{1 - G_j^r(x)} \bar{\eta}_j(0)(dx) + \lambda_j \int_0^t f(t-u) (1 - G_j^r(t-u)) du.$$

Letting $t \rightarrow \infty$, the first integral in the above converges to zero by dominated convergence (since f is bounded, $\int_0^{H_j^r} \bar{\eta}_j(0)(dx) < \infty$ and $\lim_{t \rightarrow \infty} \frac{1 - G_j^r(x+t)}{1 - G_j^r(x)} = 0$), and the second to $\lambda_j \int_0^\infty f(u) (1 - G_j^r(u)) du$. Because f is arbitrary, we then have

$$\lim_{t \rightarrow \infty} \bar{\eta}_j(t)(dx) = \lambda_j (1 - G_j^r(x)) dx, \quad \forall x \in \mathbb{R}_+, \forall j \in [J]. \quad (\text{B.96})$$

Moreover, by Lemma 10 in Puha and Ward (2022), \bar{X} satisfies Equation (35) in Puha and Ward (2022); that is, for each $j \in [J]$, and for all $t \geq 0$,

$$\bar{X}_j(t) \leq \int_0^{H_j^s} \bar{\nu}_j(t)(dx) + \int_0^{H_j^r} \bar{\eta}_j(t)(dx).$$

Dividing by t , noting that $\bar{X}_j(t) \geq 0$ for each $j \in [J]$ and for all $t \geq 0$, and taking the limit as $t \rightarrow \infty$ in the above display implies

$$\lim_{t \rightarrow \infty} \frac{\bar{X}_j(t)}{t} \leq \lim_{t \rightarrow \infty} \frac{1}{t} \left(\int_0^{H_j^s} \bar{\nu}_j(t)(dx) + \int_0^{H_j^r} \bar{\eta}_j(t)(dx) \right) = 0, \quad \forall j \in [J],$$

because $\int_0^{H_j^s} \bar{\nu}_j(t)(dx) \leq 1$ for each $t \geq 0$ (from Equation (36) in Puha and Ward (2022)), and $\int_0^{H_j^r} \bar{\eta}_j(t)(dx) \rightarrow \int_0^{H_j^r} \lambda_j (1 - G_j^r(x)) dx = \lambda_j / \theta_j$ as $t \rightarrow \infty$ (from (B.96)). Hence, $\lim_{t \rightarrow \infty} \frac{\bar{X}_j(t)}{t} = 0$ for each $j \in [J]$.

■

APPENDIX C

APPENDIX FOR CHAPTER 3

In this chapter, we provide proofs for the results stated in Chapter 3. The proofs of these results are in the order in which they appear in Chapter 3.

C.1 The Fluid Model for γ

We write the fluid model equations and write fluid model solutions for $\gamma > 0$ in this section. We refer the reader to Section 3.1 in Puha and Ward (2021) for details. Given a Polish space \mathbb{S} , we use $\mathbf{C}(\mathbb{S})$ to denote the set of functions having domain \mathbb{R}_+ and range \mathbb{S} that are continuous in time.

The fluid model for γ has as an input a non-decreasing function $E(t) = \gamma t$, $t \geq 0$. We set $\mathbb{X} := \mathbb{R}_+ \times \mathbf{M}[0, H^s] \times \mathbf{M}[0, H^r]$, endowed with the product topology in a Polish space. To define the fluid model for γ , we consider $(X, \nu, \eta) \in \mathbf{C}(\mathbb{X})$ such that

$$\langle 1_{\{x\}}, \eta(0) \rangle = 0, \text{ for all } x \in [0, H^r], \quad (\text{C.1})$$

and such that for each $t \geq 0$,

$$\langle 1, \nu(t) \rangle \leq X(t) \leq \langle 1, \nu(t) \rangle + \langle 1, \eta(t) \rangle, \quad (\text{C.2})$$

$$\langle 1, \nu(t) \rangle \leq 1, \quad (\text{C.3})$$

$$\int_0^t \langle h^s, \nu(u) \rangle du < \infty \text{ and } \int_0^t \langle h^r, \eta(u) \rangle du < \infty. \quad (\text{C.4})$$

Given $(X, \nu, \eta) \in \mathbf{C}(\mathbb{X})$ satisfying (C.1)-(C.4)), we define auxiliary functions B, Q, χ, R ,

D , and K in $\mathbf{C}(\mathbb{R}_+)$ and I in $\mathbf{C}(\mathbb{R}_+)$ as follows: for each $t \geq 0$,

$$B(t) = \langle 1, \nu(t) \rangle, \quad (\text{C.5})$$

$$Q(t) = X(t) - B(t), \quad (\text{C.6})$$

$$\chi(t) = \inf\{x \geq 0 : \langle 1_{[0,x]}, \eta(t) \rangle \geq Q(t)\}, \quad (\text{C.7})$$

$$R(t) = \int_0^t \left(\int_0^{\chi(u)} h^r(w) \eta(u)(dw) \right) du, \quad (\text{C.8})$$

$$D(t) = \int_0^t \langle h^s, \nu(u) \rangle du, \quad (\text{C.9})$$

$$K(t) = B(t) + D(t) - B(0), \quad (\text{C.10})$$

$$I(t) = 1 - B(t). \quad (\text{C.11})$$

Then B , Q , χ , R , D , K , and I are fluid analogs of the busy server, the queue length, the waiting time of the HL fluid in queue, the reneging, the departure, the entry-into-service, and the idleness processes, respectively.

Further some additional properties and equations that should be satisfied by $(X, \nu, \eta) \in \mathbf{C}(\mathbb{X})$ are as follows: for any continuous and bounded function f having domain \mathbb{R}_+ , for each $t \geq 0$,

$$K \text{ is non-decreasing,} \quad (\text{C.12})$$

$$X(t) = X(0) + E(t) - R(t) - D(t), \quad (\text{C.13})$$

$$\langle f, \nu(t) \rangle = \left\langle f(\cdot + t) \frac{\bar{G}^s(\cdot + t)}{\bar{G}^s(\cdot)}, \nu(0) \right\rangle + \int_0^t f(t-u) \bar{G}^s(t-u) dK(u), \quad (\text{C.14})$$

$$\langle f, \eta(t) \rangle = \left\langle f(\cdot + t) \frac{\bar{G}^r(\cdot + t)}{\bar{G}^r(\cdot)}, \eta(0) \right\rangle + \gamma \int_0^t f(t-u) \bar{G}^r(t-u) du. \quad (\text{C.15})$$

Definition 19. A fluid model solution for $\gamma > 0$ is (X, ν, η) that satisfies (C.1)-(C.4),

and (C.12)-(C.15).

Definition 20. A non-idling fluid model solution for $\gamma > 0$ is (X, ν, η) that satisfies Definition 19 and the following non-idling condition for each $t \geq 0$:

$$I(t) = (1 - X(t))^+. \quad (\text{C.16})$$

C.2 Proofs of Lemmas

Throughout this section, we fix $N \in \mathbb{N}$, $p \in (0, 1]$, $\pi_p^N = (\mathbb{S}^N, \{\mathbb{P}_y^N\}_{y \in \mathbb{S}^N}) \in \Pi_p^N$ and a compatible initial distribution ς^N , and we let Y^N be the state process for (π_p^N, ς^N) . For each Borel subset A of \mathbb{S}^N , define $L_0^N(A) := \mathbb{P}_\varsigma^N(Y^N(0) \in A)$ and for $t > 0$, define

$$L_t^N(A) := \frac{1}{t} \int_0^t \mathbb{P}_\varsigma^N(Y^N(s) \in A) ds.$$

Then, for each $t > 0$, L_t^N is a probability measure on \mathbb{S}^N , and we use the notation $Y_t^N(0) = (\alpha_t^N(0), X_t^N(0), \nu_t^N(0), \eta_t^N(0))$ to denote a random vector with law L_t^N . If $f : \mathbb{S}^N \rightarrow \mathbb{R}_+$ is measurable, then, for each $t \geq 0$, the expected value of $f(Y_t^N(0))$ is given by

$$\mathbb{E}_{L_t^N}^N[f(Y_t^N(0))] = \frac{1}{t} \int_0^t \mathbb{E}_\varsigma^N[f(Y^N(s))] ds.$$

Due to Lemma 4.4 in Kang et al. (2012), $\sup_{t \geq 0} \mathbb{E}_\varsigma^N[\langle 1, \eta^N(t) \rangle] < \infty$. Thus, for all $t \geq 0$,

$$\mathbb{E}_{L_t^N}^N[\langle 1, \eta_t^N(0) \rangle] < \infty,$$

In addition, as shown in the proof of Proposition 4.1 in Atar et al. (2014),

$$\lim_{t \rightarrow \infty} \mathbb{E}_{L_t^N}^N[\langle 1, \eta_t^N(0) \rangle] = p\lambda^N\theta^{-1}. \quad (\text{C.17})$$

C.2.1 Proof of Lemma 12

From Lemma 4.8 in Kang et al. (2012) (which does not require the non-idling condition), the family of probability measures $\{L_t^N\}_{t \geq 0}$ is tight. Since $\{Y^N(t)\}_{t \geq 0}$ is a Feller Markov process such that $\mathbb{E}_\zeta^N [\langle 1, \bar{\eta}^N(0) \rangle] < \infty$ (from Assumption 7), the Krylov-Bogoliubov theorem (see Corollary 3.1.2 in Da Prato et al. (1996)) implies that any limit point ξ^N of $\{L_t^N\}_{t \geq 0}$ is a stationary distribution for π_p^N such that if $Y_\infty^N(0)$ is a random variable with distribution ξ^N , then $\langle 1, \eta_\infty^N(0) \rangle$ has finite expected value under ξ^N . Moreover, the marginal distribution of $\alpha_\infty^N(0)$ is such that the corresponding arrival process E_∞^N is a stationary renewal process with rate $p\lambda^N$. Thus, $\langle 1, \eta_\infty^N(0) \rangle$ is equal in distribution to the stationary number of customers in a non-idling infinite server queue with arrival rate $p\lambda^N$ and service rate θ . Hence, for the admissible control policy π_p^N with initial distribution ξ^N , we have $\mathbb{E}_\xi^N [\langle 1, \eta_\infty^N(0) \rangle] = p\lambda^N\theta^{-1}$ by Little's law. \blacksquare

C.2.2 Proof of Lemma 13

Since π_p^N is fixed, we suppress the process dependence on π_p^N throughout the proof. From the proof of Lemma 12, for each $N \in \mathbb{N}$, any limit point ξ^N of $\{L_t^N\}_{t \geq 0}$ is a stationary distribution. Let $\{\tau(n)\}_{n \in \mathbb{N}} \subset \mathbb{R}_+$ be a strictly increasing subsequence along which $L_{\tau(n)}^N$ converges to ξ^N . On that subsequence,

$$\begin{aligned} & \lim_{n \rightarrow \infty} \frac{1}{\tau(n)} \mathbb{E}_\zeta^N \left[\int_0^{\tau(n)} g_U(\bar{B}^N(t)) dt \right] \\ & \stackrel{(1)}{=} \lim_{n \rightarrow \infty} \frac{1}{\tau(n)} \int_0^{\tau(n)} \mathbb{E}_\zeta^N [g_U(\bar{B}^N(t))] dt \\ & \stackrel{(2)}{=} \lim_{n \rightarrow \infty} \mathbb{E}_{L_{\tau(n)}}^N [g_U(\bar{B}_{\tau(n)}^N(0))] \\ & \stackrel{(3)}{=} \mathbb{E}_\xi^N [g_U(\bar{B}_\infty^N(0))]. \end{aligned} \tag{C.18}$$

where (1) follows by Fubini's theorem, (2) follows by definition of $L_{\tau(n)}^N$, and (3) follows because g_U is continuous and bounded. Since $\{\tau(n)\}_{n \in \mathbb{N}}$ is arbitrary, (C.18) implies (3.12).

For each $t \geq 0$, let

$$M^N(t) := R^N(t) - \int_0^t \left\langle 1_{[0, \chi^N(u-)]} h^r, \eta^N(u) \right\rangle du.$$

From Lemma 4 in Puha and Ward (2021), $M^N(\cdot)$ is a martingale (with respect to the filtration $\{\mathcal{F}_t^N\}_{t \geq 0}$ defined in Puha and Ward (2021)). To see this, for $(x, u) \in [0, H^r] \times \mathbb{R}_+$, let $f^N(x, u) = 1_{[0, \chi^N(u-)]}(x)$ and note that f^N is an almost surely bounded, measurable, real-valued function on $[0, H^r] \times \mathbb{R}_+$ so that Lemma 4 in Puha and Ward (2021) applies. Hence, for $n \in \mathbb{N}$,

$$\begin{aligned} \mathbb{E}_\zeta^N \left[\frac{R^N(\tau(n))}{\tau(n)} \right] &= \mathbb{E}_\zeta^N \left[\frac{1}{\tau(n)} \int_0^{\tau(n)} \left\langle 1_{[0, \chi^N(u-)]} h^r, \eta^N(u) \right\rangle du \right] \\ &\stackrel{(2)}{=} \mathbb{E}_\zeta^N \left[\frac{1}{\tau(n)} \int_0^{\tau(n)} \left\langle 1_{[0, \chi^N(u)]} h^r, \eta^N(u) \right\rangle du \right] \\ &\stackrel{(3)}{=} \frac{1}{\tau(n)} \int_0^{\tau(n)} \mathbb{E}_\zeta^N \left[\left\langle 1_{[0, \chi^N(u)]} h^r, \eta^N(u) \right\rangle \right] du \\ &\stackrel{(4)}{=} \mathbb{E}_{L_{\tau(n)}}^N \left[\left\langle 1_{[0, \chi_{\tau(n)}^N(0)]} h^r, \eta_{\tau(n)}^N(0) \right\rangle \right], \end{aligned} \quad (\text{C.19})$$

where (2) follows by noting that $\{t \geq 0 : \chi^N(t-) \neq \chi^N(t)\}$ has Lebesgue measure zero, (3) follows by Fubini's theorem, and (4) follows by definition of $L_{\tau(n)}^N$.

By assumption, $\left\langle 1, \eta_{\tau(n)}^N(0) \right\rangle \Rightarrow \left\langle 1, \eta_\infty^N(0) \right\rangle$ as $n \rightarrow \infty$. From Lemma 12, $\mathbb{E}_\xi^N \left[\left\langle 1, \eta_\infty^N(0) \right\rangle \right] = p\lambda^N \theta^{-1} < \infty$. Hence, by (C.17), $\lim_{n \rightarrow \infty} \mathbb{E}_{L_{\tau(n)}}^N \left[\left\langle 1, \eta_{\tau(n)}^N(0) \right\rangle \right] = \mathbb{E}_\xi^N \left[\left\langle 1, \eta_\infty^N(0) \right\rangle \right]$, implying that the sequence $\left\{ \left\langle 1, \eta_{\tau(n)}^N(0) \right\rangle \right\}_{n \in \mathbb{N}}$ is uniformly integrable. Note that

$$\left\langle 1_{[0, \chi_t^N(0)]} h^r, \eta_t^N(0) \right\rangle \leq \|h^r\|_\infty \left\langle 1, \eta_t^N(0) \right\rangle, \text{ for all } t \geq 0. \quad (\text{C.20})$$

Then, uniform integrability of $\left\{ \left\langle 1_{[0, \chi_{\tau(n)}^N(0)]} h^r, \eta_{\tau(n)}^N(0) \right\rangle \right\}_{n \in \mathbb{N}}$ follows from (C.20), uniform integrability of $\left\{ \left\langle 1, \eta_{\tau(n)}^N(0) \right\rangle \right\}_{N \in \mathbb{N}}$, and boundedness of h^r . Suppose we can show the following claim:

Claim 20. $\left\langle 1_{[0, \chi_{\tau(n)}^N(0)]} h^r, \eta_{\tau(n)}^N(0) \right\rangle \Rightarrow \left\langle 1_{[0, \chi_\infty^N(0)]} h^r, \eta_\infty^N(0) \right\rangle$, as $n \rightarrow \infty$.

Thus, (C.19), Claim 20 and uniform integrability of $\left\{ \left\langle 1_{[0, \chi_{\tau(n)}^N(0)]} h^r, \eta_{\tau(n)}^N(0) \right\rangle \right\}_{n \in \mathbb{N}}$ imply

$$\lim_{n \rightarrow \infty} \mathbb{E}_\zeta^N \left[\frac{R^N(\tau(n))}{\tau(n)} \right] = \mathbb{E}_\xi^N \left[\left\langle 1_{[0, \chi_\infty^N(0)]} h^r, \eta_\infty^N(0) \right\rangle \right]. \quad (\text{C.21})$$

Then, since $\{\tau(n)\}_{n \in \mathbb{N}}$ is arbitrary, (3.11) holds. ■

To complete the proof, we verify Claim 20 as follows.

C.2.2.1 Proof of Claim 20

By assumption, $\eta_{\tau(n)}^N(0) \Rightarrow \eta_\infty^N(0)$ as $n \rightarrow \infty$. Without loss of generality, we may assume that this convergence is almost sure and we may fix an w such that the stated convergence holds and evaluate all random elements at this w . We have $\eta_{\tau(n)}^N(0) = \sum_{i=1}^{\langle 1, \eta_{\tau(n)}^N(0) \rangle} \delta_{w_i^n}$, $\eta_\infty^N(0) = \sum_{i=1}^{\langle 1, \eta_\infty^N(0) \rangle} \delta_{w_i}$, and $\langle 1, \eta_{\tau(n)}^N(0) \rangle \Rightarrow \langle 1, \eta_\infty^N(0) \rangle$ as $n \rightarrow \infty$. Since $\langle 1, \eta_{\tau(n)}^N(0) \rangle$ is non-negative integer valued, it follows that there exists a finite positive \underline{n} such that for all $n \geq \underline{n}$, we have $\langle 1, \eta_{\tau(n)}^N(0) \rangle = \langle 1, \eta_\infty^N(0) \rangle$. Hence, it follows that $w_i^n \rightarrow w_i$ for all $1 \leq i \leq \langle 1, \eta_\infty^N(0) \rangle$, as $n \rightarrow \infty$. Then, due to the continuity of h^r , as $n \rightarrow \infty$.

$$\begin{aligned} \left\langle 1_{[0, \chi_{\tau(n)}^N(0)]} h^r, \eta_{\tau(n)}^N(0) \right\rangle &= \sum_{i=1}^{\langle 1, \eta_{\tau(n)}^N(0) \rangle} h^r(w_i^n) \\ &\rightarrow \sum_{i=1}^{\langle 1, \eta_\infty^N(0) \rangle} h^r(w_i) = \left\langle 1_{[0, \chi_\infty^N(0)]} h^r, \eta_\infty^N(0) \right\rangle. \end{aligned}$$

■

C.2.3 Proof of Remark 13

Let

$$\begin{aligned}\mathcal{C}_\varsigma^N(\pi_p^N, T) := & \frac{1}{T} \mathbb{E}_\varsigma^N \left[a \left(\bar{E}^N(T) - \bar{E}_p^N(T) + \bar{R}^N(T) \right) \right. \\ & \left. + \int_0^T g_U \left(\bar{B}^N(t) \right) dt \right],\end{aligned}$$

then $\mathcal{C}_\varsigma^N(\pi_p^N) = \limsup_{T \rightarrow \infty} \mathcal{C}_\varsigma^N(\pi_p^N, T)$. Let $\{\tau(n_i)\}_{i=1}^\infty$ be a subsequence of $\{\tau(n)\}_{n \in \mathbb{N}}$ such that $\{\mathcal{C}_\varsigma^N(\pi_p^N, \tau(n_i))\}_{i=1}^\infty$ converges and the limit is equal to the limit superior. Recalling that the sequence of probability measures $\{L_{\tau(n_i)}\}_{i=1}^\infty$ is tight, we denote by ξ^N a limit point. Then, there exists a further subsequence $\{\tau(n_{i_k})\}_{k=1}^\infty$ such that $L_{\tau(n_{i_k})}^N \rightarrow \xi^N$ as $k \rightarrow \infty$. From (C.18) and (C.21),

$$\begin{aligned}\mathcal{C}^N(\pi_p^N) &= \lim_{k \rightarrow \infty} \mathcal{C}_\varsigma^N(\pi_p^N, \tau(n_{i_k})) \\ &= \lim_{k \rightarrow \infty} \mathbb{E}_\varsigma^N \left[a \frac{\bar{E}^N(\tau(n_{i_k})) - \bar{E}_p^N(\tau(n_{i_k}))}{\tau(n_{i_k})} \right] + \lim_{k \rightarrow \infty} \mathbb{E}_\varsigma^N \left[a \frac{\bar{R}^N(\tau(n_{i_k}))}{\tau(n_{i_k})} \right] \\ &\quad + \lim_{k \rightarrow \infty} \frac{1}{\tau(n_{i_k})} \mathbb{E}_\varsigma^N \left[\int_0^{\tau(n_{i_k})} g_U \left(\bar{B}^N(t) \right) dt \right] \\ &= \mathbb{E}_\xi^N \left[a(1-p)\bar{\lambda}^N + a \left\langle 1_{[0, \chi_\infty^N(0)]} h^r, \bar{\eta}_\infty^N(0) \right\rangle + g_U \left(\bar{B}_\infty^N(0) \right) \right],\end{aligned}$$

which establishes the statement. ■

C.2.4 Proof of Lemma 14

Due to Remark 14, all but Assumptions 3(2) ($\lim_{N \rightarrow \infty} \mathbb{E}_\xi^N[\bar{X}^N(0)] = \mathbb{E}_\xi[X^0] < \infty$), 3(5) ($\lim_{N \rightarrow \infty} \mathbb{E}_\xi^N [\langle 1, \bar{\eta}^N(0) \rangle] = \mathbb{E}_\xi [\langle 1, \eta^0 \rangle] < \infty$) and 5(2) (η^0 has no atoms) in Puha and

Ward (2021) hold. A careful inspection of the proof of Theorem 1 in Puha and Ward (2021), which relies on Theorem 6.2 in Kang et al. (2012) (because the dynamics of η^N are not altered by the non-idling condition), shows that $\sup_{N \in \mathbb{N}} \mathbb{E}_\xi^N [\langle 1, \bar{\eta}^N(0) \rangle] < \infty$ suffices in place of Assumptions 3(2) and 3(5) in Puha and Ward (2021), and that Assumption 5(2) is not needed to establish that the limit η exists and satisfies (C.15). ■

C.2.5 Proof of Lemma 16

To see $\eta_\infty(t) = \gamma\theta^{-1}\eta_e$ for all $t \geq 0$ almost surely, note that η_∞ satisfies (C.15) almost surely. For each bounded continuous function f , the integrand of the first term of the right-hand side of (C.15) tends to zero almost surely as $t \rightarrow \infty$ and so, by the dominated convergence theorem, the integral also tends to zero almost surely as $t \rightarrow \infty$. In addition, $E_\infty(u) = \gamma u$ for all $u \geq 0$ and so the second term of the right-hand side of (C.15) converges to $\gamma \int_0^\infty f(u)(1 - G^r(u))du = \gamma\theta^{-1} \langle f, \eta_e \rangle$ almost surely, as $t \rightarrow \infty$. Thus, $\lim_{t \rightarrow \infty} \langle f, \eta_\infty(t) \rangle = \gamma\theta^{-1} \langle f, \eta_e \rangle$ for each bounded continuous function f . Since η_∞ is a stationary process, the result follows. ■

C.2.6 Proof of Lemma 17

For $t \geq 0$, we have

$$\begin{aligned} \mathbb{E}_\xi[K_\infty(t)] &\stackrel{(1)}{=} \mathbb{E}_\xi[D_\infty(t)] \stackrel{(2)}{=} \mathbb{E}_\xi \left[\int_0^t \langle h^s, \nu_\infty(u) \rangle du \right] \\ &\stackrel{(3)}{=} \mathbb{E}_\xi[\langle h^s, \nu_\infty(0) \rangle]t, \end{aligned} \tag{C.22}$$

where (1) follows from (C.10) and $\mathbb{E}_\xi[B_\infty(t)] = \mathbb{E}_\xi[B_\infty(0)]$ because B_∞ is a stationary process, (2) follows from (C.9), and (3) follows from Fubini's theorem and the stationarity of ν_∞ .

From (C.14), for each $t \geq 0$,

$$\langle 1, \nu_\infty(t) \rangle = \int_0^\infty \frac{\bar{G}^s(x+t)}{\bar{G}^s(x)} \nu_\infty(0)(dx) + \int_0^t \bar{G}^s(t-u) dK_\infty(u).$$

Taking expectation on both sides of the above display and noting that $\langle 1, \nu_\infty(t) \rangle = B_\infty(t)$ from (C.5) yields that for each $t \geq 0$

$$\begin{aligned} \mathbb{E}_\xi[B_\infty(t)] &= \mathbb{E}_\xi \left[\int_0^\infty \frac{\bar{G}^s(x+t)}{\bar{G}^s(x)} \nu_\infty(0)(dx) \right] \\ &\quad + \mathbb{E}_\xi \left[\int_0^t (\bar{G}^s(t-u)) dK_\infty(u) \right]. \end{aligned}$$

As $t \rightarrow \infty$, the first expectation converges to zero by two applications of the dominated convergence theorem and the fact that the interior integrand converges to zero almost surely as $t \rightarrow \infty$. Hence, since the left-hand side of the above is constant in t , combining this with (C.22) gives that, for all $t \geq 0$,

$$\begin{aligned} \mathbb{E}_\xi[B_\infty(t)] &= \mathbb{E}_\xi[\langle h^s, \nu_\infty(0) \rangle] \int_0^\infty (1 - G^s(u)) du \\ &= \mathbb{E}_\xi[\langle h^s, \nu_\infty(0) \rangle] \mu^{-1} = \mathbb{E}_\xi[D_\infty(t)](t\mu)^{-1}. \end{aligned} \tag{C.23}$$

Furthermore, since $X_\infty(t) = X_\infty(0) + E_\infty(t) - R_\infty(t) - D_\infty(t)$, $t \geq 0$, (from (C.13) and $\mathbb{E}_\xi[X_\infty(t)] = \mathbb{E}_\xi[X_\infty(0)]$, $t \geq 0$, due to the stationarity of X_∞ , it follows that $\mathbb{E}_\xi[D_\infty(t)] = \mathbb{E}_\xi[E_\infty(t)] - \mathbb{E}_\xi[R_\infty(t)] \leq \mathbb{E}_\xi[E_\infty(t)] = \gamma t$ for all $t \geq 0$). Then, (C.23) implies that $\mathbb{E}_\xi[B_\infty(t)] \leq \gamma/\mu$, $t \geq 0$. Also, $B_\infty(t) \leq 1$, $t \geq 0$. Hence, there exists $b \in [0, \min\{1, \gamma/\mu\}]$ such that $\mathbb{E}_\xi[B_\infty(t)] = b$ for all $t \geq 0$. This together with (C.23) gives that, for all $t \geq 0$,

$$\mathbb{E}_\xi[D_\infty(t)] = b\mu t,$$

which implies that for all $t \geq 0$,

$$\mathbb{E}_\xi[R_\infty(t)] = \mathbb{E}_\xi[E_p(t)] - \mathbb{E}_\xi[D_\infty(t)] = (\gamma - b\mu)t.$$

From (C.8), Fubini's theorem and stationarity, for all $t \geq 0$,

$$\begin{aligned}\mathbb{E}_\xi[R_\infty(t)] &= \mathbb{E}_\xi \left[\int_0^t \left\langle 1_{[0,\chi_\infty(u)]} h^r, \eta_\infty(u) \right\rangle du \right] \\ &= \mathbb{E}_\xi \left[\left\langle 1_{[0,\chi_\infty(0)]} h^r, \eta_\infty(0) \right\rangle \right] t.\end{aligned}$$

The above two displays imply that $\mathbb{E}_\xi \left[\left\langle 1_{[0,\chi_\infty(t)]} h^r, \eta_\infty(t) \right\rangle \right] = \gamma - b\mu$ for all $t \geq 0$. ■

REFERENCES

- Adlakha, Sachin, Ramesh Johari, Gabriel Y Weintraub. 2015. Equilibria of dynamic games with many players: Existence, approximation, and market structure. *Journal of Economic Theory* **156** 269–316.
- Afèche, Philipp. 2013. Incentive-compatible revenue management in queueing systems: Optimal strategic delay. *Manufacturing & Service Operations Management* **15**(3) 423–443.
- Afeche, Philipp, J Michael Pavlin. 2016. Optimal price/lead-time menus for queues with customer choice: Segmentation, pooling, and strategic delay. *Management Science* **62**(8) 2412–2436.
- Agrawal, Rajeev. 1995. Sample mean based index policies by $O(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability* **27**(4) 1054–1078.
- Agrawal, Shipra, Navin Goyal. 2012. Analysis of Thompson sampling for the multi-armed bandit problem. *Conference on Learning Theory*. JMLR Workshop and Conference Proceedings, 39–1.
- Aksin, Zeynep, Mor Armony, Vijay Mehrotra. 2007. The modern call center: A multidisciplinary perspective on operations management research. *Production and Operations Management* **16**(6) 665–688.
- Akşin, Zeynep, Barış Ata, Seyed Morteza Emadi, Che-Lin Su. 2013. Structural estimation of callers' delay sensitivity in call centers. *Management Science* **59**(12) 2727–2746.
- Akşin, Zeynep, Baris Ata, Seyed Morteza Emadi, Che-Lin Su. 2017. Impact of delay announcements in call centers: An empirical approach. *Operations Research* **65**(1) 242–265.
- Aldous, David J, Hermann Thorisson. 1993. Shift-coupling. *Stochastic Processes and Their Applications* **44**(1) 1–14.
- Allon, Gad, Achal Bassamboo, Itai Gurvich. 2011. “we will be right with you”: Managing customer expectations with vague promises and cheap talk. *Operations research* **59**(6) 1382–1394.
- Allon, Gad, Mirko Kremer. 2018. Behavioral foundations of queueing systems. K. Donohue, E. Katok, S. Leider, eds., *The Handbook of Behavioral Operations*. John Wiley & Sons, 325–366. <https://onlinelibrary.wiley.com/doi/10.1002/9781119138341>.
- Altman, Daniel, Galit B. Yom-Tov, Marcelo Olivares, Shelly Ashtar, Anat Rafaeli. 2021. Do customer emotions affect agent speed? An empirical study of emotional load in online customer contact centers. *Manufacturing & Service Operations Management* **23**(4) 854–875.

- Arapostathis, Ari, Hassan Hmedi, Guodong Pang. 2021. On uniform exponential ergodicity of markovian multiclass many-server queues in the halfin–whitt regime. *Mathematics of Operations Research* **46**(2) 772–796.
- Armony, Mor, Itai Gurvich. 2010. When promotions meet operations: Cross-selling and its effect on call center performance. *Manufacturing & Service Operations Management* **12**(3) 470–488.
- Armony, Mor, Shlomo Israelit, Avishai Mandelbaum, Yariv N Marmor, Yulia Tseytlin, Galit B Yom-Tov. 2015. On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems* **5**(1) 146–194.
- Armony, Mor, Constantinos Maglaras. 2004. Contact centers with a call-back option and real-time delay information. *Operations research* **52**(4) 527–545.
- Armony, Mor, Guillaume Roels, Hummy Song. 2017. Pooling queues with discretionary service capacity. *History* .
- Armony, Mor, Guillaume Roels, Hummy Song. 2021. Pooling queues with strategic servers: The effects of customer ownership. *Operations Research* **69**(1) 13–29.
- Asanjarani, Azam, Yoni Nazarathy, Peter Taylor. 2021. A survey of parameter and state estimation in queues. *Queueing Systems* **97**(1) 39–80.
- Ashutosh, Kumar, Jayakrishnan Nair, Anmol Kagrecha, Krishna Jagannathan. 2021. Bandit algorithms: Letting go of logarithmic regret for statistical robustness. *International Conference on Artificial Intelligence and Statistics*. PMLR, 622–630.
- Asmussen, Søren, Soren Asmussen, Sren Asmussen. 2003. *Applied probability and queues*, vol. 2. Springer.
- Atar, Rami, Amarjit Budhiraja, Paul Dupuis, Ruoyu Wu. 2019. Large deviations for the single server queue and the reneging paradox. *arXiv preprint arXiv:1903.06870* .
- Atar, Rami, Chanit Giat, Nahum Shimkin. 2010. The $c\mu/\theta$ rule for many-server queues with abandonment. *Operations Research* **58**(5) 1427–1439.
- Atar, Rami, Chanit Giat, Nahum Shimkin. 2011a. On the asymptotic optimality of the $c\mu/\theta$ rule under ergodic cost. *Queueing Systems* **67**(2) 127–144.
- Atar, Rami, Weining Kang, Haya Kaspi, Kavita Ramanan. 2021. Large-time limit of non-linearly coupled measure-valued equations that model many-server queues with reneging. *arXiv preprint arXiv:2107.05226* .
- Atar, Rami, Haya Kaspi, Nahum Shimkin. 2014. Fluid limits for many-server systems with reneging under a priority policy. *Mathematics of Operations Research* **39**(3) 672–696.

- Atar, Rami, Avi Mandelbaum, Martin I Reiman, et al. 2004. Scheduling a multi class queue with many exponential servers: Asymptotic optimality in heavy traffic. *The Annals of Applied Probability* **14**(3) 1084–1134.
- Atar, Rami, Yair Y. Shaki, Adam Shwartz. 2011b. A blind policy for equalizing cumulative idleness. *Queueing Systems* **67**(4) 275–293.
- Atar, Rami, et al. 2005. Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *The Annals of Applied Probability* **15**(4) 2606–2650.
- Auer, Peter. 2002. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* **3**(Nov) 397–422.
- Auer, Peter, Nicolo Cesa-Bianchi, Paul Fischer. 2002a. Finite-time analysis of the multiarmed bandit problem. *Machine Learning* **47**(2) 235–256.
- Auer, Peter, Nicolo Cesa-Bianchi, Yoav Freund, Robert E Schapire. 2002b. The nonstochastic multiarmed bandit problem. *SIAM journal on computing* **32**(1) 48–77.
- Avrachenkov, Konstantin, Urtzi Ayesta, Josu Doncel, Peter Jacko. 2013. Congestion control of tcp flows in internet routers by means of index policy. *Computer Networks* **57**(17) 3463–3478.
- Babbar, Sunil, David J Aspelin. 1998. The overtime rebellion: symptom of a bigger problem? *Academy of Management Perspectives* **12**(1) 68–76.
- Baccelli, F, P Boyer, G Hebuterne. 1984. Single-server queues with impatient customers. *Advances in Applied Probability* **16**(4) 887–905.
- Bandiera, O., I. Barankay, I. Rasul. 2010. Social incentives in the workplace. *The Review of Economic Studies* **77**(2) 417–458.
- Baras, JS, D-J Ma, AM Makowski. 1985. K competing queues with geometric service requirements and linear costs: The μ c-rule is always optimal. *Systems & Control Letters* **6**(3) 173–180.
- Bassamboo, Achal, Ramandeep Singh Randhawa. 2016. Scheduling homogeneous impatient customers. *Management Science* **62**(7) 2129–2147.
- Becker, Gary S. 2009. *Human capital: A theoretical and empirical analysis, with special reference to education*. University of Chicago press.
- Bell, Steven L, Ruth J Williams. 2001. Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: Asymptotic optimality of a threshold policy. *The Annals of Applied Probability* **11**(3) 608–649.
- Bennett, George. 1962. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association* **57**(297) 33–45.

- Bertsekas, D. P. 2012. *Dynamic Programming and Optimal Control, Volume II*. Athena Scientific. Third Edition.
- Bertsimas, Dimitris, Georgia Mourtzinou. 1997. Transient laws of non-stationary queueing systems and their applications. *Queueing Systems* **25**(1) 115–155.
- Besbes, Omar, Assaf Zeevi. 2009. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research* **57**(6) 1407–1420.
- Besbes, Omar, Assaf Zeevi. 2011. On the minimax complexity of pricing in a changing environment. *Operations Research* **59**(1) 66–79.
- Besbes, Omar, Assaf Zeevi. 2012. Blind network revenue management. *Operations Research* **60**(6) 1537–1550.
- Bhulai, Sandjai, Herman Blok, FM Spieksma. 2022. K competing queues with customer abandonment: optimality of a generalised $c\mu$ -rule by the smoothed rate truncation method. *Annals of Operations Research* **317**(2) 387–416.
- Bhulai, Sandjai, Ger Koole. 2003. A queueing model for call blending in call centers. *IEEE Transactions on Automatic Control* **48**(8) 1434–1438.
- Borst, Sem, Avi Mandelbaum, Martin I Reiman. 2004. Dimensioning large call centers. *Operations Research* **52**(1) 17–34.
- Boucheron, Stéphane, Gábor Lugosi, Pascal Massart. 2013. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.
- Bramson, Maury. 2008. *Stability of queueing networks*. Springer.
- Braverman, Anton, Itai Gurvich, Junfei Huang. 2020. On the Taylor expansion of value functions. *Operations Research* **68**(2) 631–654.
- Bren, Austin, Soroush Saghatelian. 2019. Data-driven percentile optimization for multiclass queueing systems with model ambiguity: Theory and application. *INFORMS Journal on Optimization* **1**(4) 267–287.
- Broder, Josef, Paat Rusmevichientong. 2012. Dynamic pricing under a general parametric choice model. *Operations Research* **60**(4) 965–980.
- Brodsky, Andrew, Teresa M Amabile. 2018. The downside of downtime: The prevalence and work pacing consequences of idle time at work. *Journal of Applied Psychology* **103**(5) 496.
- Bubeck, Sébastien, Nicolo Cesa-Bianchi. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *In Foundations and Trends in Machine Learning* **5**(1) 1–122.
- Bubeck, Sébastien, Vianney Perchet, Philippe Rigollet. 2013. Bounded regret in stochastic multi-armed bandits. *Conference on Learning Theory*. PMLR, 122–134.

- Bureau of Economic Analysis. 2020. BEA Industry Facts. <https://apps.bea.gov/industry/factsheet/factsheet.cfm>. Accessed May 12, 2020.
- Burke, Ronald JJ, Cary L Cooper. 2008. *Long work hours culture: Causes, consequences and choices*. Emerald Group Publishing.
- Burnetas, Apostolos N, Michael N Katehakis. 1996. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics* **17**(2) 122–142.
- Cachon, G.P., P.T Harker. 2002. Competition and outsourcing with scale economies. *Management Science* **48**(10) 1314–1333.
- Cachon, G.P., F. Zhang. 2007. Obtaining fast service in a queueing system via performance-based allocation of demand. *Management Science* **53**(3) 408–420.
- Chan, Carri W, Galit Yom-Tov, Gabriel Escobar. 2014. When to use speedup: An examination of service systems with returns. *Operations Research* **62**(2) 462–482.
- Chen, Xinyun, Yunan Liu, Guiyu Hong. 2023a. Online learning and optimization for queues with unknown demand curve and service distribution. *arXiv preprint arXiv:2303.03399* .
- Chen, Xinyun, Yunan Liu, Guiyu Hong. 2023b. An online learning approach to dynamic pricing and capacity sizing in service systems. *Operations Research* .
- Choudhury, Tuhinangshu, Gauri Joshi, Weina Wang, Sanjay Shakkottai. 2021. Job dispatching policies for queueing systems with unknown service rates. *arXiv preprint arXiv:2106.04707* .
- Christ, Duane, Benjamin Avi-Itzhak. 2002. Strategic equilibrium for a pair of competing servers with convex cost and balking. *Management Science* **48**(6) 813–820.
- Chung, Hakjin, Hyun-Soo Ahn, Rhonda Righter. 2020. The potentially negative effects of cooperation in service systems. *Advances in Applied Probability* **52**(1) 319–347.
- Chung, Kai Lai. 1967. Markov chains with stationary transition probabilities. *Springer-Verlag, New York* .
- Cohen, Jacob Willem. 2012. *The single server queue*. Elsevier.
- Cohen-Charash, Y., P. E. Spector. 2001. The role of justice in organizations: A meta-analysis. *Organ. Behav. Hum. Dec.* **86**(2) 278–321.
- Colquitt, J. A., D. E. Conlon, M. J. Wesson, C. O. L. H. Porter, K. Y. Ng. 2001. Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *J. Appl. Psychol.* **86**(3) 425–445.
- Combes, Richard, Chong Jiang, Rayadurgam Srikant. 2015. Bandits with budgets: Regret lower bounds and optimal algorithms. *ACM SIGMETRICS Performance Evaluation Review* **43**(1) 245–257.

- Cooper, R. B. 1981. *Introduction to queueing theory, second edition*. North Holland.
- Cox, David R. 1966. Some problems of statistical analysis connected with congestion. W.L. Smith, W.E. Wilkinson, eds., *Proceedings of the Symposium on Congestion Theory*. University of North Carolina Press Chapel Hill, NC, 289–316.
- Cox, David Roxbee, Walter Smith. 1991. *Queues*, vol. 2. CRC Press.
- Da Prato, Giuseppe, Jerzy Zabczyk, J Zabczyk. 1996. *Ergodicity for infinite dimensional systems*, vol. 229. Cambridge University Press.
- Dai, Jim G. 1995. On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *The Annals of Applied Probability* 49–77.
- Dai, Jim G, Mark Gluzman. 2020. Queueing network controls via deep reinforcement learning. *arXiv preprint arXiv:2008.01644* .
- D'Auria, B. 2012. A short note on the monotonicity of the Erlang C formula in the Halfin-Whitt regime. *Queueing Systems* **71**(4) 469–472.
- Daw, A., A. Castellanos, G. Yom-Tov, J. Pender, L. Gruendlinger. 2023. The co-production of service: Modeling service times in contact centers using Hawkes processes. Available at SSRN URL <https://ssrn.com/abstract=3817130>.
- Debo, Laurens G, Christine Parlour, Uday Rajan. 2012. Signaling quality via queues. *Management Science* **58**(5) 876–891.
- Debo, L.G, L.B. Toktay, L.N. Van Wassenhove. 2008. Queuing for expert services. *Management Science* **54**(8) 1497–1512.
- DeHoratius, Nicole, Özgür Gürerk, Dorothée Honhon, Kyle B. Hyndman. 2020. Execution failures in retail supply chains - A virtual reality experiment. Available at SSRN URL <https://ssrn.com/abstract=2676628>.
- Delasay, M., A. Ingolfsson, B. Kolfal. 2016. Modeling load and overwork effects in queueing systems with adaptive service rates. *Operations Research* **64**(4) 867–885.
- Delasay, M., A. Ingolfsson, B. Kolfal, K. Schultz. 2019. Load effect on service times. *European Journal of Operational Research* **279**(3) 673–686.
- Den Boer, Arnoud V. 2015. Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in Operations Research and Management Science* **20**(1) 1–18.
- den Boer, Arnoud V, Bert Zwart. 2015. Dynamic pricing and learning with finite inventories. *Operations Research* **63**(4) 965–978.
- Deuschel, Jean-Dominique, Daniel W Stroock. 2001. *Large Deviations*, vol. 342. American Mathematical Soc.

- Do, Hung T, Masha Shunko, Marilyn T Lucas, David C Novak. 2018. Impact of behavioral factors on performance of multi-server queueing systems. *Production and Operations Management* **27**(8) 1553–1573.
- Dong, J., P. Feldman, G. B. Yom-Tov. 2015. Service systems with slowdowns: Potential failures and proposed solutions. *Operations Research* **63**(2) 305–324.
- Dong, Jing, Rouba Ibrahim. 2020. Managing supply in the on-demand economy: Flexible workers, full-time employees, or both? *Operations Research* **68**(4) 1238–1264.
- Doroudi, S., R. Gopalakrishnan, A. Wierman. 2011. Dispatching to incentivize fast service in multi-server queues. *SIGMETRICS Performance Evaluation Review* **39**(3) 43–45.
- Down, Douglas, Sean P Meyn, Richard L Tweedie. 1995. Exponential and uniform ergodicity of Markov processes. *The Annals of Probability* **23**(4) 1671–1691.
- Down, Douglas G, Ger Koole, Mark E Lewis. 2011. Dynamic control of a single-server system with abandonments. *Queueing Systems* **67**(1) 63–90.
- Duchi, J. 2014. Lecture Notes for Statistics 311/Electrical Engineering 377. <https://web.stanford.edu/class/stats311/lecture-notes.pdf>.
- Economou, Antonis, Spyridoula Kanta. 2008. Optimal balking strategies and pricing for the single server Markovian queue with compartmented waiting space. *Queueing Systems* **59**(3-4) 237.
- Edie, L. C. 1954. Traffic delays at toll booths. *Journal of the Operations Research Society of America* **2**(2) 107–138.
- Erlang, AK. 1948. On the rational determination of the number of circuits. *The life and works of AK Erlang* 216–221.
- Falk, Armin. 2014. Fairness and motivation. *IZA World of Labor* **9**.
- Fogel, Fajwel, Rodolphe Jenatton, Francis Bach, Alexandre d'Aspremont. 2015. Convex relaxations for permutation problems. *SIAM Journal on Matrix Analysis and Applications* **36**(4) 1465–1488.
- Fralix, Brian H, Germán Riaño. 2010. A new look at transient versions of little's law, and $m/g/1$ preemptive last-come-first-served queues. *Journal of Applied Probability* **47**(2) 459–473.
- Freund, Daniel, Thodoris Lykouris, Wentao Weng. 2023. Efficient decentralized multi-agent learning in asymmetric bipartite queueing systems. *Operations Research* .
- Gaitonde, Jason, Éva Tardos. 2020. Stability and learning in strategic queueing systems. *Proceedings of the 21st ACM Conference on Economics and Computation*. 319–347.

- Gans, Noah, Ger Koole, Avishai Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* **5**(2) 79–141.
- Gans, Noah, Yong-Pin Zhou. 2003. A call-routing problem with service-level constraints. *Operations Research* **51**(2) 255–271.
- Geng, X., W. T. Huh, M. Nagarajan. 2015a. Fairness among servers when capacity decisions are endogenous. *Production and Operations Management* **24**(6) 961–974.
- Geng, X., W.T. Huh, M. Nagarajan. 2015b. Fairness among servers when capacity decisions are endogenous. *Production and Operations Management* **24**(6) 961–974.
- Getoor, Ronald K. 1980. Transience and recurrence of Markov processes. *Séminaire de probabilités de Strasbourg* **14** 397–409.
- Gilbert, S.M., Z.K. Weng. 1998. Incentive effects favor nonconsolidating queues in a service system: The principal-agent perspective. *Management Science* **44**(12) 1662–1669.
- Gilboa-Freedman, Gail, Refael Hassin, Yoav Kerner. 2013. The price of anarchy in the markovian single server queue. *IEEE Transactions on Automatic Control* **59**(2) 455–459.
- Gittins, John, Kevin Glazebrook, Richard Weber. 2011. *Multi-armed bandit allocation indices*. John Wiley & Sons.
- Gittins, John C. 1979. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)* **41**(2) 148–164.
- Gopalakrishnan, R., S. Doroudi, A. R. Ward, A. Wierman. 2014. Routing and staffing when servers are strategic. *Proceedings of the Fifteenth ACM Conference on Economics and Computation (EC)*. 713–714.
- Gopalakrishnan, R., S. Doroudi, A. R. Ward, A. Wierman. 2016a. Routing and staffing when servers are strategic. *Operations Research* **64**(4) 1033–1050.
- Gopalakrishnan, Ragavendran, Sherwin Doroudi, Amy R Ward, Adam Wierman. 2016b. Routing and staffing when servers are strategic. *Operations Research* **64**(4) 1033–1050.
- Gopalakrishnan, Ragavendran, Yueyang Zhong. 2023. Some asymptotic properties of the erlang-c formula in many-server limiting regimes. *arXiv preprint arXiv:2304.13845* .
- Gumbel, Harold. 1960. Waiting lines with heterogeneous servers. *Operations Research* **8**(4) 504–511.
- Guo, Pengfei, Paul Zipkin. 2007. Analysis and comparison of queues with different levels of delay information. *Management Science* **53**(6) 962–970.
- Gurvich, Itay, Mor Armony, Avishai Mandelbaum. 2008. Service-level differentiation in call centers with fully flexible servers. *Management Science* **54**(2) 279–294.

- Haji, B., S. M. Ross. 2015. A queueing loss model with heterogenous skill based servers under idle time ordering policies. *Journal of Applied Probability* **52**(1) 269–277.
- Harchol-Balter, Mor. 2013. *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge University Press.
- Harel, Arie. 1988. Sharp bounds and simple approximations for the erlang delay and loss formulas. *Management Science* **34**(8) 959–972.
- Harel, Arie. 2010. Sharp and simple bounds for the Erlang delay and loss formulae. *Queueing Systems* **64**(2) 119–143.
- Harrison, J Michael, N Bora Keskin, Assaf Zeevi. 2012. Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. *Management Science* **58**(3) 570–586.
- Hassin, R. 2016. *Rational Queueing*. CRC Press, Boca Raton, FL.
- Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, Boston, MA.
- Hassin, Rafael, Ran I Snitkovsky. 2020. Social and monopoly optimization in observable queues. *Operations Research* **68**(4) 1178–1198.
- Haviv, Moshe. 2014. Regulating an M/G/1 queue when customers know their demand. *Performance Evaluation* **77** 57–71.
- Haviv, Moshe, Binyamin Oz. 2016. Regulating an ovservable M/M/1 queue. *Operations Research Letters* **44**(2) 196–198.
- Hopp, W. J., S. M. R. Iravani, G. Y. Yuen. 2007. Operations systems with discretionary task completion. *Management Science* **53**(1) 61–77.
- Hopp, Wallace J, William S Lovejoy. 2012. *Hospital Operations: Principles of High Efficiency Health Care*. FT Press.
- Hsee, Christopher K, Adelle X Yang, Liangyan Wang. 2010. Idleness aversion and the need for justifiable busyness. *Psychological Science* **21**(7) 926–930.
- Hu, Ming, Yang Li, Jianfu Wang. 2018. Efficient ignorance: Information heterogeneity in a queue. *Management Science* **64**(6) 2650–2671.
- Huang, Minyi, Roland P Malhamé, Peter E Caines, et al. 2006. Large population stochastic dynamic games: closed-loop mckean-vlasov systems and the nash certainty equivalence principle. *Communications in Information & Systems* **6**(3) 221–252.
- Huh, Woonghee Tim, Ganesh Janakiraman, John A Muckstadt, Paat Rusmevichientong. 2009. An adaptive algorithm for finding the optimal base-stock policy in lost sales inventory systems with censored demand. *Mathematics of Operations Research* **34**(2) 397–416.

- Huh, Woonghee Tim, Paat Rusmevichientong. 2014. Online sequential optimization with biased gradients: theory and applications to censored demand. *INFORMS Journal on Computing* **26**(1) 150–159.
- Ibrahim, R. 2018. Managing queueing systems where capacity is random and customers are impatient. *Production and Operations Management* **27**(2) 234–250.
- Ittig, Peter T. 1994. Planning service capacity when demand is sensitive to delay. *Decision Sciences* **25**(4) 541–553.
- Jagerman, D. L. 1974. Some properties of the Erlang loss function. *The Bell System Technical Journal* **53**(3) 525–551.
- Janssen, A. J. E. M., J. S.H. Van Leeuwaarden, B. Zwart. 2011. Refining square-root safety staffing by expanding Erlang C. *Operations Research* **59**(6) 1512–1522.
- Jansson, Birger. 1966. Choosing a good appointment system—a study of queues of the type (D, M, 1). *Operations Research* **14**(2) 292–312.
- Jia, Huiwen, Cong Shi, Siqian Shen. 2022. Online learning and pricing for service systems with reusable resources. *Operations Research* .
- Jones, MC. 2008. On a class of distributions with simple exponential tails. *Statistica Sinica* 1101–1110.
- Jouini, Oualid, Yves Dallery, Rabie Nait-Abdallah. 2008. Analysis of the impact of team-based organizations in call center management. *Management Science* **54**(2) 400–414.
- Kalai, E., M. I. Kamien, M. Rubinovitch. 1992. Optimal service speeds in a competitive environment. *Management Science* **38**(8) 1154–1163.
- Kalvit, Anand, Assaf Zeevi. 2022. Dynamic learning in large matching markets. *ACM SIGMETRICS Performance Evaluation Review* **50**(2) 18–20.
- Kamakura, Wagner, Carl F Mela, Asim Ansari, Anand Bodapati, Pete Fader, Raghuram Iyengar, Prasad Naik, Scott Neslin, Baohong Sun, Peter C Verhoef, et al. 2005. Choice models and customer relationship management. *Marketing letters* **16**(3-4) 279–291.
- Kang, Weining, Guodong Pang. 2019. Equivalence of fluid models for Gt/GI/N+ GI queues. *Modeling, Stochastic Control, Optimization, and Applications*. Springer, 315–349.
- Kang, Weining, Kavita Ramanan. 2010. Fluid limits of many-server queues with reneging. *The Annals of Applied Probability* **20**(6) 2204–2260.
- Kang, Weining, Kavita Ramanan, et al. 2012. Asymptotic approximations for stationary distributions of many-server queues with abandonment. *The Annals of Applied Probability* **22**(2) 477–521.

- Karau, S. J., K. D. Williams. 1993. Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology* **65**(4) 681–706.
- Kaspi, Haya, Kavita Ramanan. 2011. Law of large numbers limits for many-server queues. *The Annals of Applied Probability* **21**(1) 33–114.
- Kc, D. S., C. Terwiesch. 2009. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* **55**(9) 1486–1498.
- Kella, Offer, Sundareswaran Ramasubramanian. 2012. Asymptotic irrelevance of initial conditions for Skorohod reflection mapping on the nonnegative orthant. *Mathematics of Operations Research* **37**(2) 301–312.
- Kelly, Frank P. 1991. Loss networks. *Annals of Applied Probability* 319–378.
- Keskin, N Bora, Assaf Zeevi. 2014. Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations Research* **62**(5) 1142–1167.
- Kim, Seong-Hee, Barry L Nelson. 2007. Recent advances in ranking and selection. *2007 Winter Simulation Conference*. IEEE, 162–172.
- Kim, Song-Hee, Ward Whitt. 2013. Estimating waiting times with the time-varying little's law. *Probability in the Engineering and Informational Sciences* **27**(4) 471.
- Kleinberg, Robert, Tom Leighton. 2003. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings..* IEEE, 594–605.
- Knudsen, N. C. 1972. Individual and social optimization in a multiserver queue with a general cost-benefit structure. *Econometrica* **40**(3) 515–528.
- Krakowski, Martin. 1973. Conservation methods in queuing theory. *Revue française d'automatique, informatique, recherche opérationnelle. Recherche opérationnelle* **7**(V1) 63–83.
- Krishnasamy, Subhashini, Ari Arapostathis, Ramesh Johari, Sanjay Shakkottai. 2018. On learning the $c\mu$ rule in single and parallel server networks. *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 153–154.
- Krishnasamy, Subhashini, Rajat Sen, Ramesh Johari, Sanjay Shakkottai. 2016. Regret of queueing bandits. *Advances in Neural Information Processing Systems* **29** 1669–1677.
- Krishnasamy, Subhashini, Rajat Sen, Ramesh Johari, Sanjay Shakkottai. 2021. Learning unknown service rates in queues: A multiarmed bandit approach. *Operations Research* **69**(1) 315–330.
- Kulkarni, Vidyadhar G. 2016. *Modeling and analysis of stochastic systems*. Crc Press.

- Kunnumkal, Sumit, Huseyin Topaloglu. 2008. Using stochastic approximation methods to compute optimal base-stock levels in inventory control problems. *Operations Research* **56**(3) 646–664.
- Laguerre, E, Stewart A Levin. 1883. On the theory of numeric equations. *Journal de Mathematiques pures et appliquees* Translation available from <http://sepwww.stanford.edu/oldsep/stew/laguerre.pdf>.
- Lai, Tze Leung, Herbert Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* **6**(1) 4–22.
- Lasry, Jean-Michel, Pierre-Louis Lions. 2007. Mean field games. *Japanese journal of mathematics* **2**(1) 229–260.
- Latane, B., K. Williams, S. Harkins. 1979. Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology* **37**(6) 822–832.
- Lee, Dabeen, Milan Vojnovic. 2021. Learning to schedule. *arXiv preprint arXiv:2105.13655*
- Lee, Nam H. 2008. *A sufficient condition for stochastic stability of an Internet congestion control model in terms of fluid model stability*. University of California, San Diego.
- Levi, Retsef, Thomas Magnanti, Yaron Shaposhnik. 2019. Scheduling with testing. *Management Science* **65**(2) 776–793.
- Levin, David A, Yuval Peres. 2017. *Markov chains and mixing times*, vol. 107. American Mathematical Soc.
- Lindvall, Torgny. 2002. *Lectures on the coupling method*. Courier Corporation.
- Liu, Bai, Qiaomin Xie, Eytan Modiano. 2019a. Reinforcement learning for optimal control of queueing systems. *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 663–670.
- Liu, Ran, Michael E Kuhl, Yunan Liu, James R Wilson. 2019b. Modeling and simulation of nonstationary non-poisson arrival processes. *INFORMS Journal on Computing* **31**(2) 347–366.
- Liu, Tie-Yan. 2011. *Learning to Rank for Information Retrieval*. Springer Verlag, Berlin Heidelberg.
- Long, Zhenghua, Nahum Shimkin, Hailun Zhang, Jiheng Zhang. 2020. Dynamic scheduling of multiclass many-server queues with abandonment: The generalized $c\mu/h$ rule. *Operations Research* **68**(4) 1218–1230.

- Lozano, Macarena, Pilar Moreno. 2008. A discrete time single-server queue with balking: Economic applications. *Applied Economics* **40**(6) 735–748.
- Lu, Yina. 2013. Data-driven system design in service operations. Ph.D. thesis, Columbia University.
- Maglaras, Costis, John Yao, Assaf Zeevi. 2018. Optimal price and delay differentiation in large-scale queueing systems. *Management Science* **64**(5) 2427–2444.
- Maglio, Paul P, Cheryl A Kieliszewski, James C Spohrer, Kelly Lyons, Lia Patrício, Yuriko Sawatani. 2019. *Handbook of Service Science, Volume II*. Springer.
- Mahajan, Aditya, Demosthenis Teneketzis. 2008. Multi-armed bandit problems. *Foundations and Applications of Sensor Management*. Springer, 121–151.
- Mandelbaum, Avi, William A Massey, Martin I Reiman, Alexander Stolyar, Brian Rider. 2002. Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommunication Systems* **21**(2) 149–171.
- Mandelbaum, Avishai, Petar Momčilović, Yulia Tseytlin. 2012. On fair routing from emergency departments to hospital wards: Qed queues with heterogeneous servers. *Management Science* **58**(7) 1273–1291.
- Mandelbaum, Avishai, Nahum Shimkin. 2000. A model for rational abandonments from invisible queues. *Queueing Systems* **36**(1-3) 141–173.
- Mandelbaum, Avishai, Alexander L Stolyar. 2004. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$ -rule. *Operations Research* **52**(6) 836–855.
- Maoui, Idriss, Hayriye Ayhan, Robert D Foley. 2009. Optimal static pricing for a service facility with holding costs. *European journal of operational research* **197**(3) 912–923.
- Marshall, Knaeale T. 1973. Linear bounds on the renewal function. *SIAM Journal on Applied Mathematics* **24**(2) 245–250.
- Mas, A., E. Moretti. 2009. Peers at work. *American Economic Review* **99**(1) 112–145.
- Meyn, Sean P, Douglas Down. 1994. Stability of generalized Jackson networks. *The Annals of Applied Probability* 124–148.
- Meyn, Sean P., Richard L. Tweedie. 1993a. Generalized resolvents and Harris recurrence of Markov processes. Harry Cohn, ed., *Contemporary Mathematics*, vol. 149. American Mathematical Society, 227–250.
- Meyn, Sean P, Richard L Tweedie. 1993b. Stability of Markovian processes II: Continuous-time processes and sampled chains. *Advances in Applied Probability* **25**(3) 487–517.

- Meyn, Sean P, Richard L Tweedie. 1993c. Stability of Markovian processes III: Foster–Lyapunov criteria for continuous-time processes. *Advances in Applied Probability* **25**(3) 518–548.
- Meyn, Sean P, Richard L Tweedie. 2012. *Markov Chains and Stochastic Stability*. Springer Verlag, London.
- Meyn, Sean P, RL Tweedie. 1994a. State-dependent criteria for convergence of Markov chains. *The Annals of Applied Probability* **4**(1) 149–168.
- Meyn, Sean P, Robert L Tweedie. 1994b. Computable bounds for geometric convergence rates of Markov chains. *The Annals of Applied Probability* 981–1011.
- Mokaddis, G. S., C. H. Matta, M. M. El Genaidy. 1998. On Poisson queue with three heterogeneous servers. *International Journal of Information and Management Sciences* **9**(4) 53–60.
- Naor, P. 1969. The regulation of queue size by levying tolls. *Econometrica* **37**(1) 15–24.
- Nelson, Barry L, Ira Gerhardt. 2011. Modelling and simulating non-stationary arrival processes to facilitate analysis. *Journal of Simulation* **5**(1) 3–8.
- Netessine, Serguei, Robert A Shumsky. 2005. Revenue management games: Horizontal and vertical competition. *Management Science* **51**(5) 813–831.
- Niño-Mora, José. 2012. Admission and routing of soft real-time jobs to multiclusters: Design and comparison of index policies. *Computers & Operations Research* **39**(12) 3431–3444.
- Parlaktürk, Ali K, Sunil Kumar. 2004. Self-interested routing in queueing networks. *Management Science* **50**(7) 949–966.
- Perchet, Vianney, Philippe Rigollet, Sylvain Chassang, Erik Snowberg. 2016. Batched bandit problems. *The Annals of Statistics* 660–681.
- Pinedo, Michael L. 2012. *Scheduling*. Springer, Boston.
- Pinelis, Iosif. 2020. Exact lower and upper bounds on the incomplete gamma function. *arXiv preprint arXiv:2005.06384* .
- Pooley, John M, Edwin A Bump. 1993. The learning performance and cost effectiveness of mentally disabled workers. *Group & Organization Management* **18**(1) 88–102.
- Prokhorov, Yu V. 1956. Convergence of random processes and limit theorems in probability theory. *Theory of Probability & Its Applications* **1**(2) 157–214.
- Puha, Amber L, Amy R Ward. 2019. Scheduling an overloaded multiclass many-server queue with impatient customers. *INFORMS TutORials in Operations Research: Operations Research & Management Science in the Age of Analytics* 189–217.

- Puha, Amber L, Amy R Ward. 2021. Fluid limits for multiclass many-server queues with general reneging distributions and head-of-the-line scheduling. *Mathematics of Operations Research* .
- Puha, Amber L, Amy R Ward. 2022. Fluid limits for multiclass many-server queues with general reneging distributions and head-of-the-line scheduling. *Mathematics of Operations Research* **47**(2) 1192–1228.
- Puterman, Martin L. 2014a. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.
- Puterman, Martin L. 2014b. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.
- Roberts, Gareth O, Jeffrey S Rosenthal. 1997. Shift-coupling and convergence rates of ergodic averages. *Stochastic Models* **13**(1) 147–165.
- Roberts, Gareth O, Jeffrey S Rosenthal. 2004. General state space Markov chains and mcmc algorithms .
- Rosenthal, Jeffrey S. 1995. Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association* **90**(430) 558–566.
- Rosokha, Yaroslav, Chen Wei. 2020. Cooperation in queueing systems. Available at SSRN URL <https://ssrn.com/abstract=3526505>.
- Rothkopf, M. H., P. Rech. 1987. Perspectives on queues: Combining queues is not always beneficial. *Operations Research* **35**(6) 906–909.
- SAKASEGAWA, HIROTAKA. 1977. An approximation formula $l_1 = ap^0/(1-p)$. *Ann. Inst. Statist. Math., Part A* **29** 67–75.
- Salomon, Antoine, Jean-Yves Audibert, Issam El Alaoui. 2013. Lower bounds and selectivity of weak-consistent policies in stochastic multi-armed bandit problem. *Journal of Machine Learning Research* **14**(1) 187–207.
- Satin, Yacov, Alexander Zeifman, Alexander Sipin, Sherif I Ammar, Janos Sztrik. 2020. On probability characteristics for a class of queueing models with impatient customers. *Mathematics* **8**(4) 594.
- Sauré, Denis, Assaf Zeevi. 2013. Optimal dynamic assortment planning with demand learning. *Manufacturing & Service Operations Management* **15**(3) 387–404.
- Schendel, Joel D, Joseph D Hagman. 1982. On sustaining procedural skills over a prolonged retention interval. *Journal of Applied Psychology* **67**(5) 605.
- Sentenac, Flore, Etienne Boursier, Vianney Perchet. 2021. Decentralized learning in online queuing systems. *Advances in Neural Information Processing Systems* **34** 18501–18512.

- Sevast'yanov, Boris Aleksandrovich. 1957. An ergodic theorem for markov processes and its application to telephone systems with refusals. *Theory of Probability & Its Applications* **2**(1) 104–112.
- Shah, Devavrat, Qiaomin Xie, Zhi Xu. 2020. Stable reinforcement learning with unbounded state space. *arXiv preprint arXiv:2006.04353* .
- Shaked, Moshe, J George Shanthikumar. 2007. *Stochastic orders*. Springer.
- Shalev-Shwartz, Shai, et al. 2011. Online learning and online convex optimization. *Foundations and Trends in Machine Learning* **4**(2) 107–194.
- Shen, Yiwen, Carri W Chan, Fanyin Zheng, Michael Argenziano, Paul Kurlansky. 2021. The impact of surgeon daily workload and its implications for operating room scheduling. URL http://www.columbia.edu/~cc3179/cardiac_workload_2021.pdf. Working paper.
- Shi, Pengyi, Mabel C Chou, Jim G Dai, Ding Ding, Joe Sim. 2016. Models and insights for hospital inpatient operations: Time-dependent ed boarding time. *Management Science* **62**(1) 1–28.
- Shortle, J. F., J. M. Thompson, D. Gross, C. M. Harris. 2018. *Fundamentals of Queueing Theory*. Wiley. Fifth Edition.
- Shunko, M., J. Niederhoff, Y. Rosokha. 2018. Humans are not machines: The behavioral impact of queueing design on service time. *Management Science* **64**(1) 453–473.
- Shwartz, Adam, Alan Weiss. 1995. *Large Deviations for Performance Analysis: Queues, Communication and Computing*, vol. 5. CRC Press.
- Simhon, Eran, Yezekael Hayel, David Starobinski, Quanyan Zhu. 2016. Optimal information disclosure policies in strategic queueing games. *Operations Research Letters* **44**(1) 109–113.
- Slivkins, Aleksandrs. 2019. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning* **12**(1-2) 1–286.
- Smith, Wayne E. 1956. Various optimizers for single-stage production. *Naval Research Logistics Quarterly* **3**(1-2) 59–66.
- Song, H., A. L. Tucker, K. L. Murrell. 2015. The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science* **61**(12) 3032–3053.
- Staats, Bradley R, Francesca Gino. 2012. Specialization and variety in repetitive tasks: Evidence from a Japanese bank. *Management Science* **58**(6) 1141–1159.
- Stahlbuhk, Thomas, Brooke Shrader, Eytan Modiano. 2018. Learning algorithms for minimizing queue length regret. *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 1001–1005.

- Stahlbuhk, Thomas, Brooke Shrader, Eytan Modiano. 2021. Learning algorithms for minimizing queue length regret. *IEEE Transactions on Information Theory* **67**(3) 1759–1781.
- Stidham, S. 2009. *Optimal design for queueing systems*. CRC Press.
- Stidham, S jr, NU Prabhu. 1974. Optimal control of queueing systems. *Mathematical methods in queueing theory*. Springer, 263–294.
- Stidham, Shaler. 1985a. Optimal control of admission to a queueing system. *IEEE Transactions on Automatic Control* **30**(8) 705–713.
- Stidham, Shaler. 1985b. Optimal control of admission to a queueing system. *IEEE Transactions on Automatic Control* **30**(8) 705–713.
- Stone, A. 2012. Why waiting is torture. <https://www.nytimes.com/2012/08/19/opinion/sunday/why-waiting-in-line-is-torture.html>.
- Sun, Xu, Jingtong Zhao. 2022. Congestion-aware matching and learning for service platforms .
- Sun, Zhankun, Nilay Tanik Argon, Serhan Ziya. 2018. Patient triage and prioritization under austere conditions. *Management Science* **64**(10) 4471–4489.
- Tan, Tom Fangyun, Serguei Netessine. 2014. When does the devil make work? An empirical study of the impact of workload on worker productivity. *Management Science* **60**(6) 1574–1593.
- Tan, Tom Fangyun, Serguei Netessine. 2019. When you work with a superman, will you also fly? An empirical study of the impact of coworkers on performance. *Management Science* **65**(8) 3495–3517.
- Tassiulas, Leandros, Anthony Ephremides. 1990. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *29th IEEE Conference on Decision and Control*. IEEE, 2130–2132.
- Tassiulas, Leandros, Anthony Ephremides. 1993. Dynamic server allocation to parallel queues with randomly varying connectivity. *IEEE Transactions on Information Theory* **39**(2) 466–478.
- Thorisson, Hermann. 1993. From coupling to shift-coupling. *Theory of Probability & Its Applications* **37**(1) 105–112.
- Thorisson, Hermann. 1994. Shift-coupling in continuous time. *Probability Theory and Related Fields* **99** 477–483.
- Thorisson, Hermann. 1995. Coupling methods in probability theory. *Scandinavian journal of statistics* 159–182.

- Thorisson, Hermann. 2000. Coupling, stationarity, and regeneration. *Probability and its Applications* .
- Tversky, Amos, Daniel Kahneman. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* **5**(4) 297–323.
- Unlu, Gorkem, Yuan Zhong. 2023. Instability and stability of parameter agnostic policies in parallel server systems. *Working paper* .
- Van Mieghem, Jan A. 1995. Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *The Annals of Applied Probability* 809–833.
- Van Mieghem, Jan A. 2003. Due-date scheduling: Asymptotic optimality of generalized longest queue and generalized largest delay rules. *Operations Research* **51**(1) 113–122.
- Varadhan, SR Srinivasa. 1984. *Large Deviations and Applications*. SIAM.
- Vermorel, Joannes, Mehryar Mohri. 2005. Multi-armed bandit algorithms and empirical evaluation. *European Conference on Machine Learning*. Springer, 437–448.
- Walton, Neil, Kuang Xu. 2021. Learning and information in stochastic networks and queues. *INFORMS TutORials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications* 161–198.
- Walton, Neil S. 2014. Two queues with non-stochastic arrivals. *Operations Research Letters* **42**(1) 53–57.
- Wang, J., Y. Zhou. 2018. Impact of queue configuration on service time: Evidence from a supermarket. *Management Science* **64**(7) 3055–3075.
- Wang, Zhongbin, Luyi Yang, Shiliang Cui, Sezer Ulku, Yong-Pin Zhou. 2022. Pooling agents for customer-intensive services. *Operations Research* doi:10.1287/opre.2022.2259.
- Ward, Amy R. 2019. Open problem—regarding static priority scheduling for many-server queues with reneging. *Stochastic Systems* **9**(3) 313–314.
- Ward, Amy R, Mor Armony. 2013. Blind fair routing in large-scale service systems with heterogeneous customers and servers. *Operations Research* **61**(1) 228–243.
- Ward, Whitt. 2006. The impact of increased employee retention upon performance in a customer contact center. *Manufacturing & Service Operations Management* .
- Weber, Richard R. 1978. On the optimal assignment of customers to parallel servers. *Journal of Applied Probability* **15**(2) 406–413.
- Weintraub, Gabriel Y, C Lanier Benkard, Benjamin Van Roy. 2008. Markov perfect industry dynamics with many firms. *Econometrica* **76**(6) 1375–1411.
- Weiss, Gideon. 1995. Scheduling: Theory, algorithms, and systems.

- Whitt, W. 2002. IEOR 6707: Advanced topics in queueing theory: Focus on customer contact centers. Homework 1e Solutions, see <http://www.columbia.edu/~ww2040/ErlangBAndCFormulas.pdf>.
- Whitt, Ward. 2006. Fluid models for multiserver queues with abandonments. *Operations Research* **54**(1) 37–54.
- Wilson, Timothy D, David A Reinhard, Erin C Westgate, Daniel T Gilbert, Nicole Ellerbeck, Cheryl Hahn, Casey L Brown, Adi Shaked. 2014. Just think: The challenges of the disengaged mind. *Science* **345**(6192) 75–77.
- Wright, T. P. 1936. Factors affecting the costs of airplanes. *Journal of the Aeronautical Sciences* **3** 122–128.
- Xie, Minge, Kesar Singh, Cun-Hui Zhang. 2009. Confidence intervals for population ranks in the presence of ties and near ties. *Journal of the American Statistical Association* **104**(486) 775–788.
- Yang, Adelle X, Christopher K Hsee. 2019. Idleness versus busyness. *Current opinion in psychology* **26** 15–18.
- Zeifman, AI. 1995. Upper and lower bounds on the rate of convergence for nonhomogeneous birth and death processes. *Stochastic Processes and Their Applications* **59**(1) 157–173.
- Zhan, D., A. R. Ward. 2019. Staffing, routing, and payment to trade off speed and quality in large service systems. *Operations Research* **67** 1738–1751.
- Zhan, Dongyuan, Amy R Ward. 2018. The M/M/1+M queue with a utility-maximizing server. *Operations Research Letters* **46**(5) 518–522.
- Zhang, Daowen. 2005. Lecture Notes for ST745 Analysis of Survival Data, Chapter 3 Likelihood and Censored (or Truncated) Survival Data. <https://www4.stat.ncsu.edu/~dzhang2/st745/chap3.pdf>.
- Zhang, Jiheng. 2013. Fluid models of many-server queues with abandonment. *Queueing Systems* **73**(2) 147–193.
- Zhong, Yueyang, John R Birge, Amy Ward. 2022a. Learning the scheduling policy in time-varying multiclass many server queues with abandonment. *Available at SSRN* .
- Zhong, Yueyang, Raga Gopalakrishnan, Amy Ward. 2023. Behavior-aware queueing: The finite-buffer setting with many strategic servers. *Operations Research* .
- Zhong, Yueyang, Raga Gopalakrishnan, Amy R Ward. 2022b. Some properties of the Erlang B and C formulae. *Working paper (Available at https://yzhong0.github.io/yueyangzhong/files/technical_file.pdf)* .
- Zhong, Yueyang, Amy R Ward, Amber L Puha. 2022c. Asymptotically optimal idling in the GI/GI/N+ GI queue. *Operations Research Letters* **50**(3) 362–369.