

Structural Methods for Large Stochastic Systems

Yige Hong

CMU-CS-26-119

August 2026

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:

Weina Wang (Chair)

Mor Harchol-Balter

Alan Scheller-Wolf

Jim Dai (Cornell University)

Yudong Chen (University of Wisconsin-Madison)

Qiaomin Xie (University of Wisconsin-Madison)

*Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science.*

Copyright © 2026 **Yige Hong**

This research was sponsored by the National Science Foundation under award numbers CNS-2007733, ECCS-2145713, CCF-2403194, ECCS-2432545, and CCF-2428569. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

Keywords: queueing theory, Markov decision processes, multiserver queues, bin packing, restless bandits, weakly-coupled MDPs, asymptotic optimality, asymptotic independence, Lyapunov analysis, basic adjoint relationship, Palm calculus

Abstract

Decision making in large stochastic systems is a central problem across computer systems, machine learning, and operations research. Such systems typically consist of a large number of interacting entities, which makes efficient and performant decision making especially challenging.

In this thesis, we study a wide variety of large stochastic systems. Despite their diverse problem definitions, these systems share similar underlying structures and admit similar fundamental techniques for performance analysis and policy design. We identify two such structures, each grouping a family of seemingly distinct problems. The two structures share a common spirit: an intractable system can be approximated by a simpler, well-understood one.

The first part studies the *weak-coupling structure*, where a system can be approximated by a collection of independent, low-dimensional subsystems. We design a control policy on this simple proxy and convert it into a near-optimal policy for the original, coupled system. Applied to stochastic bin packing, restless bandits, and weakly-coupled Markov decision processes (WCMDPs), this yields efficiently computable policies that are provably near-optimal at scale, under substantially weaker and more easily verifiable conditions than were previously required.

The second part studies the *one-dimensional structure*, where a key quantity of the system can be approximated by a one-dimensional process even when the full state is high- or infinite-dimensional. Applying this idea to multiserver queues, we prove new universal bounds for the $G/G/n$ queue, show that the Gittins policy is near-optimal for $G/G/n$ queues with setup times, and determine the best achievable delay together with a near-optimal policy for the multiserver-job model.

Acknowledgments

Before thanking everyone, let me briefly explain my learning experience and why this PhD has been the most transformative period of my life.

My inner journey has always been intertwined with my pursuit of mathematics. I have been fascinated by mathematics from an early age, drawn to its simplicity and capacity for surprise. My early study of mathematics, however, was not smooth. I spent two years training for mathematics competitions in high school but had little success. I sometimes struggled to organize my thoughts quickly. More fundamentally, my mind often wandered through an internal world, with multiple thoughts running in the background (mathematically, a G/G/n queue with large service-time variance; see Chapter 7), making it difficult for me to respond promptly to a given task.

I nevertheless chose mathematics as my major when I entered college. Although I was more fortunate in my studies than I had been in high school, the tension remained: on the one hand, I was almost certain that I wanted to spend the rest of my life on mathematics and theoretical research; on the other hand, mathematics was hard, and I did not know how I could contribute to it meaningfully. This tension filled me with self-criticism and questions about meaning, and I hoped that doing good research might help me find a stronger sense of purpose.

This led me to pursue a PhD, a long process that has turned out to be deeply rewarding and transformative. I have generally been fortunate in my research during the PhD: there were difficult times when I felt lost and frustrated, but I gradually found my own way of doing mathematics and greater confidence in what I could contribute. I was also able to move slightly beyond my original fixation on research, finding opportunities to spend time with peers, enjoy giving talks at conferences, and socialize with people I had previously known only through their papers. Moreover, by overcoming these challenges, I finally feel more in control of my life and am beginning to take my inner critic less seriously.

I could not have found my way out of the wilderness by myself. I am grateful for the generous support of so many people around me.

First of all, I would like to thank my advisor, Weina Wang. I have been very fortunate to have her as my advisor. She has guided me with excellent research taste, hands-on mentoring, integrity, and patience. I will miss the times when we discussed state-space concentration in her office from afternoon into evening, when she helped me rethink the outline of my job-talk slides in a coffee shop, and when she stayed on Zoom late into the night as I refined the algorithm section of my SIGMETRICS paper. I will also remember how encouraged I felt when she generously highlighted my work during her Rising Star Award talk and people came to discuss it with me afterward. She also introduced me to David Alan Goldberg after I improved his bounds for the $G/G/n$ queue; that work later became the basis of my first single-author paper. Those moments have motivated me over the years.

Next, I would like to thank other CMU professors who have provided generous guidance and support during my time here. Mor Harchol-Balter is on my thesis committee, and she has effectively served as my second advisor since I joined CMU. She has taught me a great deal about research, presentation, writing, and social skills. Her guidance on presentations was especially important: after each of my rough practice talks, she would work through my slides with me page by page, showing me how to turn them into a polished talk I could present with confidence. She has also taught me how to reach out to other researchers, which gave me the confidence to invite many people whose work I admire to my talk at the APS 2025 conference. Alan Scheller-Wolf is also on my thesis committee and has consistently offered thoughtful feedback on my research and presentations. We collaborated on a paper, and his comments made the work more rigorous. When I was hesitant about my job talk, he reminded me that the work was strong and worth presenting with confidence. Even on short notice, he made time for a final practice session. I am grateful for his encouragement and support. I would also like to thank Gauri Joshi, Shixiang (Woody) Zhu, Evelyn Xiaoyue Gong, and the other professors who attended the weekly SQUALL seminar (Scheduling and Queueing At LLunch), where I received detailed and constructive feedback on my presentations. Gauri also served on my writing skills committee, while Woody helped me practice for job interviews.

I would like to thank my long-term collaborators at the University of Wisconsin–Madison, Qiaomin Xie and Yudong Chen, who are also members of my thesis committee. My collaboration with Qiaomin began with the work in Chapter 2, and Yudong joined us later. Qiaomin has always been hands-on: she once worked late with me to meet a SIGMETRICS deadline, and she has

repeatedly helped me revise papers carefully under tight deadlines. Yudong has brought the same care to our later work. Their input was crucial in finding the connection between bin packing and restless bandits that led to my work in Chapter 3. They remained committed through difficult rounds of review and revision, always ready to meet, make a new plan, and improve the work. Alongside these serious efforts, I treasure our lighter discussions about naming algorithms and papers; Yudong is especially good at finding a catchy title, often through a well-placed colon. Our shared breakthroughs and setbacks eventually shaped the entire line of work in Part I, and I am deeply grateful for their patience, insights, and humor.

I would like to thank Jim Dai. Jim is my undergraduate mentor and the person who first drew me into theoretical research. Since I began my PhD, he has continued to follow my development. He visited us at CMU and hosted me during several trips to Cornell University. What I learned while working with Jim has strongly influenced the research in Part II. He joined my thesis committee, strongly advocated for me on the job market, and encouraged me throughout the process. It has always been reassuring to know that Jim has my back, and my visits to Cornell’s beautiful campus are among my best memories of the PhD.

I would also like to thank David Alan Goldberg at Cornell University. David has consistently advocated for me, followed all of my papers with interest, and given me a great deal of encouragement, which has been a strong source of motivation throughout my PhD. Whenever we met at a conference, he urged me to introduce myself to other researchers and discuss my work with them. Although I was initially reluctant, his persistence gave me the confidence to reach out, helping me enjoy conferences and engage more fully with the research community.

I would also like to thank Lei Ying, my academic grandpa. I learned a great deal from Lei’s work and enjoyed interacting with Lei and his group at conferences. Lei also supported my postdoc applications and shared valuable information that helped me prepare for job interviews.

I would like to thank my future postdoctoral advisors, Siva Theja Maguluri and Debankur Mukherjee, who encouraged me to apply for the Georgia Tech ISyE postdoctoral fellowship and supported my application, helping me secure this opportunity. Their work has been a source of inspiration for me, and I am very much looking forward to working with them.

I would also like to thank members of the academic community with whom I have interacted closely: Sid Banerjee, Anton Braverman, Xinyun Chen,

Yilun Chen, Zaiwei Chen, Sanjeeb Dash, Andrew Daw, Ton Dieker, Douglas Down, Kang Gong, Jin Guang, Guiyu Hong, Lucy Huo, Daniela Hurtado-Lange, Xiaopeng Li, Yueying Li, Yingdong Lu, Yashaswini Murthy, Hoang Nguyen, Jamol Pender, Pengyi Shi, Sean Sinclair, Mark Squillante, R. Srikant, Alexander Terenin, Sushil Varma, Jie Wang, Zedong Wang, Wentao Weng, Manxi Wu, Runhan Xie, Chen Yan, Zixian Yang, Christina Lee Yu, Jingwei Zhang, Qining Zhang, Yixuan Zhang, and many others whose company makes me look forward to conferences. I apologize to anyone I may have inadvertently omitted. Special thanks to Sushil Varma, who supported my applications and helped me prepare for interviews; Anton Braverman, who shared information that helped me prepare for the job market and offered encouragement; and Runhan Xie, who introduced me to professors in his department at the INFORMS conference.

I am grateful to my peers at CMU. Ziv Scully was the first senior CMU student I met and has since inspired me with his insight and enthusiasm for research. I have enjoyed talking with him. When I encountered difficulties, his patient listening was a great comfort. We also collaborated on a paper that later won a best paper award. Although our time together at CMU was relatively brief, Ziv has continued to root for me in academia and to spend time with me whenever we meet at conferences. Izzy Grosz was another senior student with whom I worked closely and who strongly influenced my research taste. They are remarkably knowledgeable and quick to understand new ideas; chatting with them often helped me clarify my own thoughts and research ideas. Drawing on their own experience, Izzy gave me simple, practical advice on academic writing, work habits, language, and life. Since becoming a professor, Izzy has continued to encourage and support me. Jalani Williams is my academic older brother. As the most senior student in Weina's group, he made me feel welcome from the beginning. I have always felt at ease talking with him about everything from research ideas to internet memes, anime, and manga.

I also learned from peers who brought different perspectives to my research. Shawn Shuoshuo Chen joined CMU one year after me. His interest in queueing theory sparked many stimulating discussions, and I learned a great deal from him about research on computer networks. Zhouzi Li joined CMU when I was in my fourth year. Our shared interests in restless bandits and queueing theory have given us much to discuss. I admire his quick thinking and dedication, and I have greatly enjoyed exploring challenging research questions with him. Xiangcheng Zhang visited our group in summer 2024, and we wrote a paper

together. I appreciate how quickly he tries new ideas and the clean notation choices in his notes. He made substantial contributions to the mathematics in Chapter 6.

I would also like to thank Nirav Atre, Ben Berg, Yiwei Lv, and Yuheng Qiu for their guidance, and Tuhinangshu Choudhury, Keerthana Gurushankar, and Neharika Jali for their thoughtful feedback on my talks. I am grateful to Ruihang Lai, Mingkuan Xu, and Yonghao Zhuang for the time we spent together and our many interesting conversations; Mingkuan also served on my writing skills committee. I also thank my officemates, Sanjith Athlur, André Cesar Costa, Theo Gregersen, Stephen Huan, Byungsoo Jeon, and Abhiram Kothapalli, for their companionship.

I have also been fortunate to have two wonderful roommates. I would like to thank my first roommate, Shuhua Yu, for helping me adjust to life in the United States and for sharing four wonderful years as roommates. I am especially grateful for the kindness and companionship he showed during the harder periods of my PhD. I would also like to thank my current roommate, Jingchu Gai. He invites me to cook with him every day; through our shared meals, I have learned to prepare many Chinese dishes. He is also resourceful, keeping me updated on recent developments in AI and helping me improve my workflow.

I would like to thank my friends in the anime fan circle: Rui Li, Jie Feng, Zhongxi Fang, and Shuoyuan Cui. I look forward to seeing them again after graduation. Special thanks to Rui Li. He has been my close friend since high school, and we share similar ways of thinking and a love of mathematics and anime. He provided steady support whenever I encountered difficulties, often through long conversations that helped me regain perspective. We also had a wonderful time traveling together.

I am fortunate to live in a world with great anime. Anime serves as my coffee and medicine and keeps me motivated. I am inspired by the heroism of Subaru Natsuki, admire the intelligence of Sora and Shiro, am moved by the deep emotions of Homura Akemi, and see something of myself in Hitori Gotoh. I would like to thank the creators of all of these works.

Finally, I would like to thank my parents. Although we have spent years apart, our weekly phone calls have always made me feel their presence and companionship. During these calls, we share family news, discuss my progress, and talk about interesting things happening at school. Even when they could not follow every detail of what I was trying to achieve, they paid close attention to how I felt and whether I was happy. During a particularly stressful period

of my job search, I felt misunderstood and lost. They took time to reflect, then called to tell me that they were proud of me and encouraged me to care for my health and daily routine. More recently, my father traveled to the United States to attend my defense. He took over the household chores and played table tennis with me, caring for me through ordinary actions just as he always has. My parents have always wanted the best for me without asking for anything in return or burdening me with their expectations. Their quiet love has given me the freedom to find my own path and the reassurance that I never have to walk it alone.

Contents

1	Introduction	1
1.1	Overview of problems	1
1.2	Challenges and a scalable path	4
1.3	Structure, results, and techniques	5
1.3.1	Basic setup and notation	5
1.3.2	Weak-coupling structure	6
1.3.3	One-dimensional structure	10
I	Weak-Coupling Structure: From Job Assignment to Restless Bandits and Weakly-Coupled MDPs	15
	Job assignment problem	16
	Motivation	16
	Results and impact	18
	Restless bandits and weakly-coupled MDPs	20
	Motivation	20
	Results and impact	22
2	Job Assignment with Time-Varying Item Sizes	27
2.1	Introduction	27
2.1.1	Background and motivation	27
2.1.2	Problem formulation: a simplified version	29
2.1.3	Main result	31
2.1.4	Approach overview	32
2.1.5	Chapter organization	35
2.2	Problem formulation	36
2.3	Main result and our approach	39
2.3.1	Main result	39

2.3.2	Our approach	39
2.4	Proposed meta-policy: JOIN-REQUESTING-SERVER (JRS) . .	43
2.4.1	How the single-server policy requests jobs	43
2.4.2	Description of JOIN-REQUESTING-SERVER (JRS) . .	44
2.4.3	Practical considerations in implementing JOIN-REQUESTING-SERVER (JRS)	46
2.5	Proof of Theorem 2.3 (Conversion Theorem) assuming irreducibility	48
2.5.1	Preliminaries	48
2.5.2	Steps and lemmas needed for the proof of Theorem 2.3 assuming irreducibility	49
2.5.3	Proof of Theorem 2.3 assuming irreducibility based on Lemmas 2.5–2.8.	51
2.5.4	More details on the system and proof of Lemma 2.6 .	53
2.5.5	Role of tokens and virtual jobs	59
2.6	Conclusion	60
2.A	Proof of Theorem 2.2 (Lower Bound)	62
2.B	Remaining Conversion Theorem proofs	65
2.B.1	Proof of Lemma 2.5	65
2.B.2	Proofs of Lemma 2.7 and Lemma 2.8	66
2.B.3	Proof of Conversion Theorem without assuming irreducibility	76
2.C	Solving the single-server problem	78
2.C.1	Lower bound via LP relaxation	79
2.C.2	Policy construction	82
2.C.3	Proof of Theorem 2.4	86
2.D	Performance under estimation error	87
2.D.1	Assumptions and result	87
2.D.2	Lemmas	90
2.D.3	Proof sketch for Lemma 2.23	91
2.D.4	Proof sketch for Lemma 2.24 and Lemma 2.25	93
3	Breaking the Global Attractor Assumption	97
3.1	Introduction	97
3.2	Additional related work	102
3.3	Problem setup	102
3.4	Follow-the-Virtual-Advice: A simulation-based framework	103
3.4.1	Single-armed problem	104

3.4.2	Constructing the N -armed policy	105
3.4.3	Discussion on FTVA and the role of virtual processes	106
3.5	Theoretical result on optimality gap	109
3.5.1	Synchronization Assumption	109
3.5.2	Bounds on conversion loss and optimality gap	110
3.6	Continuous-time restless bandits	112
3.7	Conclusion	115
3.A	More related work	116
3.B	Discussion on the computational costs of FTVA	118
3.C	Sufficient conditions for Synchronization Assumption (SA)	118
3.C.1	Preliminaries	119
3.C.2	Sufficient conditions based on self-loops	121
3.C.3	Sufficient conditions based on cycles	124
3.D	On relaxing the unichain condition	127
3.E	Proofs for discrete-time RBs	128
3.E.1	Overview	128
3.E.2	Lemmas and the proof of Theorem 3.2	130
3.F	Proofs for continuous-time RBs	133
3.F.1	Preliminary: finite synchronization time	133
3.F.2	Overview of the proof of Theorem 3.3	136
3.F.3	Proof of Lemma 3.19	138
3.F.4	Proof of Lemma 3.20	140
3.G	Experiment details and additional experiments	143
3.G.1	Experiment details of Figure 3.2	144
3.G.2	Experiment details of Figure 3.3	145
3.G.3	Varying the initial points of the Figure 3.2 and 3.3 examples	147
3.G.4	Visualization of the policies	149
3.H	Generalization to heterogeneous arms	154
3.H.1	Setting and algorithm	154
3.H.2	Proof for Theorem 3.27	156
4	Unichain & Aperiodicity Suffice for Asymptotic Optimality	159
4.1	Introduction	159
4.1.1	Our contributions	160
4.1.2	Chapter organization	161
4.2	Additional related work	161
4.3	Problem setup	162

4.3.1	Restless bandit problem	162
4.3.2	Scaled state-count vector	164
4.3.3	LP relaxation	164
4.3.4	Optimal single-armed policy	165
4.3.5	Additional notation	166
4.4	Main results	166
4.4.1	Algorithmic idea and focus-set policies	167
4.4.2	Set-expansion policy	168
4.4.3	ID policy	172
4.4.4	Connection with existing policies	174
4.4.5	Alternative formulation with inequality budget constraint	175
4.5	Focus-set meta-theorem and proof	176
4.5.1	Meta-theorem on $O(1/\sqrt{N})$ optimality gaps of focus-set policies	177
4.5.2	Proof of Theorem 4.9 assuming convergence to stationary distribution.	179
4.6	Proof of Theorem 4.2 (Optimality gap of set-expansion policy)	184
4.6.1	Subset Lyapunov functions	184
4.6.2	Lemmas verifying Conditions 1, 2 and 3; proof of Theorem 4.2	185
4.6.3	Proof of Lemma 4.14	186
4.6.4	Proof of Lemma 4.15	188
4.6.5	Proof of Lemma 4.16	194
4.7	Proof of Theorem 4.3 (Optimality gap of ID policy)	194
4.7.1	Subset Lyapunov functions	195
4.7.2	Focus set	196
4.7.3	Lemmas for verifying Conditions 1, 2 and 3; proof of Theorem 4.3	196
4.7.4	Proof of Lemma 4.19	198
4.7.5	Proof of Lemma 4.20	203
4.7.6	Proof of Lemma 4.21	206
4.8	Experiments	206
4.8.1	Comparing policies on two non-GAP examples	207
4.8.2	Commonness of non-GAP examples	209
4.8.3	Comparing policies on two non-SA examples	213
4.9	Conclusion and discussions	217
4.A	More SA counterexamples	218
4.B	On the aperiodic unichain assumption	219

4.B.1	Unichain conditions in prior work	219
4.B.2	Necessity of aperiodicity	221
4.C	Proof of LP relaxation upper bound	222
4.D	Proof of Theorem 4.9 in general case	224
4.E	Supplementary lemmas and proofs	226
4.E.1	Lemmas and proofs about matrix W and W -weighted L_2 norm	227
4.E.2	Lemmas and proofs about subset Lyapunov functions	229
4.E.3	Lemmas and proofs about L_1 norm	235
4.F	Set-optimization policy	238
4.F.1	Definition and asymptotic optimality	238
4.F.2	Proof of Theorem 4.27 (Optimality gap of set- optimization policy)	240
4.G	Results for inequality budget constraints	245
4.G.1	Policy modifications	246
4.G.2	Proof modifications	246
4.H	Experimental details	256
4.H.1	Implementation details of set-expansion policy	256
4.H.2	Additional details of the performance comparison ex- periments in Section 4.8	258
4.H.3	Generating and identifying non-GAP examples in Sec- tion 4.8.2	261
4.H.4	Commonness of non-GAP examples when the reward functions are not sparse	263
4.H.5	Comparing set-expansion policy with ID policy	264
5	Exponential Optimality Gap without a Global Attractor	271
5.1	Introduction	271
5.2	Problem setup	274
5.3	Two-Set Policy	276
5.3.1	LP relaxation	276
5.3.2	Overview of our main policy	279
5.3.3	Subroutine 1: π^* -Fixed-Ratio Control	280
5.3.4	Subroutine 2: LP-Priority	281
5.3.5	Constructing the Two-Set Policy	283
5.4	Main results	284
5.4.1	Optimality guarantee on Two-Set Policy	284
5.4.2	Discussion on the necessities of the assumptions	286

5.5	Analysis of two-set policy	288
5.5.1	Lyapunov framework	288
5.5.2	Warm-up: proving $O(1/\sqrt{N})$ optimality gap	289
5.5.3	Proving $O(\exp(-CN))$ optimality gap	291
5.5.4	Proof of Theorem 5.12	292
5.6	Experiments	293
5.7	Conclusion	298
5.A	Auxiliary facts for RB assumptions	298
5.A.1	Comparison of Assumption 5.2 with LP non-degeneracy	299
5.A.2	Validity of Assumption 5.3	300
5.B	Lemmas on subroutine dynamics and weighted norms	303
5.B.1	Transition dynamics under subroutines	303
5.B.2	Weighted L_2 norms for quantifying the convergence of distributions	306
5.B.3	Proof of Lemma 5.10	308
5.C	Two-set policy: exponential optimality proof	309
5.C.1	Subroutine conformity	310
5.C.2	Proof of Lemma 5.17	312
5.C.3	Proof of Lemma 5.18	326
5.C.4	Proof of Lemma 5.19	327
5.C.5	Proof of Lemma 5.20	328
5.D	Examples and proofs of violating assumptions	329
5.D.1	Examples of violating Assumptions 5.1 to 5.2	329
5.D.2	Proofs of Lemma 5.32 and Theorem 5.15	331
5.D.3	Remarks on Item (c) of regular unstable RBs (Defini- tion 5.14)	345
5.D.4	Proof of Corollary 5.16	349
5.E	Experiment details	350
5.E.1	Details of Two-Set Policy	350
5.E.2	Details of RB instances simulated in Section 5.6	357
6	Lyapunov Method for Heterogeneous WCMDPs	361
6.1	Introduction	361
6.2	Problem setup	365
6.3	ID policy with reassignment	369
6.4	Main results and technical overview	373
6.5	Experiments	378
6.A	Detailed review on related work	379

6.B	Experimental details	381
6.B.1	WCMDP instance generation	382
6.B.2	Definition of ERC policy	382
6.C	Proving the LP relaxation	382
6.D	Details of ID reassignment	385
6.E	Proof of main result	388
6.F	Lemmas and proofs for subset Lyapunov functions	391
6.F.1	Lemma for utilizing mixing time bound of transition matrices	392
6.F.2	Proof of Lemma 6.8	394
6.F.3	Properties of $h_{\text{ID}}(\cdot, \cdot)$	398
6.G	Lemmas for focus set	401
6.G.1	Proof of Lemma 6.13 (Majority conformity)	402
6.G.2	Proof of Lemma 6.14 (Almost non-shrinking)	406
6.G.3	Proof of Lemma 6.15 (Sufficient coverage)	409
6.H	Proofs of Lemma 6.9 and Lemma 6.10	410
6.H.1	Proof of Lemma 6.9	410
6.H.2	Proof of Lemma 6.10	411
6.I	Proof of the finite-time bound	414

II One-Dimensional Structure: Multiserver Queues 417

Universal bound for multiserver queues	418
Motivation	418
Results and impact	420
Gittins policy for $G/G/n$ with setup times	422
Motivation	422
Results and impact	423
Multiserver jobs	425
Motivation	425
Results and impact	426

7 Universal Bound for Multiserver Queues 429

7.1	Introduction	429
7.2	Related work	435
7.3	Problem setup	437
7.3.1	$GI/GI/n$ queue	437
7.3.2	Modified $GI/GI/n$ queue	439

7.4	Main results	441
7.5	An informal overview of the proof, with a roadmap	444
7.5.1	$M/M/1$: quadratic test function	444
7.5.2	$GI/GI/1$: “smoothed” quadratic test function	445
7.5.3	$GI/GI/n$	448
7.5.4	Roadmap for the rigorous proof	452
7.6	Technical framework and supporting lemmas	453
7.6.1	Leave-one-out systems	453
7.6.2	Basic Adjoint Relationship	455
7.6.3	Other supporting lemmas	458
7.7	Proof of main theorem	460
7.7.1	Step 1: exact expression of $\mathbb{E}[Q]$	461
7.7.2	Step 2: analyzing the covariance terms	464
7.8	Generalization to heterogeneous servers	467
7.8.1	$GI/GI/n$ queue and modified $GI/GI/n$ queue with heterogeneous servers	467
7.8.2	Assumptions and results	468
7.8.3	Technical framework and supporting lemmas	470
7.8.4	Proof of Theorem 7.25	473
7.9	Conclusion	477
7.A	Bounded mean residual time implies exponential tail	478
7.B	Stability of modified $GI/GI/n$	479
7.C	Proof of modified $GI/GI/n$ queue’s dominance	482
7.D	Proofs of supporting lemmas	484
7.D.1	Additional notation and definitions	484
7.D.2	Proof of BAR	486
7.D.3	Proof of bounded mean residual times (Lemmas 7.14 and 7.29)	491
7.D.4	Proof of the uncorrelation lemmas (Lemmas 7.15 and 7.30)	495
7.D.5	Proofs of some steady-state equalities and finiteness results	500
8	The Gittins Policy in $G/G/1$ and $G/G/k$ Queues	507
8.1	Introduction	507
8.1.1	Results: Performance Bounds and Heavy-Traffic Optimality	509
8.1.2	Main Obstacles and Key Ideas	510

8.1.3	Outline	512
8.2	Related Work	513
8.2.1	Optimal Scheduling in Queues	513
8.2.2	Setup Times	514
8.2.3	Decomposition Laws in Queues	515
8.3	Model	515
8.3.1	Core Queueing Models: $G/G/k$, $G/G/1$, $M/G/k$, and $M/G/1$	515
8.3.2	Setup Times	518
8.3.3	What the Scheduler Knows About Jobs' Sizes	519
8.3.4	The Gittins Policy	520
8.4	Main Results	521
8.4.1	Remarks on Main Results	523
8.5	Proof Overview	524
8.5.1	New Tool: Work Decomposition Law for G/G Arrivals	525
8.5.2	From Work Decomposition to Suboptimality Gap Bounds	526
8.5.3	From Suboptimality Gap Bounds to Heavy-Traffic Optimality	527
8.6	Background on WINE and r -Work	527
8.6.1	Additional Definitions for Reasoning About r -Work	528
8.7	Work Decomposition Law for G/G Arrivals	530
8.7.1	Palm Expectation and Notation	530
8.7.2	Statement and Proof of Work Decomposition Law	531
8.8	Bounding Gittins's Suboptimality Gap	534
8.8.1	Analysis of the Residual Interarrival Cost	535
8.8.2	Analysis of the Recycling Cost	535
8.8.3	Analysis of the Idleness Cost	535
8.8.4	Analysis of the Setup Cost	536
8.8.5	Proofs of Main Results	538
8.9	Heavy-traffic Optimality	538
8.10	Potential Extensions	541
8.10.1	Multiserver Jobs	542
8.10.2	Batch Arrivals	543
8.10.3	Generalized Vacations	544
8.11	Conclusion	545
8.A	Summary of Notation	546
8.B	Additional Model Details	548
8.B.1	Processor Sharing	548

8.B.2	General Definition of the Gittins Rank Function . . .	549
8.B.3	Technical Considerations for the Job Markov Process .	550
8.C	Deferred Proofs	550
8.D	Towards Stability of G/G/k/setup	554
8.D.1	Partial Proof Sketch	554
8.D.2	Remaining Obstacles	555
9	Sharp Waiting-Time Bounds for Multiserver Jobs	557
9.1	Introduction	557
9.1.1	Model and Performance Metric	559
9.1.2	Scaling Regimes	560
9.1.3	Results	562
9.1.4	Technical Challenges	563
9.1.5	Organization of the Chapter	564
9.2	Related Work	564
9.3	Model	567
9.4	Main Results	571
9.5	Proof Roadmap and Drift Method Preliminaries	572
9.6	Workload Bounds	573
9.6.1	Preliminaries for Lemma 9.7 and Lemma 9.9	574
9.6.2	Proofs of Lemma 9.7 and Lemma 9.9	576
9.7	Proof of Theorem 9.4 (Waiting Times under FCFS)	584
9.7.1	Proof Overview and Lemmas	584
9.7.2	Proof of Theorem 9.4	585
9.8	Proof of Mean Waiting Time Lower Bound	586
9.9	Proof of Theorem 9.6 (Mean Waiting Time under SNF)	588
9.10	Simulation Results	591
9.11	Conclusion and Future Work	598
9.A	Discussion on Settings without Commonness Assumption	599
9.B	Conditions for Applying Drift Method	600
9.B.1	Applying Drift Method to a Generic Markov Chain	600
9.B.2	Applying Drift Method to the Multiserver-Job System	601
9.C	Coupling with Infinite-Server Systems	602
9.D	Proofs of the Lemmas for Theorem 9.4	605
9.E	Proofs of Lemmas for SNF Waiting Time	609
9.E.1	A Tail Bound for Generic Markov Chains Based on Lyapunov Conditions	609
9.E.2	Tail Bounds for the Multiserver-Job System	610

<i>CONTENTS</i>	xix
9.F More on Experiment 2	621
Bibliography	625

Chapter 1

Introduction

Decision making in large stochastic systems is a central problem across computer systems, machine learning, and operations research. Such systems typically consist of a large number of interacting entities, which makes efficient and performant decision making especially challenging.

In this thesis, we study a wide variety of large stochastic systems. Despite their diverse problem definitions, these systems share similar underlying structures and admit similar fundamental techniques for performance analysis and policy design. By studying this set of example systems (each important on its own), we aim to demonstrate a broader philosophy: the essential behavior of a stochastic system is often governed by a minimal set of structural properties, and identifying these properties connects a wide range of seemingly different systems.

1.1 Overview of problems

We start by giving a high-level overview of the systems that we study, focusing on the motivation and models. A more detailed discussion of our approaches and results will be given in Section 1.3.

Stochastic bin packing with departures

Consider the following *job-assignment* problem in data centers (Chapter 2): A data center maintains tens of thousands of servers and constantly receives requests from users around the world. Each request creates a “job” that

demands a random amount of resources, such as compute, memory, and I/O bandwidth, from a server. Upon arrival, each job must be assigned immediately to a server that has the available resources. The goal is to pack the jobs as compactly as possible to maximize the utilization of the servers.

This problem can naturally be studied through the lens of stochastic bin packing, where each job is an item whose size is its resource requirement, and each server is a bin whose size is its resource capacity. Traditional job-assignment algorithms use bin packing heuristics, such as First-Fit or Best-Fit, which view each server as a bin and each job as an item. However, these heuristics are effective only when a job's resource requirement does not change during service, so that the job can be viewed as an item of fixed size.

In reality, however, a job's resource requirement often *fluctuates over time*. For example, a machine learning job may require more compute in one phase and more memory in another. Thus, a job's life cycle is better modeled as a *Markov chain* that transitions among several states, each corresponding to a different resource requirement. These time-varying requirements not only invalidate the simple bin packing heuristics, but also give rise to a *high-dimensional* stochastic control problem, where each job in service could be at different stages and have different resource requirements. It has remained open how to achieve near-optimal utilization in such a variant of stochastic bin packing problem.

Multiserver queues

Consider a bank with several tellers: Customers arrive at random time points and join the end of a single shared line, and each teller calls the next customer from the front of the line whenever they become free. The customer leaves the bank after a random amount of service time. We would like to understand how long the customers will wait in line, and if possible, how we could improve the average waiting times of the customers.

This is the simplest example of a *multiserver queue*: abstractly, the customers are the *jobs*, the tellers are the *servers*, and the waiting line is the *queue*. The performance metrics of interest are the mean waiting time and the mean queue length, which depend on the randomness in the jobs' arrival and service times.

The multiserver queue model applies whenever multiple servers work together on the same stream of jobs that cannot all be served immediately. An emerging application is the modeling of computer systems, where a pool

of servers (or CPU/GPU cores) works on a stream of jobs submitted by users over time.

Despite their simple structure, multiserver queues can exhibit highly complex behavior. When the number of servers is large and the jobs have general service-time distributions, the result is a high-dimensional stochastic system, in which each job may have received a different amount of service and thus have a different residual service-time distribution (Chapters 7 and 8). Further complications arise when a job must be served by multiple servers simultaneously (Chapter 9), or when servers can become temporarily unavailable (Chapter 8). These complexities capture essential aspects of real systems but remain poorly understood.

Weakly-coupled Markov decision processes

A large fraction of dynamic decision problems in stochastic systems fit into the framework of Markov decision process (MDP). The MDP abstracts a stochastic decision problem into several elements: state, action, transition probability, reward. At each state, the decision-maker picks an action, which causes the system to randomly transition to a new state. A certain amount of reward is also generated, depending on the current state and action. The objective is to maximize the reward in a proper sense; we are particularly interested in maximizing the infinite-horizon average reward. MDP plays a central role in operations research and machine learning.

In real life, instead of controlling one MDP, we sometimes need to control *multiple MDPs* simultaneously. One such formulation is called weakly-coupled MDPs (WCMDPs): the decision-maker controls a set of MDPs, each with its own state-transition dynamics and reward structure, evolving independently. However, the actions are subject to some joint constraints. To give an example, imagine the wireless control of multiple drones, where the controller can only send signals to steer the movement of a subset of them at every time step, due to the communication bandwidth constraints.

The WCMDP is a very general formulation, with many well-known special cases, such as (rested) Markovian bandits and restless bandits, each playing an important role in resource allocation and machine learning. Applications of WCMDPs include machine maintenance (deciding which machines to repair), wireless scheduling (allocating channels to users), online advertising (allocating a budget across many campaigns), and healthcare (selecting adaptive

treatments across many patients), among many others. We defer a detailed review of the literature to the corresponding chapters (Chapters 3 to 6).

Despite their importance, WCMDPs are extremely hard: even for the special case of restless bandits, computing an exact optimal policy is known to be intractable. Intuitively, the complexity arises from the joint constraints, which require the decision-maker to coordinate the states and actions of the component MDPs jointly, rather than maximizing their reward individually. This leads to a high-dimensional control problem where the system’s state space is the Cartesian product of the component MDPs’ state spaces. Prior work has developed scalable policies for systems with a large number of component MDPs, but provable near-optimality has been established only under strong assumptions, leaving open whether it can be achieved more generally.

1.2 Challenges and a scalable path

In studying these large stochastic systems, two major challenges arise:

- **Curse of dimensionality.** A large stochastic system consists of many interacting components, so its state space grows exponentially with the number of components.
- **Diversity of problems.** Different applications give rise to a wide range of problems, and small changes in a problem often change its solution significantly. For example, a queueing system can behave very differently under the exponential distribution than under other distributions.

Approaches that sit at one of two extremes tend to struggle with one of these challenges. Structure-agnostic methods (such as vanilla dynamic programming) apply generally, but they inherit the curse of dimensionality and do not scale well. Methods that exploit very specific structure can be efficient, but their solutions are problem-specific, so there are always more problems than solutions.

We look for a scalable path between these two extremes, aiming to efficiently solve as large a class of problems as possible. The guiding idea is to identify generalizable structures shared among many problems and to develop techniques adapted to each structure, so that seemingly different problems can be grouped together and addressed by similar techniques. This idea is not new in itself; the purpose of the thesis is rather to provide a concrete demonstration of it.

1.3 Structure, results, and techniques

We identify two structures, each shared by a family of problems with different definitions; each structure reduces a high-dimensional system to a tractable proxy. The first is the *weak-coupling structure*, where the system is approximated by a collection of independent subsystems. The second is the *one-dimensional structure*, where a key quantity of the system is approximated by a one-dimensional process, even when the full state is high- or infinite-dimensional.

These two structures are studied in Part I and Part II, respectively; each part opens with an overview of the motivation, results, and impact of *each chapter*. By contrast, the two subsections below give a *cross-cutting view*: they explain each structure and the techniques common to its chapters, illustrating them with one result in detail and summarizing the rest more briefly.

1.3.1 Basic setup and notation

To ground the discussion, we introduce a small amount of notation in this section.

Here we use the language of *Markov decision processes* to describe our problems. For illustration, we suppose the system runs in discrete time; the continuous-time setting is analogous. At each time step $t = 0, 1, 2, \dots$, the system has a *state* $\mathbf{S}_t$ in a state space $\mathcal{S}$, and a decision-maker chooses an *action* $\mathbf{A}_t$ from an action set $\mathcal{A}$. Given the current state $\mathbf{S}$ and action $\mathbf{A}$, the system moves to a new state $\mathbf{S}'$ with *transition probability*

$$P(\mathbf{S}' \mid \mathbf{S}, \mathbf{A}) = \mathbb{P}(\mathbf{S}_{t+1} = \mathbf{S}' \mid \mathbf{S}_t = \mathbf{S}, \mathbf{A}_t = \mathbf{A}),$$

and earns a *reward* $r(\mathbf{S}, \mathbf{A})$.

A *policy* $\pi(\mathbf{A} \mid \mathbf{S})$ samples the action given the state. When the policy is fixed, the state alone evolves as a *Markov process*. Running the system under a fixed policy produces a trajectory $\mathbf{S}_0, \mathbf{S}_1, \mathbf{S}_2, \dots$, and the performance objective is the long-run average reward

$$\bar{r}(\pi) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} r(\mathbf{S}_t, \mathbf{A}_t).$$

This long-run average can model the *first moment* of any steady-state quantity of interest.

All the problems introduced in Section 1.1 (bin packing, multiserver queues, and WCMDPs) can be cast as a single MDP of this form, where the state $\mathbf{S}_t$ is the *entire* system state and the reward is chosen so that $\bar{r}$ is the performance metric of interest. For a WCMDP, $\mathbf{S}_t$ collects the states of all its component MDPs, and the reward is the total reward of all component MDPs. For multiserver queues, $\mathbf{S}_t$ may include the number of jobs of each type, and the residual service times for the jobs that have received any service; the reward could be utilization, the number of job completions in the time slot, or the (negative) queue length.

The problems in this thesis come in two flavors:

- *Performance analysis.* The policy is fixed, and we want to understand how well the system performs, ideally through a simple, closed-form characterization of $\bar{r}$.
- *Policy design.* We seek a policy that (approximately) maximizes $\bar{r}$, while remaining simple and efficient to compute.

A recurring theme of this thesis is that a large, high-dimensional system can often be well approximated by a *proxy* system that is far easier to analyze. The two structures we study differ in *which* proxy the system admits. In the *weak-coupling structure* (Part I), the system is approximated by a collection of independent low-dimensional subsystems. In the *one-dimensional structure* (Part II), its key quantities of interest are approximated by a single one-dimensional process. We introduce each in turn below.

1.3.2 Weak-coupling structure

In Part I of this thesis, we consider the *weak-coupling structure*. The main motivation for considering this structure is to address the curse of dimensionality when the system state $\mathbf{S}_t$ is represented as a high-dimensional vector $(S_{t,1}, S_{t,2}, \dots, S_{t,n})$ for some large n .

We say a system has the *weak-coupling structure* if it can be approximated by a set of independent subsystems in a proper sense. Weak coupling is a loose suggestive term and its specific meaning depends on how we will use it:

- For performance analysis, it means the original system state $\mathbf{S}_t$ can be approximated by treating its coordinates $S_{t,i}$ as independent Markov processes.

- For policy design, it means there exists a “good-enough” policy π under which $\mathbf{S}_t$ can be approximated by treating its coordinates $S_{t,i}$ as independent Markov processes.

The weak-coupling structure allows us to analyze the performance or design policies through a proxy system with independent components, which can be done efficiently. This circumvents the curse of dimensionality at the cost of a small approximation error.

Weak coupling can sometimes be non-trivial to identify. Take the bin packing problem introduced in Section 1.1 as an example: Recall that in this problem, a centralized scheduler needs to assign incoming jobs to one of the servers. Consider the state of the i -th server, $S_{t,i}$, which represents the service phases of the jobs that it is currently hosting. From the problem definition, $S_{t,i}$ does *not* seem to be approximately independent at all. In particular, the scheduler reads the system state $\mathbf{S}_t$ to assign a new job to one of the servers, which then changes the state of the target server $S_{t,i}$ and creates a correlation between $S_{t,i}$ and the other coordinates of $\mathbf{S}_t$; the correlation can be arbitrarily strong, depending on the job-assignment policy. Nevertheless, we show that there exists a good policy π under which $S_{t,i}$ ’s are approximately independent and the system achieves near-optimal utilization.

We identify and utilize the weak-coupling structure in the bin packing problem through the following *local-to-global policy conversion* framework (See Section 2.1.4):

1. First, we construct a truly decoupled system where each server receives jobs from an independent pool.
2. Then we solve for the optimal policy in the decoupled system (in this case, a set of “optimal single-server policies”).
3. Finally, we construct a policy in the original coupled system that approximately matches the optimal dynamics of the decoupled system.

This process establishes the near-optimality of the constructed policy, after further showing that the optimal utilization of the decoupled system is always better than that of the original system. We have thus found the first near-optimal policy for stochastic bin packing with time-varying item sizes. We refer the readers to Chapter 2 for the details of this work.

To summarize this approach, for a system that we suspect has the weak-coupling structure, we first construct a proxy system that consists of truly independent components, and try to match the dynamics of the original

system with the proxy system. This would then help us simultaneously design a good policy and establish its performance guarantee.

This approach turns out to be surprisingly general, and we have used it to resolve some long-standing puzzles for weakly-coupled MDPs (WCMDPs). In Chapters 3 to 6, we study the following aspects of restless bandits (RBs), the canonical special case of WCMDPs, and then of WCMDPs in full generality:

- **When do scalable, near-optimal policies exist?** Prior work achieves near-optimality in RBs under a restrictive and hard-to-verify assumption called the *global attractor property (GAP)*. This assumption had long been believed necessary, but we find a way to remove it completely (Chapters 3 and 4).
- **How fast does the optimality gap vanish?** We examine a stronger notion of near-optimality for RBs, called *exponential asymptotic optimality*, in which the suboptimality vanishes exponentially fast as the number of MDPs grows. We establish a minimal set of conditions under which it can be achieved (Chapter 5).
- **How much heterogeneity can be tolerated?** Most prior work assumes the component MDPs, or *arms*, are identical, or come from a fixed number of types. This assumption enables mean-field-based policies and analysis, which break down when the system is fully heterogeneous, with every arm having distinct dynamics. We establish the first near-optimality result for RBs and WCMDPs under full heterogeneity (Chapter 6).

A closer look at our techniques. The rest of this subsection comments on the techniques and how they relate to prior work; a reader interested only in the results may skip it.

First, our approach coincides with those in the prior work in the first step, i.e., we consider the same decoupled system obtained from a linear program relaxation. Nevertheless, we develop completely different ideas to match the original system with the decoupled system, leading to different policies and stronger performance guarantees.

Second, our techniques have gradually evolved along this line of work on RBs and WCMDPs. Chapter 3 uses the same policy as the one in Chapter 2, which only partially removes the GAP assumption; Chapter 4 uses a different (but simpler) policy, which better handles the more general chain structure of MDPs and completely removes GAP. Chapter 6 further improves on the

Lyapunov function to derive a bound in the fully heterogeneous setting. Below we highlight some novel techniques developed along the way, which could be of independent interest:

- Stein’s method for proving approximate independence (Section 2.5.2). We show that the stationary distribution of the bin packing system is close to that of a system with independent bins through Stein’s method. It demonstrates the usefulness of Stein’s method in comparing the stationary distributions of high-dimensional Markov chains.
- Control by simulating a coupled virtual system (Section 2.4.2 and Section 3.4). This method demonstrates an interesting policy-design idea: we can simulate a virtual system with a desired property (such as independence) and use it to guide policy design. Specifically, the policy should keep the original system as tightly coupled with the virtual system as possible.
- Bivariate Lyapunov functions (Sections 4.5 and 5.5). This technique combines a set of “partial” Lyapunov functions into one. It is useful for analyzing convergence that consists of multiple phases.
- Lyapunov function for a set of independent but heterogeneous Markov chains (Section 6.4). This function quantifies the “distance” between the state realizations of these Markov chains and their stationary distributions, and it decreases as these Markov chains mix. The challenge lies in comparing a vector of *state realizations* with a vector of *state distributions*. The Lyapunov function is constructed using a novel *projection-based* idea, which may be useful for more general heterogeneous many-body systems.

Relation to mean-field methods. Finally, we comment on the relationship between our approach and mean-field approximation [202]. Mean-field approximation is the default approach for analyzing large systems with many interacting components; roughly speaking, it approximates the empirical state distribution of $\{S_{t,1}, S_{t,2}, \dots, S_{t,n}\}$ by running a deterministic approximation of the original stochastic dynamics. Prior work indeed establishes performance guarantees using the mean-field approximation [107, 108, 266, 274], but these results suffer from two limitations inherent to mean-field approximation:

- First, the approximation is asymptotically accurate only on finite horizons, and an additional condition (the global attractor property) is needed to certify the long-term error.
- Second, the mean-field approximation usually requires $S_{t,i}$'s dynamics to be homogeneous across i , so that the empirical state distribution contains all information of the system state $\mathbf{S}_t$.

Exceptions exist: Some follow-up work of ours [e.g., 121, 294] is able to prove GAP under some new policies and establish near-optimality, but their results are restricted to the homogeneous setting; Allmeier and Gast [8, 9] extended the mean-field analysis to a large class of heterogeneous large stochastic systems, but the results are only for transient distributions. Our approach can address these two limitations simultaneously for RBs and WCMDPs, mainly because the local-to-global conversion approach allows us to identify the hidden independence structure in RBs and WCMDPs that was not utilized by prior work, and we directly work with $\mathbf{S}_t$ rather than the empirical state distribution.

1.3.3 One-dimensional structure

In Part II of this thesis, we consider the second type of structure, which we refer to as the *one-dimensional structure*. The one-dimensional structure loosely refers to the fact that some key quantities associated with the target system under study are approximated by a one-dimensional Markov process, such as a birthdeath process or a single-server queue. Since a one-dimensional process is often easier to study, this allows us to approximate those key quantities efficiently.

In this part of the thesis, we focus on the multiserver queues. For a multiserver queue, the quantities of interest are total queue length, work (amount of server-time to complete all jobs in the queue), or other variants, which ultimately translate to the delay performance of the queue. Depending on the specific system, some of these quantities may be approximated by a one-dimensional process. However, proving such an approximation could be highly non-trivial, especially for the multiserver queues with high-dimensional states.

We take the $G/G/n$ queue as an example. The $G/G/n$ queue (more rigorously, $GI/GI/n$ queue) is the n -server queue with general i.i.d. interarrival-time and service-time distributions. There is a long pursuit in queueing theory

for a simple, accurate upper bound for the steady-state mean queue length of $G/G/n$ under First-Come-First-Served scheduling. When $n = 1$, Kingman’s bound shows that the steady-state mean queue length $\mathbb{E}[Q]$ is bounded as

$$\mathbb{E}[Q] \leq \frac{\rho^2 \text{Var}[\mu S] + \text{Var}[\lambda A]}{1 - \rho},$$

where $\text{Var}[\lambda A]$ and $\text{Var}[\mu S]$ denote the variances of normalized interarrival and service times, and $\rho < 1$ denotes the load or utilization of the system. People suspect that a similar bound holds for $n > 1$, but the generality and tightness of the prior bounds are far from Kingman’s bound. Specifically, most prior work is restricted to certain limiting regimes where n and ρ scale jointly in a particular way. A recent breakthrough by Li and Goldberg [180] shows that $\mathbb{E}[Q] \leq C/(1 - \rho)$ for some constant C , which achieves the right order $O(1/(1 - \rho))$ universally for all n and ρ . However, the constant C in this bound is astronomically large.

Intuitively, the $G/G/n$ queue is challenging to analyze due to its high dimensionality: its state needs to track the remaining service times of all n jobs in service. When n is large, it is very hard to solve this high-dimensional system and get a closed-form formula for its mean queue length.

In our work, we make the observation that the queue length of $G/G/n$ can be approximated by a $G/G/1$ queue — the latter is a one-dimensional process, whose mean queue length can be bounded via Kingman’s bound. This observation is not new — prior work has extensively used this idea to establish $O(1/(1 - \rho))$ bounds in the regime when n is fixed and $\rho \rightarrow 1$ [e.g., 128, 245]. Intuitively, when ρ is close 1, the queue length is long, keeping the n servers busy for most of the time. When the n servers are all busy, they process the jobs at a constant total rate, so we can treat them as a single “superserver” that is n times faster than each server. However, this intuition only applies when n is fixed and ρ converges to 1. Yet it has remained unclear whether $G/G/n$ still behaves like $G/G/1$ when n is large — in particular, when $n > 1/(1 - \rho)$.

Despite the challenges when $n > 1/(1 - \rho)$, we show that the queue length of $G/G/n$ can be approximated by $G/G/1$ for all n and ρ . This allows us to get a much simpler and tighter $O(1/(1 - \rho))$ bound than the one in [180]. For light-tailed service times, the leading constant in our bound is orders

of magnitude smaller than that in prior work. For instance, we bound the $M/GI/n$ queue's mean queue length by simply

$$\mathbb{E}[Q] \leq \frac{1}{1 - \rho}$$

for Poisson arrivals and any New-Better-than-Used-in-Expectation (NBUE) service times. Similarly clean bounds are available for gamma, phase-type, and bounded service times. We also extend our results to $GI/GI/n$ queues with *fully heterogeneous* service-time distributions, a setting not addressed by prior universal bounds. The key intuition is that when n is large, the service processes and the queue lengths are all *approximately independent* in a proper sense. We prove this asymptotic independence via a novel *leave-one-out coupling* technique. We refer the readers to Chapter 7 for details of this work.

The one-dimensional structure also shows up in our other works on multi-server queues, even though the detailed techniques and settings differ. Their results and techniques are summarized below:

- In Chapter 8, we study the optimal scheduling of the $G/G/n$ queue (or the “ $G/G/k$ ” queue under the convention of the original paper). We show that the Gittins index policy achieves heavy-traffic optimality when n is fixed and $\rho \rightarrow 1$. This system has an infinite-dimensional state representation, as the number of jobs that have received some service under Gittins policy could be unbounded. Nevertheless, we show that a quantity called r -work, a generalization of the total work defined above, has similar dynamics in $G/G/n$ and $G/G/1$, and that a linear combination of the r -work and the residual arrival time is almost a one-dimensional Markov process, which lets us approximate the mean number of jobs in the system.
- In Chapter 9, we study the optimal scheduling of the *multiserver-job model*. Here, jobs occupy multiple servers simultaneously during service, and the *server need* of the jobs can vary across many orders of magnitude. For this model, we show that the First-Come-First-Served (FCFS) policy is strictly suboptimal in mean waiting time, prove a lower bound on the mean waiting time of all policies, and prove the order-wise optimality of a policy that we call Smallest-Need-First (SNF). The key argument in this work is a sequence of state-space concentration arguments, which restricts the system to a good subset of the state space. In this subset, a *one-dimensional quantity* called *workload* has a simple characterization.

The characterization of workload then allows us to analyze the mean waiting times of FCFS and SNF, and prove a lower bound.

Finally, we comment on the relationship between our techniques and those in the literature. Approximating a complex system with a one-dimensional process is not a novel idea. In particular, the relationship between the n -server queue and single-server queue has been exploited in a large number of prior works; it naturally arises when one tries to generalize results from single-server queues to multiserver queues. Recently, the comparison between multi- and single-server queues has also been found useful for studying complex scheduling policies [e.g. 125, 242, 245], or complex job models such as multiserver jobs [e.g. 128, 129]. Beyond multiserver queues, there has also been a large body of literature on the Complete Resource Pooling condition, under which a complex queueing system with multiple types of jobs or servers behaves like a single-server queue in heavy traffic [e.g. 140, 158]. Our work can be viewed as further evidence of the effectiveness of the one-dimensional approximation idea for systems that are seemingly high-dimensional, such as the $G/G/n$ queue with a large n , or any queueing model with a general service-time distribution, where an unbounded number of jobs could be partially served but not yet completed.

Part I

Weak-Coupling Structure: From Job Assignment to Restless Bandits and Weakly-Coupled MDPs

This part of the thesis develops the *weak-coupling structure*, where a high-dimensional system can be approximated by a collection of independent subsystems. We apply this idea to stochastic bin packing (Chapter 2), restless bandits (Chapters 3 to 5), and fully heterogeneous weakly-coupled MDPs (Chapter 6). Below, we give an overview of each chapter in this part: first stochastic bin packing, then the three chapters on restless bandits and weakly-coupled MDPs together. Each overview consists of a motivation followed by results and impact.

Job assignment problem

Motivation

What is the problem and why we care? Today’s data centers consume enormous amounts of energy yet run at low utilization. For example, Google’s Borg clusters have been reported to operate at CPU and memory utilization below 60% [258]. The low utilization is caused by multiple factors, with an important one being the efficiency of job assignment decisions. Specifically, each job demands a certain amount of resources, such as compute, memory, and I/O bandwidth, from a server of finite capacity. When a server’s remaining capacity is too small to accommodate another job, that capacity is wasted. What makes job assignment especially challenging is that a job’s resource requirements are often not fixed but vary over time [28, 80, 190, 234, 240, 258]. For example, a container job hosting a website requires different amounts of CPU and memory as its traffic rises and falls.

Accounting for this time variability is crucial for efficient job assignment. If we ignore the time variability and reserve conservatively according to the peak, we waste capacity. At the other extreme, if we reserve according to the off-peak values, then when the jobs later transition to more resource-consuming phases, they compete for resources and service quality degrades. Therefore, efficient packing must account for the full dynamics of the jobs’ resource requirements, which calls for a new model and a new algorithm.

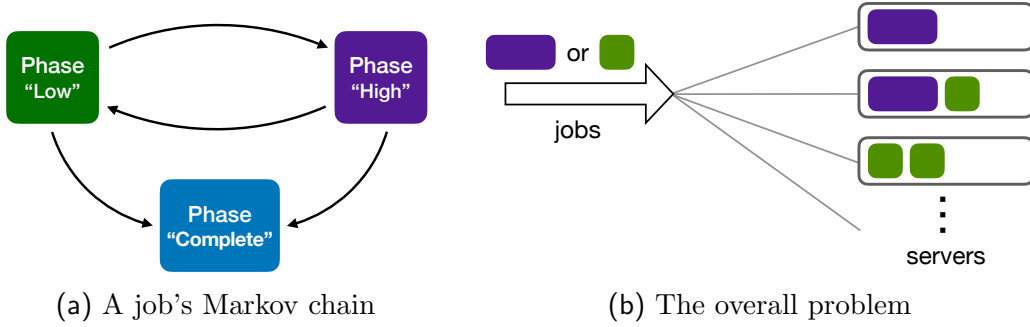


Figure I.1: The model of stochastic bin packing with time-varying item sizes. Each job is a Markov chain over a high- and a low-requirement phase, absorbed upon completion (a); such jobs arrive over time and are assigned to servers of finite capacity (b).

Stochastic bin packing with time-varying item sizes To study the job assignment problem, we propose the following variant of *stochastic bin packing with departures*, where we view each job as an item whose size is its resource requirement, and each server is a bin whose size is its resource capacity. In this model, items arrive over time, and each item leaves the system after a random amount of time. Since item sizes typically vary stochastically, we use Markov chains to model their dynamics. A simple example of the job model is shown in Figure I.1a, where a job transitions between a high-requirement phase and a low-requirement phase, and jumps to an absorbing state when it completes. The goal is to maximize server utilization, which is equivalent to minimizing the number of servers in use, subject to a resource contention constraint on each server. The overall model is illustrated in Figure I.1.

Prior work and what makes the problem hard In prior work, many variants of stochastic bin packing have been studied, but none consider time-varying item sizes. The classical model is *stochastic bin packing without departures*, which has a rich history in operations research and theoretical computer science [e.g., 25, 65, 66, 68, 94, 132]. Another line of work, more closely related to ours, is *stochastic bin packing with departures*, proposed more recently in the context of service systems [110, 194, 196, 197, 225–228, 252–256, 291]. For stochastic bin packing with departures, prior work has constructed job assignment policies that achieve *asymptotic optimality* in the following sense: when the number of bins and the jobs’ arrival rate

grow to infinity proportionally, the utilization under the policy converges to optimality [110, 252–256]. However, since all of this work requires the item sizes to be fixed over time, its policies and techniques do not generalize to the time-varying setting of interest.

The main challenge with time-varying requirements is that the current size of an item reveals little about its future size: an item that looks small now may soon enter a resource-demanding phase and consume much of the bin’s capacity. A good policy must therefore anticipate how that size will evolve. Yet an item’s future evolution is a whole stochastic process, and it is unclear how to condense such rich, forward-looking information into good assignment decisions.

Results and impact

Contribution 1 We propose a stochastic bin packing model with time-varying item sizes, which captures how the resource requirements of data-center jobs change over time. This allows us to quantitatively study and optimize job assignment decisions.

Contribution 2 We propose the first policy for this problem that is provably asymptotically optimal. The policy is efficiently computable, only requiring us to solve a subproblem related to the dynamics of a single server.

What is new in our approach? We propose a novel *local-to-global policy conversion* framework to construct our policy and establish performance guarantees. This framework allows us to efficiently solve for a near-optimal policy in two steps:

- First, we identify a single-server problem, which, roughly speaking, decides “when a single server should request jobs to maximize the throughput”.
- Second, we convert the optimal policy for the single-server problem to a policy for the original problem.

Concretely, the server decides when to *request* a job using only its own jobs’ phases and remaining capacity. The conversion then lets each server *independently* run this single-server rule, and a central dispatcher routes each arriving job to a server that is currently requesting one.

Intuitively, under this global policy, the dispatcher is the *only* coupling among the servers. This coupling *vanishes as the system grows*: with many servers, their requests arrive at rates that match the job stream, so nearly every request is filled without delay, and each server behaves as if running its single-server policy *in isolation*.

This large-system behavior is, in fact, half of our optimality proof; it establishes the *achievability* direction, showing that any single-server policy can be approximately matched by an original-system policy when the system is large. The other half is a *converse*: every policy of the original system is itself no better than some single-server policy, which has more flexibility because it can request a job whenever it is ready, rather than waiting for one to arrive. Together, these two directions constitute a *bi-directional mapping* between the two systems’ policies, and establish asymptotic optimality.

Practical impact Modern jobs in computing systems exhibit increasing time variability in their compute, memory, and bandwidth requirements. An emerging application is the serving of Large Language Model (LLM) inference jobs, whose memory usage grows dramatically during service [12, 62, 160, 209], and this growth is becoming increasingly significant as the context length and model size continue to scale. Our work provides an algorithmic tool that could better schedule such jobs and ultimately improve the energy efficiency of data centers.

Broader applicability Our local-to-global conversion framework is essentially independent of the details of the servers’ dynamics. This suggests that we may be able to tackle new, more complex problems with the same framework. In principle, a large-scale, many-component system could be analyzed through a single-component proxy, whose definition is flexible enough to capture richer dynamics and operations.

Two such examples arise within the job assignment problem itself.

- The first is *job migration*: moving a job from one server to another partway through its service, which can improve efficiency in some settings [204] but has rarely been studied. With migration, we would optimize not only where to place arriving jobs, but also which jobs to move and where to move them. To study this, we might give the single-server problem two new decisions, evicting a job and admitting a

migrated one, on top of its original decision of when to accept a new job.

- The second is *queueing*, where a job assigned to a server may have to wait in a local queue when the server lacks the capacity to start it immediately. With a queue, we would optimize not only where to place arriving jobs, but also the order in which each server serves its queued jobs. To study this, we might add a queue to the single-server problem, so that, on top of deciding when to accept a new job, it also decides the order in which its queued jobs are served.

Pushed further, by viewing each server as a full Markov decision process, the same viewpoint has inspired our line of work on restless bandits and weakly-coupled Markov decision processes [149, 151, 153, 299], which we present next.

Restless bandits and weakly-coupled MDPs

Motivation

What is the problem and why do we care? Many resource allocation and control problems consist of a set of *weakly-coupled* components. For example, consider a hospital ward with many patients but only a few nurses. Each patient’s condition changes over time, partly on its own and partly in response to the care the patient receives; at each moment the nurses can attend to only a few patients and must decide which. The goal is to allocate this limited attention over time so as to keep the patients as healthy as possible.

The same logic recurs across many domains: a maintenance crew servicing a fleet of machines that wear out at different rates, a limited advertising budget spread across many users, content moderators reviewing a stream of posts, or an interviewer deciding which applicants to meet. In every case, we repeatedly allocate a scarce resource among many components that would each evolve independently were it not for a shared budget that ties the decisions together.

This structure is captured by *weakly-coupled Markov decision processes (WCMDPs)*. Each component is modeled as a *Markov decision process (MDP)*, which has a state that evolves stochastically and can be steered by actions. Crucially, the components are *independent in their dynamics* and are coupled

only through joint constraints on the actions, such as a budget on how many can be served at each step. The task is to choose actions over time to maximize the *long-run average reward over an infinite horizon*, subject to the budget constraints.

A canonical special case of WCMDPs is the *restless bandit*, where each component, or *arm*, has only two actions, “activate” or “keep passive”, and the budget limits how many arms can be activated at once.

RBs and WCMDPs have a rich history and find applications across a spectrum of domains, including wireless communication [3], congestion control [23], inventory control [13], crawling web content [22], machine maintenance [115], and clinical trials [269]. We refer the reader to the recent survey [218] for a detailed discussion of these applications.

Research questions The problem is challenging because of the curse of dimensionality: with a large number of stateful components, the number of possible combinations of their states is exponentially large. While the optimal policy for the special case of *rested bandits* is known (the famous Gittins index policy), computing the optimal policy for general RBs is computationally challenging, and so is the more general WCMDP problem.

Therefore, we aim not for exact optimality but for policies that are efficiently computable and provably *asymptotically optimal when the system is large*, that is, the optimality gap diminishes as the number of components grows. This goal raises the questions that drive our line of work on this topic: under what conditions can asymptotic optimality be achieved, and how weak can those conditions be made? Can we handle fully *heterogeneous* components? And how *fast* does the optimality gap vanish as the system grows?

Prior work and what makes the problem hard Prior work pursues this same goal, but achieves asymptotic optimality only under restricted conditions: Whittle index policy [274, 283] and LP-Priority policy [266] have been proved to be asymptotically optimal for RBs under a strong assumption called Global Attractor Property (GAP), which is hard to verify and could fail in many cases. Moreover, existing results rely on homogeneity, i.e., the arms are identical or can be partitioned into a few groups that are internally identical. These results do not apply to the fully heterogeneous case, where the arms have mutually distinct transition dynamics or reward functions.

A separate line of work has focused on the finite-horizon total-reward objective (or the infinite-horizon discounted-reward objective) for RBs and WCMDPs. Most of these works achieve asymptotic optimality for large systems under a fixed horizon, but the optimality gap of their policies grows quadratically or even exponentially as the (effective) time horizon increases. The complexity of computing these policies also grows with the (effective) time horizon.

Results and impact

Contribution 1: asymptotic optimality under minimal conditions

We obtain a better theoretical understanding of this fundamental class of problems.

- **Asymptotic optimality without GAP** Prior work achieves asymptotic optimality under the *global attractor property (GAP)* assumption, which asserts global convergence of a non-linear system; it is hard to verify and could fail in many cases. We completely remove the long-standing GAP assumption, achieving asymptotic optimality under only a weaker-than-standard aperiodic unichain condition (Chapters 3 and 4). Our policy-design approach is significantly different from prior approaches, which are mainly index or priority policies. These policies rely on a static ranking of arm states, and greedily activate the highest-ranked arms. Our policy, in contrast, takes a more global approach: it actively controls the empirical distribution of the arm states to ensure the global convergence that GAP would otherwise have to assume.
- **Minimal conditions for exponential asymptotic optimality** We examine a stronger notion of asymptotic optimality for RBs, called *exponential asymptotic optimality*, in which the optimality gap vanishes exponentially fast as the number of arms grows. We establish a minimal set of conditions under which exponential asymptotic optimality can be achieved. In particular, these conditions do not involve GAP, are easy to verify, and are proved to be necessary with current techniques (Chapter 5).
- **Asymptotic optimality under full heterogeneity** We establish the first asymptotic optimality result for WCMDPs under full heterogeneity, allowing the arms to have mutually distinct reward functions and transition dynamics (Chapter 6).

Contribution 2: a rich set of policies We obtain a collection of new policies, suitable for a wide range of scenarios, that greatly enrich the existing toolbox. These policies can be computed as efficiently as those in prior work (the Whittle index and LP-Priority policies). Specifically,

- The first one is the *ID policy* (Section 4.4.3). It is simple to implement, demonstrates robust performance across a large class of RB problems, and has asymptotic optimality guarantees under only a weaker-than-standard aperiodic unichain assumption. Moreover, a modified version of the basic ID policy works for fully heterogeneous WCMDPs (Section 6.3).
- Another policy is called the *set-expansion policy* (Section 4.4.2). It has similar performance and the same performance guarantees as the ID policy for homogeneous RBs, and treats different arms more symmetrically.
- The third policy is called the *two-set policy* (Section 5.3). It has stronger performance than the ID policy on instances where two additional conditions are satisfied, and degrades to the same performance as the ID or set-expansion policy without these conditions. Theoretically, this policy achieves exponential asymptotic optimality under the weakest known conditions. It is slightly more complex to implement.
- The fourth policy is called *Follow-the-Virtual-Advice (FTVA)*. It has slightly weaker performance than the previous three policies and requires stronger conditions for asymptotic optimality. However, it comes with a clean and explicit optimality-gap bound.

What is new in our approach? We further develop the local-to-global conversion framework from the job assignment problem. The local problem is now a *single-armed problem* for each arm: the optimal control of that arm under long-run average budget constraints. Its solution is a *single-armed policy*, and our conversion constructs a global policy that *matches the long-run behavior* of the decoupled system in which every arm independently runs its own single-armed policy.

This departs from the classical index and priority policies (i.e., Whittle index and LP-Priority). Those methods start from the same relaxation, but extract from it a scalar index for each state measuring the benefit of activating an arm in that state, and then *greedily* activate the arms of highest index. Our approach instead uses the single-armed policy to match the long-run behavior

of the two systems. This lets our policies attain asymptotic optimality under much weaker conditions.

Beyond the local-to-global conversion framework, we also contribute to the toolbox of Lyapunov functions, a fundamental technique for proving convergence and concentration in stochastic systems. In particular,

- We develop a new Lyapunov-function technique, which we call *bivariate Lyapunov functions* (Sections 4.5 and 5.5). This technique combines a set of “partial” Lyapunov functions into one, and is useful for analyzing convergence that proceeds in multiple phases.
- In studying fully heterogeneous WCMDPs, we also develop a Lyapunov function for a set of independent but heterogeneous Markov chains (Section 6.4). The Lyapunov function is constructed using a novel *projection-based* idea, which may be useful for more general heterogeneous many-body systems.

Practical impact RBs and WCMDPs underpin applications ranging from wireless scheduling and machine maintenance to online advertising and clinical trials. In these settings, the standard policies are the Whittle index, LP-Priority, and their variants, whose performance guarantees hold only under restrictive assumptions, such as GAP (and indexability, for the Whittle index). Our policies achieve asymptotic optimality without these assumptions, yet remain just as efficient to compute. Any application currently relying on the Whittle index or LP-Priority can therefore switch to ours at no additional computational cost, gaining provable guarantees over a far broader range of instances.

Impact on subsequent work Our work has inspired a number of follow-up papers in RBs and WCMDPs.

- Goldsztajn and Avrachenkov [121], Yan [294] further relax our conditions for asymptotic optimality in RBs established in Chapter 4 [153].
- Inspired by the ideas in Chapter 3 [149], Brown and Zhang [58] designed a policy for the finite-horizon and discounted-reward settings whose optimality gap grows only *linearly* in the (effective) time horizon, scaling better over long horizons than prior policies, whose gaps grow quadratically or even exponentially.

- Wu et al. [290] propose learning algorithms for WCMDPs with unknown parameters and obtain an explicit non-asymptotic optimality-gap bound. Their policies and techniques are heavily inspired by our work.

Theoretical impact Our work can potentially have an impact on the study of many queueing models, thanks to the modeling power of RBs and WCMDPs. Each MDP in an RB or WCMDP can model a job in a queueing system with complex features, such as unknown service times, time-varying or unknown holding cost, or adjustable service rates (as in malleable jobs). This connection has been extensively studied and remains an active research direction [see, e.g., 11, 119, 176, 182, 265]. Niño-Mora [218] provides a comprehensive survey of restless bandits, including their many applications to queueing and scheduling.

We may contribute to this line of research by providing new policy-design ideas and more general performance guarantees. Prior work relies mainly on index or priority policies, which impose a static ranking on arm states and greedily activate the highest-ranked arms; our policies instead actively control the empirical distribution of arm states, ensuring the convergence that prior policies must simply assume. This new design achieves asymptotic optimality under much weaker conditions and, in some cases, yields explicit, non-asymptotic optimality-gap bounds. Such policies and guarantees could be helpful for studying queueing systems.

Some gaps remain. For instance, jobs typically have infinite state spaces (tracking quantities such as remaining size or waiting time), and they arrive dynamically rather than forming a fixed population. Nevertheless, neither of these gaps appears out of reach: our guarantees for heterogeneous WCMDPs (Chapter 6) already avoid an explicit dependence on the size of the state space, which suggests they may extend to continuous state spaces. Existing techniques can also handle dynamically arriving arms [e.g., 266].

Chapter 2

Near-Optimal Stochastic Bin-Packing in Large Service Systems with Time-Varying Item Sizes

2.1 Introduction

2.1.1 Background and motivation

In modern computing systems, a job often takes the form of a virtual machine (VM) or a container [92, 123]. Such a job comes with a resource requirement, such as a certain number of CPUs and amount of memory, while in service. Each server in the system offers a limited amount of these resources. When a job arrives at the system, the job dispatch policy needs to decide which server the job should be assigned to, given the job’s resource requirement and servers’ current job configurations. This job scheduling problem can be approached as a *Stochastic Bin-Packing (SBP) problem*, where jobs are viewed as items, job resource requirements as item sizes, and servers as bins. A traditional SBP setting considers a finite set of jobs that arrive online but do not depart from the system. The objective is to minimize the number of servers that have jobs on them, or ‘non-empty bins’, subject to the resource capacities of the servers. SBP, with a rich history in operations research and

theoretical computer science [65, 66, 68], is a field of continuous developments and advancements [25, 94, 132].

To incorporate job *departures* into the problem formulation, a setting referred to as *stochastic bin-packing in service systems* has been proposed recently [110, 252–256]. In this setting, jobs not only arrive but also depart over time. More specifically, jobs are assumed to arrive according to Poisson processes, and each job is assumed to stay in the system for an exponentially distributed service time. The service time of a job remains unknown until the job departs. Before delving further into SBP in service systems, it is worth mentioning that there is a parallel thread of research on the so-called dynamic bin-packing problem that also handles job departures (see, e.g., [59, 64, 181], and references therein), but it is primarily from a worst-case analysis perspective. Additionally, the virtual machine scheduling problem with objectives different from minimizing the number of active servers has also been widely studied (see, e.g., [194, 196, 197, 225–228, 291]).

For SBP in service systems, the goal is to design a job dispatch policy σ that minimizes the expected number of active servers in *steady state*, denoted as $N(\sigma)$. The *optimality gap* of a policy σ is defined as $N(\sigma) - N(\sigma^*)$, where σ^* is the optimal policy. Since SBP in service systems aims to model today’s large-scale computing systems, the optimality gap of a policy is usually studied in the regime where the total job arrival rate becomes large. As we scale up the total job arrival rate linearly with a scaling factor, r , the optimal value $N(\sigma^*)$ can be shown to be $\Theta(r)$.¹ Therefore, we say a policy is asymptotically optimal if its optimality gap is $o(r)$.

The optimality gap for SBP in service systems has been characterized in the line of work [252–256]. In particular, Stolyar [252], Stolyar and Zhong [254] propose greedy policies that are asymptotically optimal, but the scheduler that executes these policies needs to know detailed state information, which is in a high-dimensional space. Stolyar and Zhong [255, 256] later develop policies that use much less state information and achieve $\Theta(r)$ (with an arbitrarily small constant) and $o(r)$ optimality gaps, respectively.

While prior work on SBP in service systems has provided substantial insights into scheduling virtual-machine-type jobs, it primarily focuses on job resource requirements that remain fixed over time. However, in modern

¹We use the standard Bachmann–Landau notation. Consider two functions $a(r)$ and $b(r)$, where $b(r)$ is positive for large enough r . Then $a = O(b)$ if $\limsup_{r \rightarrow +\infty} \frac{|a(r)|}{b(r)} < \infty$; $a = o(b)$ if $\lim_{r \rightarrow +\infty} \frac{a(r)}{b(r)} = 0$; $a = \Theta(b)$ if $a = O(b)$ and $b = O(a)$.

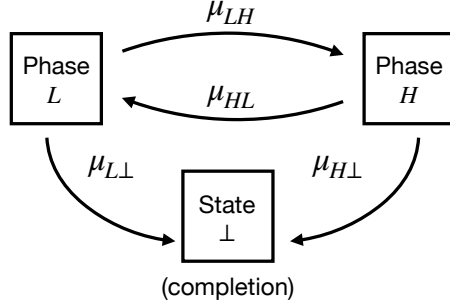
computing systems, jobs' resource requirements often vary over time [28, 80, 190, 234, 240, 258]. For example, when a job involves providing user-facing services, the instantaneous requirement on CPUs and memory depends on the service demand, which is subject to fluctuation over time [80, 190]. Time-varying job resource requirements pose significant challenges in optimizing system efficiency, particularly when aiming to minimize the number of active servers, thereby improving server utilization. It is pertinent to note that low utilization has been recognized as a significant obstacle to the continued scaling of today's computing systems.

Motivated by this gap, in this chapter, we propose a *new setting of stochastic bin-packing in service systems* that allows job resource requirements, or 'item sizes', to vary over time.

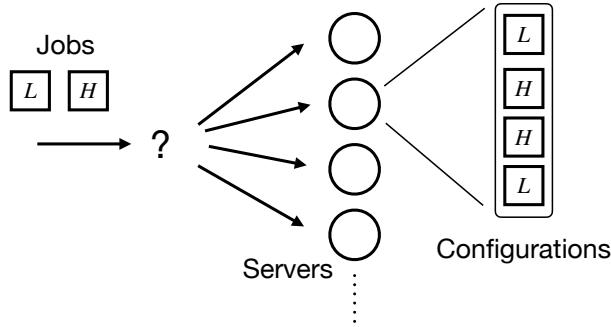
2.1.2 Problem formulation: a simplified version

We first describe our job model that features time-varying resource requirements. For ease of exposition, here we present a simplified setting where each job in service can be in one of the two *phases*, L and H , associated with *low* and *high* resource requirements, respectively. Our full model, presented in Section 2.2, allows *more than two phases and more than one type of resources*. To model the temporal variation in the resource requirement, we assume that each job transitions between the two phases while in service until it is completed, following a continuous-time Markov chain illustrated in Figure 2.1a. We use an absorbing state $\perp$ to denote that the job is completed. A job can initialize in either phase L or phase H , and they are referred to as *type L* and *type H* jobs, respectively. Note that the setting where a job's resource requirement does not vary over time is a special case of our job model where the transition rates between phases are 0.

We consider a system with an infinite number of identical servers, illustrated in Figure 2.1b. We assume jobs arrive according to a Poisson process as existing work on SBP in service systems. In particular, we assume that the two types of jobs arrive at the system following two independent Poisson processes, with rates Λ_L and Λ_H , respectively; i.e., the interarrival times of type L and type H jobs are i.i.d. following exponential distributions with means $1/\Lambda_L$ and $1/\Lambda_H$, respectively. Upon arrival, a job needs to be dispatched to a server according to a *dispatch policy*, and the job enters service immediately. The goal is to design a policy σ to minimize the expected number of *active*



(a) A job alternates between the low- and high-resource phases, L and H , according to a continuous-time Markov chain with transition rates $\mu_{ii'}$. State $\perp$ denotes completion.



(b) Arriving jobs are immediately dispatched to one of infinitely many identical servers. A server configuration records the number of jobs in each phase.

Figure 2.1: Job model and system model.

servers (servers currently serving a positive number of jobs) in steady state, denoted as $N(\sigma)$.

As job resource requirements vary over time, situations can arise where the total job resource requirement on a server exceeds the server's resource capacity, resulting in resource contention. Modern computing systems can tolerate temporary overruns of resource capacity, though they often incur performance degradation or other costs [93, 124]. In our model, we incorporate a rate at which the cost accumulates due to resource contention. We first represent the state of a server by its *configuration*, a vector $\mathbf{k} = (k_L, k_H)$ where k_L and k_H are the numbers of jobs in phase L and phase H , respectively. Then a *cost rate function* $h(\cdot)$ maps a server's configuration to a rate of cost. For example, the cost rate can be proportional to how much the total resource

requirement of the jobs on the server exceeds this server's resource capacity. A more general definition of $h(\cdot)$ is given in Section 2.2. We assume that the resource contention does not affect the transition rates in the job model nor prompt jobs to be terminated, suitable for the application scenarios where the contention level is low and manageable. Let $C(\sigma)$ denote the average expected cost rate per server.

Now our bin-packing problem can be formulated as follows:

$$\begin{aligned} & \underset{\sigma}{\text{minimize}} && N(\sigma) \\ & \text{subject to} && C(\sigma) \leq \epsilon, \end{aligned} \tag{2.1}$$

where $\epsilon > 0$ is a budget for the cost rate of resource contention. We are interested in solving this problem in the asymptotic regime where the arrival rates (Λ_L, Λ_H) scale to infinity [253–256], motivated by the ever-increasing computing demand that drives today's computing systems to be large-scale. Specifically, we assume $(\Lambda_L, \Lambda_H) = (\lambda_L r, \lambda_H r)$ for some fixed coefficients λ_L and λ_H and a scaling factor r , and we study the asymptotic regime where r increases.

2.1.3 Main result

Our main result is an *asymptotically optimal* policy, named JOIN-REQUESTING-SERVER (JRS), for this new setting of SBP in service systems with time-varying job resource requirements. The asymptotic optimality is in the sense that under our proposed policy JRS, the expected number of active servers is at most $(1 + O(r^{-0.5}))$ times the optimal objective value of the optimization problem in (2.1), while the cost rate incurred is at most $(1 + O(r^{-0.5})) \cdot \epsilon$ (i.e., exceeding the budget by at most a diminishing fraction). This asymptotic optimality result translates into an *additive optimality gap* of $O(\sqrt{r})$ in the objective value (expected number of active servers), since the optimal objective value can be shown to be $\Theta(r)$. This main result is formally presented in Theorem 2.1.

Our model can be specialized to the traditional setting of SBP in service systems where jobs' resource requirements remain fixed over time. For this specialization, we replace the constraint $C(\sigma) \leq \epsilon$ in the problem formulation (2.1) with a capacity constraint, which requires the total resource requirement by jobs on a server to be within the server's resource capacity. Our proposed policy JRS can then be adapted into one that has an $O(\sqrt{r})$ optimality gap in

the objective value, which improves upon the state-of-the-art $o(r)$ optimality gap. A discussion on the implementation complexity of JRS and how it compares with existing policies for the traditional setting of SBP in service systems is provided in Section 2.4.3. To be clear, this setting is not a strict special case of the formulation in (2.1) because $\epsilon > 0$ is required there, but our approach and proof carry over.

From a technical approach perspective, our contribution is a novel approach that decomposes the policy design into two steps: defining a single-server sub-problem, and then converting the solution of the sub-problem into a policy in the original problem. This decomposition not only reduces the complexity of policy design but also makes the analysis tractable. We provide an overview of this approach in Section 2.1.4.

2.1.4 Approach overview

To motivate our approach, we ask two questions:

How should we design a good dispatch policy for this system?

How can we prove that a dispatch policy is asymptotically optimal?

Before presenting our answers to these two questions, we first comment on why they are challenging to answer. On the one hand, solving this problem directly via dynamic programming is intractable due to the unbounded state space resulting from the infinitely many servers. Even if we restrict ourselves to the servers that are active, the state space is still prohibitively large. On the other hand, we can consider designing a heuristic policy. However, unlike traditional SBP problems where we can simply seek to pack the servers as compactly as possible, here the question of *how many jobs should be put on a server* is complicated. The complexity comes from the time-varying resource requirements of jobs, which affect future resource contention.

A policy-conversion framework.

We answer the two questions at the same time with a novel *policy-conversion* framework. The framework has two steps:

1. Define the *single-server problem*, which is an easy-to-solve low-dimensional subproblem;
2. Convert the optimal policy of the single-server problem into a policy in the original problem.

This framework allows us to break down the complicated policy design problem into two components. In defining the single-server problem in Step 1, our goal is to quantify the throughput of each server under the resource contention constraint; in the policy conversion in Step 2, our goal is to dispatch jobs optimally based on each single server's characteristics. As we will show, a careful construction of the single-server problem and the conversion procedure naturally leads to a policy for the original system and a proof of its asymptotic optimality. Below we give a quick overview of how we carry out these two steps and the motivation for the design choices.

Single-server problem. To define the single-server problem, consider the following setting: suppose that our goal is to maximize the throughput of *one specific server* while keeping its expected cost rate of resource contention below ϵ , then how should we send jobs to this server? Observe that even though we want to send jobs to the server as frequently as possible, the frequency is fundamentally limited by how fast the server is able to serve jobs and how many jobs can be packed on the server. This motivates us to consider the single-server system illustrated in Figure 2.2. The system has one server and an infinite supply of jobs of all types, so the server can start the service of any number of new jobs of any type at any time. We say the server *requests* a job from the infinite supply whenever it starts serving a new job. We assume the same job model and cost model as in the original infinite-server system. The single-server problem aims to find a job-requesting policy that maximizes the throughput (the number of each type of job that can be served) along the direction of the arrival rate vector $(\Lambda_L, \Lambda_H) = (\lambda_L r, \lambda_H r)$, while maintaining the steady-state expected cost rate of resource contention below ϵ .

How is the single-server problem related to the original problem?

Let $\bar{N}^*$ be the number such that the total throughput of $\bar{N}^*$ single-server systems under the optimal job-requesting policy is equal to $(\lambda_L r, \lambda_H r)$ (assuming $\bar{N}^*$ is an integer for simplicity). Consider the following policy in the original system: let each of the first $\bar{N}^*$ servers in the original system adopt the optimal job-requesting policy and send requests to the dispatcher based on its current configuration. If the requested jobs were to arrive as soon as the dispatcher received the requests, the dispatcher would be able to fulfill the requests immediately. In this case, the first $\bar{N}^*$ servers in the original system would have the same dynamics as $\bar{N}^*$ i.i.d. single-server systems, achieving

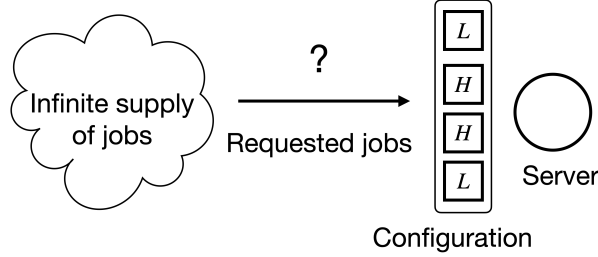


Figure 2.2: A single-server system with an infinite supply of jobs. A single-server policy decides when to request jobs and how many jobs of each type to request.

the largest possible throughput and satisfying the constraint on resources contention. So the original system would have achieved the optimal number of active servers.

However, in the actual model, the dispatcher cannot immediately fulfill a job request because jobs arrive stochastically over time. Nevertheless, the dispatcher can still find a suitable way to match each job arrival with the requests. To see this, note that the time points when the dispatcher receives type i requests result from the superposition of $\bar{N}^* = \Theta(r)$ independent point processes, each with the average rate $\lambda_i r / \bar{N}^*$. As $r \rightarrow \infty$, the instantaneous rates of requesting type i jobs concentrate around the arrival rate $\lambda_i r$ for each i . As a result, most job requests can be fulfilled within a diminishing delay, so most servers in the original system can closely track the optimal single-server dynamics.

A meta-policy, Join-Requesting-Server (JRS), and its asymptotic optimality. Based on the single-server problem and the idea of tracking the optimal single-server dynamics, we propose a meta-policy, JOIN-REQUESTING-SERVER (JRS), which converts a single-server policy $\bar{\sigma}$ to a dispatch policy in the original infinite-server system. We say that JRS takes $\bar{\sigma}$ as a subroutine. The full definition of JRS is given in Section 2.4, along with discussions on various practical considerations in its implementation.

We show that the asymptotic performance of JRS is related to its subroutine in the sense described in Theorem 2.3, which we refer to as the *conversion theorem*. In particular, JRS with the optimal single-server policy (which we refer to as SINGLE-OPT) as the subroutine is asymptotically optimal for the original infinite-server problem as $r \rightarrow \infty$.

In order to track the optimal single-server dynamics, JRS uses a more sophisticated mechanism to control the long-term consequences of missing job requests or fulfilling requests with delays. The mechanism involves the auxiliary variables of *tokens* and *virtual jobs*, which regularize the process of generating requests and matching arrivals with requests. These auxiliary variables play a crucial role in the proof of Theorem 2.3 in Section 2.5, where a novel Stein’s method argument is carried out. We discuss the role of tokens and virtual jobs and their necessity at the end of Section 2.4 and in Section 2.5.5.

Finally, we comment that our policy conversion framework can be applicable to other systems with similar structures. Specifically, we can try to define a suitable single-server problem, solve for its optimal policy, and convert the optimal single-server policy to the original problem. A similar conversion theorem should hold as long as the servers are weakly coupled in some sense. See Section 2.6 for a discussion of such systems.

Relation to the mean-field approach. We remark that the mean-field approach often studies the empirical distribution of configurations on all servers [110, 252–256], which can be viewed as a probability distribution of a single server’s configuration. However, the mean-field approach is typically used to *analyze* this empirical distribution under a given policy for the original system. In contrast, our approach solves the single-server problem to *design* a single-server policy, and then converts it to a policy in the original system with performance guarantee.

2.1.5 Chapter organization

In Section 2.2, we present the general problem formulation, which generalizes the simplified version in this section. In Section 2.3, we give a more detailed overview of our main result and approach, with a short proof of Theorem 2.1 (main result) based on Theorems 2.2–2.4 at the end of the section. Section 2.4 provides a detailed description of our meta-policy, JOIN-REQUESTING-SERVER (JRS), along with discussions on practical considerations in its implementation. In Section 2.5, we prove the performance guarantee of JRS (Theorem 2.3) under an irreducibility assumption. The proof for the general case and other proofs are deferred to the appendices. We conclude the chapter and discuss some future directions in Section 2.6.

2.2 Problem formulation

Job model. As described in Section 2.1, we consider a job model where each job in service can be in one of multiple *phases*, each phase associated with a different resource requirement. Here the resource requirement can be a multi-dimensional vector, with each coordinate specifying the requirement of one type of resource. To model the temporal variation in the resource requirement, we assume that each job transitions between phases while in service until it is completed. The phase transition process is described by a continuous-time Markov chain on the state space $\mathcal{I} \cup \{\perp\}$, where $\mathcal{I}$ is the set of phases and $\perp$ is the absorbing state that denotes the completion of the job. We call a transition between two phases in $\mathcal{I}$ an internal transition, and let $\mu_{ii'}$ denote the transition rate from phase i to phase i' ; the departure of a job then corresponds to a transition from a phase $i \in \mathcal{I}$ to $\perp$, whose transition rate is denoted as $\mu_{i\perp}$. The phase transitions of different jobs are assumed to be independent of each other.

We classify a job as a type i job if it starts from phase $i \in \mathcal{I}$ when entering service. Jobs of each type i arrive to the system according to an independent Poisson process with rate Λ_i .

Server model. We consider an infinite-server system with identical servers. As soon as a job arrives to the system, the job needs to be dispatched to a server to start service immediately. Note that this is always feasible because there are an infinite number of servers in the system. We assume that jobs cannot be preempted or migrated. To describe the state of a server, we define the *configuration* of a server as an $|\mathcal{I}|$ -dimensional vector $\mathbf{k} = (k_i)_{i \in \mathcal{I}}$, whose i -th entry k_i is the number of jobs in phase i on the server. Each server has a limit on the total number of jobs in service at the same time. This limit is denoted as $K_{\max}$ and referred to as the *service limit*. Then the set of feasible server configurations is $\mathcal{K} \triangleq \{\mathbf{k}: \sum_{i \in \mathcal{I}} k_i \leq K_{\max}\}$.

System dynamics. The system state can be represented by the concatenation of the configurations of all servers. Specifically, we index the servers by positive integer numbers and denote the configuration of server ℓ at time t as $\mathbf{K}^\ell(t)$. Then the state of the entire system can be represented by the infinite vector $(\mathbf{K}^\ell(t))_{\ell \in \mathbb{Z}_+}$.

Suppose that the system is in state $(\mathbf{k}^\ell)_{\ell \in \mathbb{Z}_+}$. Let $\mathbf{e}_i$ be an $|\mathcal{I}|$ -dimensional vector whose i -th entry is 1 and all other entries are 0. Then the following state transitions can happen:

- $\mathbf{k}^\ell \rightarrow \mathbf{k}^\ell + \mathbf{e}_i, \mathbf{k}^{\ell'} \rightarrow \mathbf{k}^{\ell'} \forall \ell' \neq \ell$: a type i job arrives and is dispatched to server ℓ ;
- $\mathbf{k}^\ell \rightarrow \mathbf{k}^\ell + \mathbf{e}_{i'} - \mathbf{e}_i, \mathbf{k}^{\ell'} \rightarrow \mathbf{k}^{\ell'} \forall \ell' \neq \ell$: a job on server ℓ transits from phase i to phase i' ;
- $\mathbf{k}^\ell \rightarrow \mathbf{k}^\ell - \mathbf{e}_i, \mathbf{k}^{\ell'} \rightarrow \mathbf{k}^{\ell'} \forall \ell' \neq \ell$: a job on server ℓ departs the system from phase i .

The specifics of the system dynamics depend on the employed *dispatch policy* that decides which server to dispatch to when a job arrives.

Active servers. We are interested in the number of *active servers*, i.e., servers currently serving a positive number of jobs. Note that given the arrival rates of jobs, the smaller the number of active servers, the better the system is utilized. Let $X_{\mathbf{k}}(t)$ be the number of servers in configuration $\mathbf{k}$ at time t , i.e., $X_{\mathbf{k}}(t) = \sum_{\ell=1}^{\infty} \mathbb{1}_{\{\mathbf{K}^\ell(t)=\mathbf{k}\}}$. Then the number of active servers can be written as $\sum_{\mathbf{k} \neq \mathbf{0}} X_{\mathbf{k}}(t)$, where $\mathbf{0} \in \mathbb{R}^{|\mathcal{I}|}$ is the zero vector.

Cost of resource contention. Recall that the cost rate function $h(\cdot)$ maps a server's configuration to a rate of cost. We assume that $h(\cdot)$ is any function that is Γ -Lipschitz continuous with respect to the L^1 distance for some constant $\Gamma > 0$ and satisfies $h(\mathbf{0}) = 0$.

Performance goal. Our high-level goal is to design dispatch policies that minimize the number of active users while keeping the cost rate of resource contention within a certain budget. Specifically, we consider policies that are allowed to be randomized and non-Markovian (i.e., the policies can make history-dependent decisions). We further focus on policies that induce a unique stationary distribution on the configuration process $\{(\mathbf{K}^\ell(t))_{\ell \in \mathbb{Z}_+}\}$, assuming that the configuration process is embedded in a Markov chain that has a unique stationary distribution. We are interested in such policies because the resulting time averages of quantities related to the configurations are equal to the corresponding expectations under the unique stationary distribution regardless of the initial state. Let σ be a policy of interest, $(\mathbf{K}^\ell)_{\ell \in \mathbb{Z}_+}$ be a random element that follows the stationary distribution of the

system state induced by σ , and $X_{\mathbf{k}}$ be the corresponding number of servers in configuration $\mathbf{k}$ in steady state under σ . Then the expected number of active servers is given by

$$N(\sigma) \triangleq \sum_{\mathbf{k} \neq \mathbf{0}} \mathbb{E}[X_{\mathbf{k}}].$$

We define the expected cost rate per expected active server as

$$C(\sigma) \triangleq \frac{\sum_{\mathbf{k} \neq \mathbf{0}} h(\mathbf{k}) \mathbb{E}[X_{\mathbf{k}}]}{\sum_{\mathbf{k} \neq \mathbf{0}} \mathbb{E}[X_{\mathbf{k}}]}.$$

Note that if $C(\sigma) \leq \epsilon$, we have $\sum_{\mathbf{k} \neq \mathbf{0}} h(\mathbf{k}) \mathbb{E}[X_{\mathbf{k}}] \leq \epsilon \sum_{\mathbf{k} \neq \mathbf{0}} \mathbb{E}[X_{\mathbf{k}}]$. Now our goal can be formulated as the following optimization problem, referred to as problem $\mathcal{P}((\Lambda_i)_{i \in \mathcal{I}}, \epsilon)$:

$$\begin{aligned} & \underset{\sigma}{\text{minimize}} && N(\sigma) \\ & \text{subject to} && C(\sigma) \leq \epsilon, \end{aligned} \tag{2.2}$$

where ϵ is a budget for the cost rate of resource contention.

Asymptotic optimality. We focus on the asymptotic regime where for all $i \in \mathcal{I}$, the arrival rate is given by $\Lambda_i = \lambda_i r$ for some constant coefficient λ_i and a positive *scaling factor* $r \rightarrow +\infty$. To define asymptotic optimality, we first define the following notion of approximation to the optimization problem $\mathcal{P}((\Lambda_i)_{i \in \mathcal{I}}, \epsilon)$ in (2.2): a policy σ is said to be (α, β) -optimal if $N(\sigma) \leq \alpha \cdot N^*((\Lambda_i)_{i \in \mathcal{I}}, \epsilon)$ and $C(\sigma) \leq \beta \cdot \epsilon$, where $N^*((\Lambda_i)_{i \in \mathcal{I}}, \epsilon)$ is the optimal objective value in (2.2). Now consider a family of policies $\sigma^{(r)}$ indexed by the scaling factor r . We say that the policy $\sigma^{(r)}$ is *asymptotically optimal* if it is $(\alpha^{(r)}, \beta^{(r)})$ -optimal to the optimization problem $\mathcal{P}((\lambda_i r)_{i \in \mathcal{I}}, \epsilon)$ with $\alpha^{(r)}, \beta^{(r)} \rightarrow 1$ as $r \rightarrow \infty$. We will suppress the superscript (r) for simplicity when it is clear from the context.

We note that under any policy σ , $N(\sigma) = \Theta(r)$. This can be proven using the renowned Little's Law [172] in the following way. The total job arrival rate is $\Theta(r)$ and the expected time that a job spends in the system is $O(1)$. So by Little's Law, the expected number of jobs in the system in steady state is $\Theta(r)$. Since each server can accommodate a constant number of jobs, the expected number of active servers is $\Theta(r)$. Given this, $(\alpha^{(r)}, \beta^{(r)})$ -optimality implies an optimality gap of $\alpha^{(r)} \cdot r$ in the objective value.

2.3 Main result and our approach

2.3.1 Main result

Our main result, Theorem 2.1, is the asymptotic optimality of our proposed policy JOIN-REQUESTING-SERVER (JRS), with a subroutine we call SINGLE-OPT, as briefly discussed in Section 2.1. This asymptotic optimality result implies an $O(\sqrt{r})$ optimality gap in the expected number of active servers. We defer the detailed descriptions of JRS and SINGLE-OPT to Section 2.4 and Appendix 2.C. Theorem 2.1 follows immediately from Theorems 2.2–2.4 to be introduced in Section 2.3.2; a short proof is included at the end of Section 2.3.2 for clarity.

Theorem 2.1 (Asymptotic Optimality). *Consider a stochastic bin-packing problem in service systems with time-varying job resource requirements. Let the arrival rates be $(\lambda_i r)_{i \in \mathcal{I}}$ and the cost rate budget be $\epsilon > 0$. Then the policy JOIN-REQUESTING-SERVER (JRS) with the subroutine SINGLE-OPT is $(1 + O(r^{-0.5}), 1 + O(r^{-0.5}))$ -optimal. That is, the expected number of active servers under JRS with SINGLE-OPT is at most $(1 + O(r^{-0.5}))$ times the optimal value of the problem $\mathcal{P}((\lambda_i r)_{i \in \mathcal{I}}, \epsilon)$, while the cost rate incurred is at most $(1 + O(r^{-0.5})) \cdot \epsilon$.*

Specialization to non-time-varying resource requirements. As mentioned in Section 2.1, we can specialize this result to the setting where the resource requirement of a job does not vary over time. To do that, we remove the cost constraint in $\mathcal{P}((\lambda_i r)_{i \in \mathcal{I}}, \epsilon)$, and redefine the set of feasible server configurations, $\mathcal{K}$, to also incorporate hard capacity constraints for each type of resources. The rest of the analysis is almost identical to that of the analysis for time-varying resource requirements; we omit the details due to the space limit. This specialization results in a policy that is $(1 + O(r^{-0.5}))$ -optimal in the expected number of active servers.

2.3.2 Our approach

In a nutshell, our approach is to reduce the original optimization problem in an infinite-server system to an optimization problem in a single-server system, which is defined below.

A single-server system. Consider a single-server system serving jobs with time-varying resource requirements. The system has *an infinite supply of jobs* of all types. As a result, the server can *request* any number of new jobs of any type at any time. Once a job is requested, it immediately enters service.

We represent the server configuration at time t using a vector $\bar{\mathbf{K}}(t) = (\bar{K}_i(t))_{i \in \mathcal{I}}$, whose i -th entry denotes the number of jobs in phase i . We assume that the single-server system has the same service limit $K_{\max}$ and cost rate function $h(\cdot)$ as a server in the original infinite-server system. Therefore, the server configuration $\bar{\mathbf{K}}(t)$ is also in the set $\mathcal{K} = \{\mathbf{k}: \sum_{i \in \mathcal{I}} k_i \leq K_{\max}\}$, and the cost rate at time t is $h(\bar{\mathbf{K}}(t))$.

A single-server policy $\bar{\sigma}$ determines when and how many jobs of each type to request. We allow the single-server policy to be randomized and assume it is Markovian, i.e., it makes decisions only based on the current configuration. Note that allowing non-Markovian policies will not change the optimal value of the single-server problem that we will consider (see Appendix 2.C). Let $\pi \triangleq (\pi(\mathbf{k}))_{\mathbf{k} \in \mathcal{K}}$ be a stationary distribution of the server configuration under the policy $\bar{\sigma}$, and let $\bar{\mathbf{K}}(\infty)$ be a random variable with the distribution π . When we consider a policy $\bar{\sigma}$ and its stationary distribution π , we assume that the system is initialized from π . The policy $\bar{\sigma}$ together with π defines the *request rate* of type i jobs $\bar{\lambda}_i$, which is the expected number of type i jobs requested per unit time in steady state. Note that $\bar{\lambda}_i$ is the throughput of type i jobs since the system has a finite state space.

We consider the following single-server problem, denoted as $\bar{\mathcal{P}}((\lambda_i r)_{i \in \mathcal{I}}, \epsilon)$:

$$\begin{aligned} & \underset{\bar{N}, \bar{\sigma}, \pi}{\text{minimize}} && \bar{N} \\ & \text{subject to} && \mathbb{E} [h(\bar{\mathbf{K}}(\infty)) | \bar{\mathbf{K}}(\infty) \neq \mathbf{0}] \leq \epsilon, \\ & && \bar{N} \bar{\lambda}_i = \lambda_i r, \quad \forall i \in \mathcal{I}. \end{aligned} \tag{2.3}$$

The single-server problem can be interpreted as follows. We can think of $\bar{N}$ as the number of copies of the single-server system under $\bar{\sigma}$ needed to support the arrival rates $(\lambda_i r)_{i \in \mathcal{I}}$ in the infinite-server system. To minimize $\bar{N}$, it is equivalent to maximizing the throughput $(\bar{\lambda}_i)_{i \in \mathcal{I}}$ in each single-server system, while maintaining their proportions as $(\lambda_i r)_{i \in \mathcal{I}}$.

We remark that for the problem $\bar{\mathcal{P}}((\lambda_i r)_{i \in \mathcal{I}}, \epsilon)$, we only need to consider policies that do not depend on the scaling factor r . To see this, we can replace

the decision variable $\bar{N}$ with $\bar{n} \triangleq \bar{N}/r$ and the optimization problem can be equivalently formulated as follows, which does not involve r :

$$\begin{aligned} & \underset{\bar{n}, \bar{\sigma}, \pi}{\text{minimize}} && \bar{n} \\ & \text{subject to} && \mathbb{E} [h(\bar{\mathbf{K}}(\infty)) | \bar{\mathbf{K}}(\infty) \neq \mathbf{0}] \leq \epsilon, \\ & && \bar{n} \bar{\lambda}_i = \lambda_i, \quad \forall i \in \mathcal{I}. \end{aligned} \tag{2.4}$$

Lower bound. The single-server problem gives a lower bound to the original problem in (2.2) as stated in the following theorem. The proof is given in Appendix 2.A.

Theorem 2.2 (Lower Bound). *Consider a stochastic bin-packing problem in service systems with time-varying job resource requirements. Let the arrival rates be $(\lambda_i r)_{i \in \mathcal{I}}$ and the cost rate budget be $\epsilon > 0$. Let N^* be the optimal value of the original infinite-server problem in (2.2), and let $\bar{N}^*$ be the optimal value of the single-server problem $\bar{\mathcal{P}}((\lambda_i r)_{i \in \mathcal{I}}, \epsilon)$, then $N^* \geq \bar{N}^*$.*

Converting from the single-server system to the infinite-server system. Having established a lower bound on the infinite-server problem $\mathcal{P}((\lambda_i r)_{i \in \mathcal{I}}, \epsilon)$ in terms of the optimal value of the single-server problem $\bar{\mathcal{P}}((\lambda_i r)_{i \in \mathcal{I}}, \epsilon)$, next we focus on finding an asymptotically optimal policy. We will characterize the performance guarantee of a class of policies and then show that the best policy within the class is asymptotically optimal. Specifically, we consider a meta-policy called JOIN-REQUESTING-SERVER (JRS), which converts a Markovian single-server policy $\bar{\sigma}$ into an infinite-server policy. We call the policy resulting from the conversion a *JRS policy with a subroutine $\bar{\sigma}$* . Through analyzing the meta-policy JRS, we show that the performance of each JRS policy can be characterized by the performance of its subroutine, as stated in Theorem 2.3 below. The proof of Theorem 2.3 under an irreducibility assumption is given in Section 2.5, and the proof for the full version is given in Appendix 2.B.3.

Theorem 2.3 (Conversion Theorem). *Consider a stochastic bin-packing problem in service systems with time-varying job resource requirements. Let the arrival rates be $(\lambda_i r)_{i \in \mathcal{I}}$ and the cost rate budget be $\epsilon > 0$. Let $(\bar{N}, \bar{\sigma}, \pi(\mathbf{k}))$ be a solution feasible to the single-server problem $\bar{\mathcal{P}}((\lambda_i r)_{i \in \mathcal{I}}, \epsilon)$. In addition,*

we assume that the policy $\bar{\sigma}$ is Markovian. Let the infinite-server policy σ be JRS with a subroutine $\bar{\sigma}$. Then under σ , we have

$$\left| \sum_{\mathbf{k} \neq \mathbf{0}} \mathbb{E}[X_{\mathbf{k}}] - \lceil \bar{N} \rceil \cdot \mathbb{P}(\bar{\mathbf{K}} \neq \mathbf{0}) \right| = O(\sqrt{r}), \quad (2.5)$$

$$\left| \sum_{\mathbf{k} \neq \mathbf{0}} h(\mathbf{k}) \mathbb{E}[X_{\mathbf{k}}] - \lceil \bar{N} \rceil \cdot \mathbb{E}[h(\bar{\mathbf{K}})] \right| = O(\sqrt{r}). \quad (2.6)$$

As a result,

$$N(\sigma) \leq (1 + O(r^{-0.5})) \cdot \bar{N}, \quad (2.7)$$

$$C(\sigma) \leq (1 + O(r^{-0.5})) \cdot \epsilon. \quad (2.8)$$

Optimal single-server policy. Theorem 2.3 together with the lower bound in Theorem 2.2 reduces the infinite-server problem $\mathcal{P}((\lambda_i r)_{i \in \mathcal{I}}, \epsilon)$ in (2.2) to the single-server problem $\bar{\mathcal{P}}((\lambda_i r)_{i \in \mathcal{I}}, \epsilon)$ in (2.3). We can obtain the optimal single-server policy, SINGLE-OPT, by solving a linear program, as stated in the theorem below.

Theorem 2.4 (Optimality of Single-OPT, Informal). *There exists a linear program $\bar{\mathcal{L}}\mathcal{P}((\lambda_i)_{i \in \mathcal{I}}, \epsilon)$ that is equivalent to the single-server problem $\bar{\mathcal{P}}((\lambda_i r)_{i \in \mathcal{I}}, \epsilon)$. In particular, we can construct an optimal Markovian policy for $\bar{\mathcal{P}}((\lambda_i r)_{i \in \mathcal{I}}, \epsilon)$ from the optimal solution of $\bar{\mathcal{L}}\mathcal{P}((\lambda_i)_{i \in \mathcal{I}}, \epsilon)$.*

Proof of Theorem 2.4 and details on the construction of the optimal policy are given in Appendix 2.C.

Proof of the Theorem 2.1 Based on Theorems 2.2–2.4. Because SINGLE-OPT (along with an optimal stationary distribution) optimally solves $\bar{\mathcal{P}}((\lambda_i r)_{i \in \mathcal{I}}, \epsilon)$, it achieves the optimal value $\bar{N}^*$. Let σ be JRS with a subroutine SINGLE-OPT, then according to Theorem 2.3, we have $N(\sigma) \leq (1 + O(r^{-0.5})) \cdot \bar{N}^*$ and $C(\sigma) \leq (1 + O(r^{-0.5})) \cdot \epsilon$. By Theorem 2.2, we also have $N^* \geq \bar{N}^*$. So we conclude that JRS with a subroutine SINGLE-OPT is $(1 + O(r^{-0.5}), 1 + O(r^{-0.5}))$ -optimal. $\square$

2.4 Proposed meta-policy: Join-Requesting-Server (JRS)

In this section, we describe our meta-policy, JOIN-REQUESTING-SERVER (JRS), in full detail. For ease of presentation, we focus on the case where the subroutine policy $\bar{\sigma}$ for JRS is $\mathbf{k}^0$ -irreducible, i.e., under $\bar{\sigma}$, there exists a configuration $\mathbf{k}^0$ such that the single-server system can return to $\mathbf{k}^0$ from any other configurations (which is equivalent to assuming that the configuration of the single-server system under policy $\bar{\sigma}$ forms a *unichain*). The algorithm for the general case is given in Appendix 2.B.3.

2.4.1 How the single-server policy requests jobs

Before going into the definition of JRS, we first take a closer look at how the Markovian *single-server* policy $\bar{\sigma}$ requests jobs, to avoid potential ambiguity caused by the fact that a single-server policy can request jobs at *any time*. Let a_i denote the number of type i jobs requested, and let $\mathbf{a} \triangleq (a_i)_{i \in \mathcal{I}}$. We say $\mathbf{a}$ is *feasible* if the total number of jobs on the server does not exceed $K_{\max}$ after adding the jobs. The policy $\bar{\sigma}$ performs one of the following two types of requests based on the current configuration.

- **Reactive requests.** A *reactive request* is triggered by *either an internal transition or a departure*. The changes in the configuration when a reactive request is made can be represented by the diagram

$$\mathbf{k} \rightarrow \mathbf{k}' \rightarrow \mathbf{k}' + \mathbf{a},$$

where $\mathbf{k} \rightarrow \mathbf{k}'$ is due to the internal transition or departure, and $\mathbf{k}' \rightarrow \mathbf{k}' + \mathbf{a}$ happens since the policy immediately requests $\mathbf{a}$ jobs. The policy $\bar{\sigma}$ specifies a probability distribution over all feasible $\mathbf{a}$ when it decides to perform reactive requests for the configuration $\mathbf{k}'$.

- **Proactive requests.** A *proactive request* happens on its own, and it happens at a *finite rate* depending on the current configuration of the server. The change of the configuration when a proactive request happens can be represented by the diagram

$$\mathbf{k} \rightarrow \mathbf{k} + \mathbf{a}.$$

More specifically, suppose the policy $\bar{\sigma}$ decides to perform proactive requests for a configuration $\mathbf{k}$. Then for each feasible $\mathbf{a}$, the policy $\bar{\sigma}$ specifies a

rate and runs a timer with an exponentially distributed duration with the specified rate. When the timer ticks, $\mathbf{a}$ jobs are requested. When the configuration changes, all the timers are canceled and restarted with new rates based on the new configuration.

2.4.2 Description of Join-Requesting-Server (JRS)

The inputs of JRS include: (i) the single-server policy $\bar{\sigma}$, (ii) the objective value of $\bar{\sigma}$ in the single-server problem (2.3), denoted as $\bar{N}$, and (iii) the transition rates in the job model.

We first divide the infinite server pool into two sets based on the server index ℓ . Let $L = \lceil \bar{N} \rceil$. We call servers with index $\ell \leq L$ *normal servers*; we call servers with index $\ell > L$ *backup servers*. The normal servers are responsible for serving most of the jobs, while the backup servers are activated only to handle overflow jobs (jobs that are not dispatched to normal servers).

The JRS is specified in two steps.

- **Step 1 (Job Requesting on a Normal Server):** We let each normal server request jobs using its subroutine, the single-server policy $\bar{\sigma}$. The input to the policy $\bar{\sigma}$ is what we refer to as the *observed configuration* of the server, which will be further explained below. When $\bar{\sigma}$ requests $\mathbf{a} = (a_i)_{i \in \mathcal{I}}$ jobs, a_i type i *tokens* are generated for each $i \in \mathcal{I}$ to store the job requests. The server *pauses* the job requesting process if it already has any type of tokens, and resumes when all the tokens that it generated are removed.
- **Step 2 (Arrival Dispatching):**
 - **Real jobs.** When a type i job arrives, the dispatcher chooses a type i token uniformly at random, removes the token, and assigns the job to the corresponding server. When there are no type i tokens, the dispatcher sends the job to an idle backup server.
 - **Virtual jobs.** When the total number of type i tokens throughout the system exceeds the limit $\eta_{\max} = \lceil \sqrt{L} \rceil$ (called the *token limit*), a type i *virtual arrival* is triggered, which causes the dispatcher to choose a type i token uniformly at random, remove the token, and assign a *virtual job* to the corresponding server. A virtual job has the same transition dynamics as a real job but does not consume physical resources.

The *observed configuration* of a normal server in Step 1 is the configuration resulting from real jobs and virtual jobs combined. That is, it is a vector

whose i -th entry represents the total number of real and virtual jobs in phase i on this server. The observed configuration changes when there is a new real or virtual job arrival assigned to the server, or when a real or virtual job on the server has a phase transition or departs. We update the input to the policy $\bar{\sigma}$ when the observed configuration changes. Whenever the observed configuration changes, the policy $\bar{\sigma}$ cancels the exponential timers in progress; but a reactive request from the policy $\bar{\sigma}$ can only be triggered when a real or virtual job on the server has a phase transition or departs.

Intuition behind JRS

To provide a better understanding of the main design ideas of JRS, here we give an intuitive description of how it works. Broadly, servers generate job requests and store unfulfilled requests as tokens; the dispatcher assigns jobs to servers according to the tokens to fulfill job requests. This is the mechanism for matching job arrivals with requests, which is referenced at the end of Section 2.1.4. However, rather than matching all tokens with job arrivals, JRS opts to convert some of the tokens into virtual jobs to keep the total number of tokens within an upper limit $\eta_{\max}$. By capping the number of tokens, JRS ensures that the job requests generated by each server get fulfilled quickly (either by a real job or a virtual job), and thus the observed configurations of servers maintain proximity to i.i.d. copies of the single-server systems.

The choice of the token limit $\eta_{\max} = \Theta(\sqrt{r})$ balances two key considerations. On the one hand, a smaller $\eta_{\max}$ brings the observed configurations closer to i.i.d. copies of single-server systems. On the other hand, if $\eta_{\max}$ is overly small, the rate of generating virtual jobs becomes high and the probability for a job arrival to see no tokens is also high. As a result, the observed configurations, which include both real and virtual jobs, deviate from the real-job configurations. A more in-depth discussion on the role of tokens and virtual jobs and whether they are fundamental is in Section 2.5.5.

2.4.3 Practical considerations in implementing Join-Requesting-Server (JRS)

Computational complexity of JRS

The computational complexity of JRS consists of two components: the *offline* component that computes a single-server policy $\bar{\sigma}$ and its objective value $\bar{N}$, and the *online* component that carries out the two steps of JRS.

The offline component reduces to solving the linear program given in (2.61) in Appendix 2.C.1, whose number of optimization variables is linear in the number of feasible configurations times the number of phases, i.e., $|\mathcal{K}| \times |\mathcal{I}|$, on a single server. Admittedly, $|\mathcal{K}|$ can be large when a single server has a large quantity of resources and there are many job phases. However, we opt for the view that a single server is not excessively large and the system's scale is primarily captured by the scaling factor r . Therefore, it is advantageous that the computational complexity of this offline component is independent of r .

In the online component, the bulk of the computation is in job requesting and virtual job simulation, which can be executed distributedly on the normal servers. Specifically, each normal server monitors its observed configuration and generates tokens according to the single-server policy $\bar{\sigma}$; additionally, when a virtual job is assigned to the server, the server simulates the dynamics of the virtual job, i.e., generates random variables corresponding to phase transitions and job departure. Backup servers do not need to perform any computation beyond serving jobs.

The scheduler, which stores all the tokens, has two responsibilities in the online component: (i) the scheduler matches each newly arrived job to a token of the same type, chosen uniformly at random, or sends the job to a backup server when there are no tokens of the same type; (ii) the scheduler monitors the number of tokens of each type and assigns virtual jobs when the number of tokens exceeds the limit $\eta_{\max}$.

It is informative to compare the computational complexity of JRS with existing algorithms designed for the traditional setting of stochastic bin-packing in service systems, where the resource requirements are non-time-varying [110, 252–256]. At a high level, these existing algorithms function as follows: upon the arrival of a job, the scheduler checks the current configurations of all servers and assigns the job to a server whose configuration optimizes certain predefined criteria. Among these, the GRAND algorithm [253, 255, 256]

stands out for its simplicity and asymptotic optimality. Under GRAND, the scheduler only needs to identify configurations that can accommodate the incoming job and then randomly assigns the job to one of these feasible servers, along with some idle servers. Compared to JRS, GRAND does not have an offline planning component, and individual servers do not perform computation beyond serving jobs. The scheduler’s role in GRAND is slightly more complex than in JRS. Consequently, when considering using JRS in settings where job resource requirements are non-time-varying, practitioners should weigh whether the additional computational complexity is warranted.

Model parameter estimation.

A limitation of JRS is its dependency on known model parameters, including job arrival rates and phase transition rates. Such dependency is not present in existing algorithms designed for the setting with non-time-varying resource requirements. In real-world applications, the model parameters can be estimated from workload traces such as [258, 285]. Estimation errors can impact the system’s performance, an issue that merits further in-depth investigation in future work. Here, we provide a preliminary result on the performance degradation due to parameter estimation errors. Roughly speaking, suppose that the estimation error in the job arrival rate coefficients λ_i ’s and the phase transition rates $\mu_{ii'}$ ’s and $\mu_{i\perp}$ ’s are bounded by $\delta \geq 0$ (along with an insensitivity assumption on the single-server problem). Then if we use JRS where the single-server policy is obtained by solving for the optimal single-server policy under the estimated parameters, the resulting JRS is $(1 + B\delta + O(r^{-0.5}), 1 + B\delta + O(r^{-0.5}))$ -optimal for any $\delta \leq \delta_{\max}$, where B and $\delta_{\max}$ are positive constants independent of r . The exact statement is given in Proposition 2.20 in Appendix 2.D, along with a proof.

Connection to practical algorithms.

Recent progress has been made in addressing the issue of low utilization due to time-varying job resource requirements, notably within Google’s data centers, as discussed in [28]. The approach in [28] makes predictions on the future resource requirements of jobs, which lead to a further prediction on the future peak resource requirement on a server if a newly arrived job were to be sent to that server (assuming no future job arrivals). This prediction categorizes each server as either feasible or infeasible for the new job, and

this binary outcome is subsequently used by a separate scheduler for job assignment.

Our proposed JRS policy can be viewed as giving more detailed predictions on whether it is suitable for a server to take on new jobs, represented by the tokens. The predictions are optimized by taking into account future job arrivals and the stochastic dynamics of jobs.

2.5 Proof of Theorem 2.3 (Conversion Theorem) assuming irreducibility

In this section, we prove Theorem 2.3 to establish the performance guarantee of JRS. For ease of presentation, we focus on the case where the subroutine policy $\bar{\sigma}$ is $\mathbf{k}^0$ -irreducible. The proof for the general case is in Appendix 2.B.3.

This section is organized as follows. We first provide some preliminaries in Section 2.5.1. Then we outline the steps and lemmas needed for the proof in Section 2.5.2. In Section 2.5.3, we prove Theorem 2.3 based on the lemmas. In Section 2.5.4, we prove one of the lemmas, Lemma 2.6, where we devise a novel approach to employ Stein’s method. Finally, in Section 2.5.5, we discuss the role of tokens and virtual jobs and their necessity from a proof perspective. The proofs of the rest of the lemmas presented in this section are given in Appendix 2.B.

2.5.1 Preliminaries

Consider an infinite-server system under the JRS policy. For each normal server ℓ , we describe its status at time t using the following variables: *configuration of real jobs* $\mathbf{K}^\ell(t)$ (referred to simply as configuration in previous sections), *tokens* $\boldsymbol{\eta}^\ell(t)$, *configuration of virtual jobs* $\boldsymbol{\zeta}^\ell(t)$, and *observed configuration* $\widehat{\mathbf{K}}^\ell(t) \triangleq \mathbf{K}^\ell(t) + \boldsymbol{\zeta}^\ell(t)$. We use the superscript “1 : L” to refer to a certain descriptor of all normal servers, for example, $\widehat{\mathbf{K}}^{1:L}(t) \triangleq (\widehat{\mathbf{K}}^\ell(t))_{\ell=1,2,\dots,L}$. The system under JRS is a Markov chain with a unique Markovian representation $((\mathbf{K}^\ell(t))_{\ell=1,2,\dots}, \boldsymbol{\zeta}^{1:L}(t), \boldsymbol{\eta}^{1:L}(t))$. The following lemma shows that the system has a unique stationary distribution (the proof is provided in Appendix 2.B.1).

Lemma 2.5 (Unique Stationary Distribution). *Consider an infinite-server system under the JRS policy with $\bar{\sigma}$ as its subroutine, where $\bar{\sigma}$ is a single-*

server policy that is Markovian and $\mathbf{k}^0$ -irreducible. Then the state of the system $((\mathbf{K}^\ell(t))_{\ell=1,2,\dots}, \boldsymbol{\zeta}^{1:L}(t), \boldsymbol{\eta}^{1:L}(t))$ has a unique stationary distribution.

Let $\bar{\mathbf{K}}^{1:L}(t) \triangleq (\bar{\mathbf{K}}^\ell(t))_{\ell=1,2,\dots,L}$ be the configuration vector of L i.i.d. copies of the single-server system under $\bar{\sigma}$. As discussed in Section 2.4.2, we will show that $\mathbf{K}^{1:L}(\infty)$ can be approximated by $\bar{\mathbf{K}}^{1:L}(\infty)$. In the remainder of this section, we omit the steady-state symbol (∞) for simplicity.

To rigorously discuss the approximation of the steady-state random variables, we define some metrics. Recall that $\mathcal{K} \triangleq \{\mathbf{k}: \sum_{i \in \mathcal{I}} k_i \leq K_{\max}\}$ is the set of feasible single-server configurations. Let $\mathcal{K}^L \triangleq \{\mathbf{k}^{1:L}: \mathbf{k}^\ell \in \mathcal{K}, \forall \ell\}$ be the set of feasible configurations for all normal servers. We use $\|\cdot\|$ to denote the L^1 norm in both space $\mathcal{K}$ and space $\mathcal{K}^L$:

$$\begin{aligned} \|\mathbf{k} - \mathbf{k}'\| &= \sum_{i \in \mathcal{I}} |k_i - k'_i|, \quad \text{for } \mathbf{k}, \mathbf{k}' \in \mathcal{K}, \\ \|\mathbf{k}^{1:L} - \mathbf{k}'^{1:L}\| &= \sum_{\ell=1}^L \|\mathbf{k}^\ell - \mathbf{k}'^\ell\|, \quad \text{for } \mathbf{k}^{1:L}, \mathbf{k}'^{1:L} \in \mathcal{K}^L. \end{aligned}$$

For any two random variables $\mathbf{U}^a, \mathbf{U}^b \in \mathcal{K}^L$, their closeness will be measured in terms of Wasserstein distance as follows:

$$d(\mathbf{U}^a, \mathbf{U}^b) \triangleq \sup_{f \in \text{Lip}(1)} \{ \mathbb{E}[f(\mathbf{U}^a)] - \mathbb{E}[f(\mathbf{U}^b)] \},$$

where the supremum is taken over the all Lipschitz-1 functions from $\mathcal{K}^L$ to $\mathbb{R}$.

2.5.2 Steps and lemmas needed for the proof of Theorem 2.3 assuming irreducibility

Our goal is to show that the steady-state distribution of the normal servers' real-job configurations $\mathbf{K}^{1:L}$ is close to the steady-state distribution of i.i.d. copies of the single-server systems $\bar{\mathbf{K}}^{1:L}$ in Wasserstein distance, and that the backup servers are almost empty as the arrival rate gets large. More formally, we aim to show that $d(\mathbf{K}^{1:L}, \bar{\mathbf{K}}^{1:L}) = O(\sqrt{r})$ and $\sum_{\ell=L+1}^\infty \sum_{i \in \mathcal{I}} K_i^\ell = O(\sqrt{r})$ as $r \rightarrow \infty$. These two bounds provide the performance guarantee claimed in Theorem 2.3.

Instead of directly looking into the distribution of real-job configuration $\mathbf{K}^{1:L}$, we show that the distribution of each of the three sums, $\mathbf{K}^{1:L} + \boldsymbol{\zeta}^{1:L} + \boldsymbol{\eta}^{1:L}$, $\mathbf{K}^{1:L} + \boldsymbol{\zeta}^{1:L}$, and $\mathbf{K}^{1:L}$, can be approximated by the distribution of $\bar{\mathbf{K}}^{1:L}$ in Wasserstein distance. The approximation result for each sum helps us derive the approximation result for the sum with one fewer term. The result that

the backup servers are almost empty also follows from these approximations. This sequence of approximations is illustrated in the figure below, where recall that $\widehat{\mathbf{K}}^\ell(t) \triangleq \mathbf{K}^\ell(t) + \boldsymbol{\zeta}^\ell(t)$ is the observed configuration.

$$\widehat{\mathbf{K}}^{1:L} \longrightarrow \boxed{\mathbf{K}^{1:L} + \boldsymbol{\zeta}^{1:L} + \boldsymbol{\eta}^{1:L}} \approx \overline{\mathbf{K}}^{1:L}$$

A crucial observation that leads to this stepwise proof is that the process $(\widehat{\mathbf{K}}^{1:L}(t), \boldsymbol{\eta}^{1:L}(t))$ forms a Markov chain on its own. This is because real jobs and virtual jobs have the same transition dynamics and are indistinguishable by the subroutine when requesting jobs. Moreover, by the construction of JRS, the Markov chain $(\widehat{\mathbf{K}}^{1:L}(t), \boldsymbol{\eta}^{1:L}(t))$ governs the dynamics of the virtual-job configurations $\boldsymbol{\zeta}^{1:L}(t)$ and the configurations on backup servers.

Our proof consists of two steps. In **Step 1**, we focus on the process $(\widehat{\mathbf{K}}^{1:L}(t), \boldsymbol{\eta}^{1:L}(t))$. We show that $d(\widehat{\mathbf{K}}^{1:L} + \boldsymbol{\eta}^{1:L}, \overline{\mathbf{K}}^{1:L}) = O(\sqrt{r})$, which immediately implies $d(\widehat{\mathbf{K}}^{1:L}, \overline{\mathbf{K}}^{1:L}) = O(\sqrt{r})$ because we have limited the total number of tokens to $O(\sqrt{r})$. In **Step 2**, we use the approximation result for $\widehat{\mathbf{K}}^{1:L}$ in **Step 1** to show that the total number of virtual jobs, $\sum_{i \in \mathcal{I}} \sum_{\ell=1}^L \zeta_i^\ell$, and the total number of jobs on backup servers are both $O(\sqrt{r})$. Recall that $\mathbf{K}^{1:L} = \widehat{\mathbf{K}}^{1:L} - \boldsymbol{\zeta}^{1:L}$, so we get $d(\mathbf{K}^{1:L}, \overline{\mathbf{K}}^{1:L}) = O(\sqrt{r})$.

Next, we state the specific lemmas.

Step 1. Lemma 2.6 below bounds the Wasserstein distance between $\widehat{\mathbf{K}}^{1:L}$ and $\overline{\mathbf{K}}^{1:L}$.

Lemma 2.6. *Under the conditions of Theorem 2.3 and $\bar{\sigma}$ being $\mathbf{k}^0$ -irreducible, we have*

$$d(\widehat{\mathbf{K}}^{1:L}, \overline{\mathbf{K}}^{1:L}) = O(\sqrt{r}).$$

The key challenge for proving Lemma 2.6 is that the job dispatching decisions are based on the configurations of *all normal servers*, which creates dependencies among the transitions of different servers. The key idea that helps us overcome this challenge is to consider the sum $\widehat{\mathbf{K}}^{1:L} + \boldsymbol{\eta}^{1:L}$, which remains unchanged under job arrivals regardless of dispatching decisions. Observe that $\widehat{\mathbf{K}}^{1:L} + \boldsymbol{\eta}^{1:L}$ has *decoupled* dynamics across servers because it is only changed by internal phase transitions, departures, and requests of new jobs, which happen independently on each server. This helps us prove $d(\widehat{\mathbf{K}}^{1:L} + \boldsymbol{\eta}^{1:L}, \overline{\mathbf{K}}^{1:L}) = O(\sqrt{r})$, which implies Lemma 2.6, as argued earlier in the section.

Formally, the proof of Lemma 2.6 makes use of Stein's method (see, e.g., [46, 49, 51]) to compare $\widehat{\mathbf{K}}^{1:L} + \boldsymbol{\eta}^{1:L}$ with $\overline{\mathbf{K}}^{1:L}$. Stein's method usually consists of three steps: generator comparison, Stein factor bound, and moment bound. In our case, due to the finiteness of the state space $\mathcal{K}$, we only need to do the generator comparison and the Stein factor bound. In the generator comparison step, we show that the instantaneous transition rates of $\widehat{\mathbf{K}}^{1:L} + \boldsymbol{\eta}^{1:L}$ match with those of $\overline{\mathbf{K}}^{1:L}$; in the Stein factor bound step, we show that small difference in the transition rates of $\widehat{\mathbf{K}}^{1:L} + \boldsymbol{\eta}^{1:L}$ and $\overline{\mathbf{K}}^{1:L}$ does not cause much increase in the overall distance of the distributions. The detailed proof is in Section 2.5.4.

Step 2. We establish Lemma 2.7 and Lemma 2.8 below, which bound the steady-state expected number of virtual jobs and the jobs on backup servers.

Lemma 2.7. *Under the conditions of Theorem 2.3 and $\bar{\sigma}$ being $\mathbf{k}^0$ -irreducible, for each $i \in \mathcal{I}$, the steady-state expected number of virtual jobs of type i is of the order $O(\sqrt{r})$, i.e.,*

$$\mathbb{E} \left[\sum_{\ell=1}^L \zeta_i^\ell \right] = O(\sqrt{r}).$$

Lemma 2.8. *Under the conditions of Theorem 2.3 and $\bar{\sigma}$ being $\mathbf{k}^0$ -irreducible, for each $i \in \mathcal{I}$, the steady-state expected number of type i jobs on backup servers is of the order $O(\sqrt{r})$, i.e.,*

$$\mathbb{E} \left[\sum_{\ell=L+1}^{\infty} K_i^\ell \right] = O(\sqrt{r}).$$

The key idea for proving Lemma 2.7 and Lemma 2.8 is that by the characterization of $\widehat{\mathbf{K}}^{1:L}$ in Lemma 2.6 and the fact that the job requests are made based on $\widehat{\mathbf{K}}^{1:L}$, we can show that the rate of requesting jobs is approximately equal to the arrival rate for each job type. Therefore, the number of tokens rarely reaches 0 or $\eta_{\max}$. This implies the rarity of virtual jobs and jobs on backup servers. The proofs are provided in Appendix 2.B.2.

2.5.3 Proof of Theorem 2.3 assuming irreducibility based on Lemmas 2.5–2.8.

Proof. First we show that Lemmas 2.6 and 2.7 imply the closeness between $\mathbf{K}^{1:L}$ and $\overline{\mathbf{K}}^{1:L}$. By Lemma 2.6, for any $f \in \text{Lip}(1)$, we have $\mathbb{E}[f(\overline{\mathbf{K}}^{1:L})] -$

$\mathbb{E}[f(\widehat{\mathbf{K}}^{1:L})] = O(\sqrt{r})$. By Lemma 2.7, $\mathbb{E}\left[\sum_{\ell=1}^L \zeta_i^\ell\right] = O(\sqrt{r})$. Recall that $\widehat{\mathbf{K}}^\ell = \mathbf{K}^\ell + \boldsymbol{\zeta}^\ell$, so $\mathbb{E}[f(\widehat{\mathbf{K}}^{1:L})] - \mathbb{E}[f(\mathbf{K}^{1:L})] = O(\sqrt{r})$. Therefore,

$$\mathbb{E}\left[f(\overline{\mathbf{K}}^{1:L})\right] - \mathbb{E}\left[f(\mathbf{K}^{1:L})\right] = O(\sqrt{r}). \quad (2.9)$$

Now we prove (2.5), the bound on the expected number of the active servers, by taking a suitable f in (2.9). Observe that

$$\begin{aligned} \sum_{\mathbf{k} \neq \mathbf{0}} \mathbb{E}[X_{\mathbf{k}}] - L \cdot \mathbb{P}(\overline{\mathbf{K}} \neq \mathbf{0}) &= \mathbb{E}\left[\sum_{\ell=1}^L \mathbb{1}_{\{\mathbf{K}^\ell \neq \mathbf{0}\}}\right] - \mathbb{E}\left[\sum_{\ell=1}^L \mathbb{1}_{\{\overline{\mathbf{K}}^\ell \neq \mathbf{0}\}}\right] \\ &\quad + \mathbb{E}\left[\sum_{\ell=L+1}^{\infty} \mathbb{1}_{\{\mathbf{K}^\ell \neq \mathbf{0}\}}\right], \end{aligned} \quad (2.10)$$

where the last term on RHS is $O(\sqrt{r})$ by Lemma 2.8. To show that the difference between the first two terms on the RHS of (2.10) are also $O(\sqrt{r})$, consider $f_1(\mathbf{k}^{1:L}) \triangleq \sum_{\ell=1}^L \mathbb{1}_{\{\mathbf{k}^\ell \neq \mathbf{0}\}}$. Because

$$\begin{aligned} |f_1(\mathbf{k}^{1:L}) - f_1(\mathbf{k}'^{1:L})| &= \left| \sum_{\ell=1}^L (\mathbb{1}_{\{\mathbf{k}^\ell \neq \mathbf{0}\}} - \mathbb{1}_{\{\mathbf{k}'^\ell \neq \mathbf{0}\}}) \right| \\ &\leq \sum_{\ell=1}^L \mathbb{1}_{\{\mathbf{k}^\ell \neq \mathbf{k}'^\ell\}} \leq \|\mathbf{k}^{1:L} - \mathbf{k}'^{1:L}\|, \end{aligned}$$

for any $\mathbf{k}^{1:L}, \mathbf{k}'^{1:L} \in \mathcal{K}^L$, we have $f_1 \in \text{Lip}(1)$. By (2.9), $\mathbb{E}[f_1(\mathbf{K}^\ell)] - \mathbb{E}\left[\sum_{\ell=1}^L f_1(\overline{\mathbf{K}}^\ell)\right] = O(\sqrt{r})$. Therefore, $\sum_{\mathbf{k} \neq \mathbf{0}} \mathbb{E}[X_{\mathbf{k}}] - L \cdot \mathbb{P}(\overline{\mathbf{K}} \neq \mathbf{0}) = O(\sqrt{r})$. Recall that $L = \lceil N \rceil$, so we get (2.5).

Similarly, to prove (2.6), we observe that

$$\begin{aligned} \sum_{\mathbf{k} \neq \mathbf{0}} h(\mathbf{k}) \mathbb{E}[X_{\mathbf{k}}] - L \cdot \mathbb{E}[h(\overline{\mathbf{K}})] &= \mathbb{E}\left[\sum_{\ell=1}^L h(\mathbf{K}^\ell)\right] - \mathbb{E}\left[\sum_{\ell=1}^L h(\overline{\mathbf{K}}^\ell)\right] \\ &\quad + \mathbb{E}\left[\sum_{\ell=L+1}^{\infty} h(\mathbf{K}^\ell)\right]. \end{aligned} \quad (2.11)$$

The last term of (2.11) can be bounded as

$$\mathbb{E}\left[\sum_{\ell=L+1}^{\infty} h(\mathbf{K}^\ell)\right] \leq \mathbb{E}\left[\sum_{\ell=L+1}^{\infty} \mathbb{1}_{\{\mathbf{K}^\ell \neq \mathbf{0}\}}\right] \cdot \max_{\mathbf{k} \in \mathcal{K}} h(\mathbf{k}),$$

which is $O(\sqrt{r})$ by Lemma 2.8 and the fact that $\mathcal{K}$ is a finite set. To show that the difference between the first two terms on the RHS of (2.11) is also $O(\sqrt{r})$, consider $f_2(\mathbf{k}^{1:L}) = \frac{1}{\Gamma} \sum_{\ell=1}^L h(\mathbf{k}^\ell)$, where Γ is the Lipschitz constant of $h(\cdot)$. Because

$$\begin{aligned} |f_2(\mathbf{k}^{1:L}) - f_2(\mathbf{k}'^{1:L})| &= \frac{1}{\Gamma} \left| \sum_{\ell=1}^L (h(\mathbf{k}^\ell) - h(\mathbf{k}'^\ell)) \right| \\ &\leq \sum_{\ell=1}^L \|\mathbf{k}^\ell - \mathbf{k}'^\ell\| = \|\mathbf{k}^{1:L} - \mathbf{k}'^{1:L}\|, \end{aligned}$$

for any $\mathbf{k}^{1:L}, \mathbf{k}'^{1:L} \in \mathcal{K}^L$, we have $f_2 \in \text{Lip}(1)$. By (2.9), $\mathbb{E}[f_2(\mathbf{K}^\ell)] - \mathbb{E}[\sum_{\ell=1}^L f_2(\bar{\mathbf{K}}^\ell)] = O(\sqrt{r})$. Therefore, $\sum_{\mathbf{k} \neq \mathbf{0}} h(\mathbf{k}) \mathbb{E}[X_{\mathbf{k}}] - L \cdot \mathbb{E}[h(\bar{\mathbf{K}})] = O(\sqrt{r})$. Recall that $L = \lceil \bar{N} \rceil$, so we get (2.6).

To show (2.7) and (2.8), noting that $\lceil \bar{N} \rceil = \Theta(r)$, we have

$$\begin{aligned} N(\sigma) &= \sum_{\mathbf{k} \neq \mathbf{0}} \mathbb{E}[X_{\mathbf{k}}] \leq \lceil \bar{N} \rceil + O(\sqrt{r}) = (1 + O(r^{-0.5})) \cdot \bar{N}, \\ C(\sigma) &= \frac{\sum_{\mathbf{k} \neq \mathbf{0}} h(\mathbf{k}) \mathbb{E}[X_{\mathbf{k}}]}{\sum_{\mathbf{k} \neq \mathbf{0}} \mathbb{E}[X_{\mathbf{k}}]} = \frac{\sum_{\mathbf{k} \neq \mathbf{0}} h(\mathbf{k}) \pi(\mathbf{k}) + O(r^{-0.5})}{1 - \pi(\mathbf{0}) + O(r^{-0.5})} \\ &\leq (1 + O(r^{-0.5})) \cdot \epsilon, \end{aligned}$$

where in the last inequality we have used the fact that $\epsilon > 0$. This completes the proof. $\square$

2.5.4 More details on the system and proof of Lemma 2.6

To bound the distance between $\bar{\mathbf{K}}^{1:L}$ and $\widehat{\mathbf{K}}^{1:L}$, observe that because $f \in \text{Lip}(1)$ and $\sum_{\ell=1}^L \sum_{i \in \mathcal{I}} \eta_i^\ell = O(\sqrt{r})$, it suffices to bound the Wasserstein distance between $\bar{\mathbf{K}}^{1:L}$ and $\widehat{\mathbf{K}}^{1:L} + \boldsymbol{\eta}^{1:L}$, i.e.,

$$\sup_{f \in \text{Lip}(1)} \left\{ \mathbb{E} \left[f(\bar{\mathbf{K}}^{1:L}) \right] - \mathbb{E} \left[f(\widehat{\mathbf{K}}^{1:L} + \boldsymbol{\eta}^{1:L}) \right] \right\} = O(\sqrt{r}), \quad (2.12)$$

where $f(\widehat{\mathbf{K}}^{1:L} + \boldsymbol{\eta}^{1:L})$ is a valid expression because $\widehat{\mathbf{K}}^{1:L} + \boldsymbol{\eta}^{1:L} \in \mathcal{K}$ as discussed in Remark 2.9 below.

More details on system dynamics and generator

To prepare for the proof, we first look into the dynamics of the two systems under study. In particular, we write out the *generators* of $\bar{\mathbf{K}}^{1:L}(t)$ and $(\widehat{\mathbf{K}}^{1:L}(t), \boldsymbol{\eta}^{1:L}(t))$, which are used in the Stein's method arguments.

We first examine the dynamics of the single-server system under Markovian policy $\bar{\sigma}$. Four types of events change a single-server system's configuration: internal transitions, departures, reactive requests, and proactive requests (see Section 2.4.1). The change of configuration due to any event can be represented by the diagram

$$\mathbf{k} \rightarrow \mathbf{k}' \rightarrow \mathbf{k}' + \mathbf{a},$$

where the arrow $\mathbf{k} \rightarrow \mathbf{k}'$ denotes an internal transition or a departure from configuration $\mathbf{k}$ to $\mathbf{k}'$ if $\mathbf{k} \neq \mathbf{k}'$; the arrow $\mathbf{k}' \rightarrow \mathbf{k}' + \mathbf{a}$ denotes a reactive request that adds $\mathbf{a}$ jobs to the system if $\mathbf{k} \neq \mathbf{k}'$, and denotes a proactive request if $\mathbf{k} = \mathbf{k}'$. We call the above change of configuration a *transition*, and denote its rate as $\gamma_{\mathbf{k},(\mathbf{k}',\mathbf{a})}$. Let $E(\mathbf{k})$ denote the set of possible $(\mathbf{k}', \mathbf{a})$ pairs in a transition.

We define the total transition rate at configuration $\mathbf{k}$ as $\gamma_{\mathbf{k}} \triangleq \sum_{(\mathbf{k}',\mathbf{a}) \in E(\mathbf{k})} \gamma_{\mathbf{k},(\mathbf{k}',\mathbf{a})}$, and define the maximal transition rate $\gamma_{\max} = \max_{\mathbf{k} \in \mathcal{K}} \gamma_{\mathbf{k}}$. Since $\mathcal{K}$ is a finite set, we have $\gamma_{\max} < \infty$. Also, observe that the request rate of type i jobs is given by

$$\bar{\lambda}_i \triangleq \sum_{\mathbf{k}} \sum_{(\mathbf{k}',\mathbf{a}) \in E(\mathbf{k})} \gamma_{\mathbf{k},(\mathbf{k}',\mathbf{a})} a_i \cdot \pi(\mathbf{k}), \quad (2.13)$$

where π denotes the stationary distribution of single-server configuration under policy $\bar{\sigma}$.

Next, we focus on the dynamics of L i.i.d. copies of single-server systems. Consider the generator $\bar{G}$ of the corresponding Markov chain $\{\bar{\mathbf{K}}^{1:L}(t)\}$, which is a linear operator on functions $g: \mathcal{K}^L \rightarrow \mathbb{R}$ defined as:

$$\bar{G}g(\mathbf{k}^{1:L}) \triangleq \frac{d}{dt} \mathbb{E} \left[g \left(\bar{\mathbf{K}}^{1:L}(t) \right) \middle| \bar{\mathbf{K}}^{1:L}(0) = \mathbf{k}^{1:L} \right] \Big|_{t=0}, \quad (2.14)$$

and we call the resulting function $\bar{G}g(\cdot)$ the *drift* of $g(\cdot)$. Based on the transition rates defined above, we have

$$\bar{G}g(\mathbf{k}^{1:L}) = \sum_{\ell=1}^L \sum_{(\mathbf{k}',\mathbf{a}) \in E(\mathbf{k}^\ell)} \gamma_{\mathbf{k}^\ell,(\mathbf{k}',\mathbf{a})} \left(g(\cdot, \mathbf{k}' + \mathbf{a}, \cdot) - g(\cdot, \mathbf{k}^\ell, \cdot) \right), \quad (2.15)$$

where $g(\cdot, \mathbf{k}' + \mathbf{a}, \cdot) - g(\cdot, \mathbf{k}^\ell, \cdot)$ is a shorthand for $g(\mathbf{k}^1, \dots, \mathbf{k}^{\ell-1}, \mathbf{k}' + \mathbf{a}, \mathbf{k}^{\ell+1}, \dots, \mathbf{k}^L) - g(\mathbf{k}^{1:L})$, i.e., we use $\cdot$ to omit the entries that agree with $\mathbf{k}^{1:L}$.

Similarly, for the infinite-server system, consider the generator $\widehat{G}$ of $(\widehat{\mathbf{K}}^{1:L}(t), \boldsymbol{\eta}^{1:L}(t))$ defined as

$$\begin{aligned} & \widehat{G}\psi(\mathbf{k}^{1:L}, \boldsymbol{\eta}^{1:L}) \\ & \triangleq \frac{d}{dt} \mathbb{E} \left[\psi \left(\widehat{\mathbf{K}}^{1:L}(t), \boldsymbol{\eta}^{1:L}(t) \right) \middle| \widehat{\mathbf{K}}^{1:L}(0) = \mathbf{k}^{1:L}, \mathbf{K}^{1:L}(0) = \boldsymbol{\eta}^{1:L} \right] \Big|_{t=0}, \end{aligned} \quad (2.16)$$

for any function $\psi: (\mathcal{K} \times \mathcal{K})^L \rightarrow \mathbb{R}$. The drift of ψ under $\widehat{G}$ turns out to have a similar decoupled form as $\overline{G}g$: observe that for each ℓ , the transition of $(\widehat{\mathbf{K}}^\ell(t), \boldsymbol{\eta}^\ell(t))$ from $(\mathbf{k}, \boldsymbol{\eta})$ to $(\mathbf{k}', \boldsymbol{\eta} + \mathbf{a} \mathbb{1}_{\{\boldsymbol{\eta}=\mathbf{0}\}})$ occurs at the rate $\gamma_{\mathbf{k}, (\mathbf{k}', \mathbf{a})}$ for each $(\mathbf{k}', \mathbf{a}) \in E(\mathbf{k})$, and any real or virtual job arrivals do not change the sum $\widehat{\mathbf{K}}^\ell(t) + \boldsymbol{\eta}^\ell(t)$. Consider any function $g: \mathcal{K}^L \rightarrow \mathbb{R}$ and the function $\psi(\mathbf{k}^{1:L}, \boldsymbol{\eta}^{1:L}) = g(\mathbf{k}^{1:L} + \boldsymbol{\eta}^{1:L})$.

$$\begin{aligned} \widehat{G}\psi(\mathbf{k}^{1:L}, \boldsymbol{\eta}^{1:L}) &= \sum_{\ell=1}^L \sum_{(\mathbf{k}', \mathbf{a}) \in E(\mathbf{k}^\ell)} \gamma_{\mathbf{k}^\ell, (\mathbf{k}', \mathbf{a})} (g(\cdot, \mathbf{k}' + \mathbf{a}, \cdot) - g(\cdot, \mathbf{k}^\ell, \cdot)) \mathbb{1}_{\{\boldsymbol{\eta}^\ell=\mathbf{0}\}} \\ &+ \sum_{\ell=1}^L \sum_{(\mathbf{k}', \mathbf{a}) \in E(\mathbf{k}^\ell)} \gamma_{\mathbf{k}^\ell, (\mathbf{k}', \mathbf{a})} \\ &\quad \cdot (g(\cdot, \mathbf{k}' + \boldsymbol{\eta}^\ell, \cdot) - g(\cdot, \mathbf{k}^\ell + \boldsymbol{\eta}^\ell, \cdot)) \mathbb{1}_{\{\boldsymbol{\eta}^\ell \neq \mathbf{0}\}}. \end{aligned} \quad (2.17)$$

In this context $g(\cdot, \mathbf{k}' + \boldsymbol{\eta}^\ell, \cdot) - g(\cdot, \mathbf{k}^\ell + \boldsymbol{\eta}^\ell, \cdot)$ is a shorthand for $g(\mathbf{k}^1 + \boldsymbol{\eta}^1, \dots, \mathbf{k}^{\ell-1} + \boldsymbol{\eta}^{\ell-1}, \mathbf{k}' + \boldsymbol{\eta}^\ell, \mathbf{k}^{\ell+1} + \boldsymbol{\eta}^{\ell+1}, \dots, \mathbf{k}^L + \boldsymbol{\eta}^L) - g(\mathbf{k}^{1:L} + \boldsymbol{\eta}^{1:L})$. In other words, we use $\cdot$ to omit the entries of g 's input that agree with the corresponding entries of $\mathbf{k}^{1:L} + \boldsymbol{\eta}^{1:L}$.

Remark 2.9. In (2.17), although g is only defined on the domain $\mathcal{K}^L$, it is valid to write $\mathbf{k}^{1:L} + \boldsymbol{\eta}^{1:L}$ as its input because we always have $\widehat{\mathbf{K}}^\ell + \boldsymbol{\eta}^\ell \in \mathcal{K}$, i.e., the total number of real jobs, virtual jobs, and tokens on a normal server never exceeds $K_{\max}$. To see why this is true, the single-server policy $\bar{\sigma}$ requests jobs only when there are no tokens on the server, and it will not request more than $K_{\max} - n$ jobs if there are already n real and virtual jobs on the server.

Proof of Lemma 2.6.

Proof. Generator Comparison. For any $f \in \text{Lip}(1)$, consider the Poisson equation (see, e.g., [46]) that solves for $g_f: \mathcal{K}^L \rightarrow \mathbb{R}$:

$$\mathbb{E} \left[f(\bar{\mathbf{K}}^{1:L}) \right] - f(\mathbf{k}^{1:L}) = \bar{G}g_f(\mathbf{k}^{1:L}). \quad (2.18)$$

We let $\mathbf{k}^{1:L} = \widehat{\mathbf{K}}^{1:L} + \boldsymbol{\eta}^{1:L}$ in (2.18) and take the expectation. This results in

$$\mathbb{E} \left[f(\bar{\mathbf{K}}^{1:L}) \right] - \mathbb{E} \left[f(\widehat{\mathbf{K}}^{1:L} + \boldsymbol{\eta}^{1:L}) \right] = \mathbb{E} \left[\bar{G}g_f(\widehat{\mathbf{K}}^{1:L} + \boldsymbol{\eta}^{1:L}) \right]. \quad (2.19)$$

On the other hand, because $(\widehat{\mathbf{K}}^{1:L}(t), \boldsymbol{\eta}^{1:L}(t))$ is a finite-state Markov chain, we have

$$\mathbb{E} \left[\widehat{G}\psi_f(\widehat{\mathbf{K}}^{1:L}, \boldsymbol{\eta}^{1:L}) \right] = 0, \quad (2.20)$$

where ψ_f is given by $\psi_f(\widehat{\mathbf{K}}^{1:L}, \boldsymbol{\eta}^{1:L}) = g_f(\widehat{\mathbf{K}}^{1:L} + \boldsymbol{\eta}^{1:L})$. Subtracting (2.20) from (2.19), we get

$$\begin{aligned} & \mathbb{E} \left[f(\bar{\mathbf{K}}^{1:L}) \right] - \mathbb{E} \left[f(\widehat{\mathbf{K}}^{1:L} + \boldsymbol{\eta}^{1:L}) \right] \\ &= \mathbb{E} \left[\bar{G}g_f(\widehat{\mathbf{K}}^{1:L} + \boldsymbol{\eta}^{1:L}) - \widehat{G}\psi_f(\widehat{\mathbf{K}}^{1:L}, \boldsymbol{\eta}^{1:L}) \right]. \end{aligned} \quad (2.21)$$

We want to show that $\bar{G}$ and $\widehat{G}$ are close so that we can bound the RHS of (2.21).

Now we plug the formula of the generators in (2.15) and (2.17) into the RHS of (2.21) and get

$$\begin{aligned} & \left| \bar{G}g_f(\widehat{\mathbf{K}}^{1:L} + \boldsymbol{\eta}^{1:L}) - \widehat{G}\psi_f(\widehat{\mathbf{K}}^{1:L}, \boldsymbol{\eta}^{1:L}) \right| \\ & \stackrel{(i)}{=} \left| \sum_{\ell=1}^L \sum_{(\mathbf{k}', \mathbf{a}) \in E(\widehat{\mathbf{K}}^\ell + \boldsymbol{\eta}^\ell)} \gamma_{\widehat{\mathbf{K}}^\ell + \boldsymbol{\eta}^\ell, (\mathbf{k}', \mathbf{a})} \right. \\ & \quad \cdot \left(g_f(\cdot, \mathbf{k}' + \mathbf{a}, \cdot) - g_f(\cdot, \widehat{\mathbf{K}}^\ell + \boldsymbol{\eta}^\ell, \cdot) \right) \cdot \mathbb{1}_{\{\boldsymbol{\eta}^\ell \neq \mathbf{0}\}} \\ & \quad - \sum_{\ell=1}^L \sum_{(\mathbf{k}', \mathbf{a}) \in E(\widehat{\mathbf{K}}^\ell)} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} \\ & \quad \cdot \left(g_f(\cdot, \mathbf{k}' + \boldsymbol{\eta}^\ell, \cdot) - g_f(\cdot, \widehat{\mathbf{K}}^\ell + \boldsymbol{\eta}^\ell, \cdot) \right) \cdot \mathbb{1}_{\{\boldsymbol{\eta}^\ell \neq \mathbf{0}\}} \left. \right| \end{aligned}$$

$$\begin{aligned}
 & \stackrel{\text{(ii)}}{\leq} \sum_{\ell=1}^L \gamma_{\widehat{\mathbf{K}}^\ell + \boldsymbol{\eta}^\ell} \cdot \sup_{(\mathbf{k}', \mathbf{a}) \in E(\widehat{\mathbf{K}}^\ell + \boldsymbol{\eta}^\ell)} \left| g_f(\cdot, \mathbf{k}' + \mathbf{a}, \cdot) - g_f(\cdot, \widehat{\mathbf{K}}^\ell + \boldsymbol{\eta}^\ell, \cdot) \right| \cdot \mathbb{1}_{\{\boldsymbol{\eta}^\ell \neq \mathbf{0}\}} \\
 & \quad + \sum_{\ell=1}^L \gamma_{\widehat{\mathbf{K}}^\ell} \cdot \sup_{(\mathbf{k}', \mathbf{a}) \in E(\widehat{\mathbf{K}}^\ell)} \left| g_f(\cdot, \mathbf{k}' + \boldsymbol{\eta}^\ell, \cdot) - g_f(\cdot, \widehat{\mathbf{K}}^\ell + \boldsymbol{\eta}^\ell, \cdot) \right| \cdot \mathbb{1}_{\{\boldsymbol{\eta}^\ell \neq \mathbf{0}\}} \\
 & \stackrel{\text{(iii)}}{\leq} 2\gamma_{\max} \cdot \sum_{\ell=1}^L \sup_{\mathbf{k}' \in \mathcal{K}} \left| g_f(\cdot, \mathbf{k}', \cdot) - g_f(\cdot, \widehat{\mathbf{K}}^\ell + \boldsymbol{\eta}^\ell, \cdot) \right| \cdot \mathbb{1}_{\{\boldsymbol{\eta}^\ell \neq \mathbf{0}\}} \\
 & \leq 2\gamma_{\max} \cdot \sup_{\mathbf{k}, \mathbf{k}' \in \mathcal{K}} \left| g_f(\cdot, \mathbf{k}', \cdot) - g_f(\cdot, \mathbf{k}, \cdot) \right| \cdot \sum_{\ell=1}^L \mathbb{1}_{\{\boldsymbol{\eta}^\ell \neq \mathbf{0}\}}, \tag{2.22}
 \end{aligned}$$

where in $g_f(\cdot, \mathbf{k}', \cdot) - g_f(\cdot, \mathbf{k}, \cdot)$ we have omitted the entries that agree with $\mathbf{k}^{1:L}$. The equality (i) is true because each of the ℓ -th terms in $\overline{G}$ and $\widehat{G}$ is equal if $\boldsymbol{\eta}^\ell = \mathbf{0}$. For the inequalities (ii) and (iii), recall that $\gamma_{\mathbf{k}}$ is the total transition rate given by $\gamma_{\mathbf{k}} \triangleq \sum_{(\mathbf{k}', \mathbf{a}) \in E(\mathbf{k})} \gamma_{\mathbf{k}, (\mathbf{k}', \mathbf{a})}$, and $\gamma_{\max} = \max_{\mathbf{k} \in \mathcal{K}} \gamma_{\mathbf{k}}$. Observe that

$$\sum_{\ell=1}^L \mathbb{1}_{\{\boldsymbol{\eta}^\ell \neq \mathbf{0}\}} \leq \sum_{\ell=1}^L \sum_{i \in \mathcal{I}} \eta_i^\ell \leq |\mathcal{I}| \cdot \eta_{\max} = O(\sqrt{r}).$$

Therefore (2.22) can be further bounded by

$$\begin{aligned}
 & \left| \overline{G} g_f(\widehat{\mathbf{K}}^{1:L} + \boldsymbol{\eta}^{1:L}) - \widehat{G} \psi_f(\widehat{\mathbf{K}}^{1:L}, \boldsymbol{\eta}^{1:L}) \right| \\
 & \leq 2\gamma_{\max} \cdot \sup_{\mathbf{k}, \mathbf{k}' \in \mathcal{K}} \left| g_f(\cdot, \mathbf{k}', \cdot) - g_f(\cdot, \mathbf{k}, \cdot) \right| \cdot \sum_{\ell=1}^L \sum_{i \in \mathcal{I}} \eta_i^\ell \\
 & \leq 2\gamma_{\max} \cdot \sup_{\mathbf{k}, \mathbf{k}' \in \mathcal{K}} \left| g_f(\cdot, \mathbf{k}', \cdot) - g_f(\cdot, \mathbf{k}, \cdot) \right| \cdot O(\sqrt{r}).
 \end{aligned}$$

To prove (2.12), it remains to show that

$$\sup_{\mathbf{k}, \mathbf{k}' \in \mathcal{K}} \left| g_f(\cdot, \mathbf{k}', \cdot) - g_f(\cdot, \mathbf{k}, \cdot) \right| = O(1). \tag{2.23}$$

Stein Factor Bound. To prove (2.23), observe that the following $g_f(\cdot)$ is a solution to the Poisson equation (2.18):

$$g_f(\mathbf{k}^{1:L}) = \int_0^\infty \mathbb{E} \left[\left(f(\overline{\mathbf{K}}^{1:L}(t)) - \mathbb{E} \left[f(\overline{\mathbf{K}}^{1:L}) \right] \right) \middle| \overline{\mathbf{K}}^{1:L}(0) = \mathbf{k}^{1:L} \right] dt. \tag{2.24}$$

This allows us to bound the difference of g_f using coupling. Specifically, we define the coupling of two systems, each consisting of L i.i.d. copies of the single-server system under $\bar{\sigma}$. The two systems are initialized with configurations $(\cdot, \mathbf{k}', \cdot)$ and $(\cdot, \mathbf{k}, \cdot)$ that only differ at the ℓ -th server, where we omit the entries that agree with $\mathbf{k}^{1:L}$. Let $(\bar{\mathbf{K}}^{1:L,1}(t), \bar{\mathbf{K}}^{1:L,2}(t))$ be the joint configuration of the two systems, which is actually $2L$ i.i.d. copies of the single-server system. As a result, we can specify the couplings $(\bar{\mathbf{K}}^{\ell',1}(t), \bar{\mathbf{K}}^{\ell',2}(t))$ for different ℓ' separately. For $\ell' \neq \ell$, the corresponding server in the two systems have the same initial configurations, so we can always keep their configurations identical. For the ℓ -th servers, we let them evolve independently following their own dynamics until a stopping time τ_{mix} when their configurations become the same. After that, we can use coupling to keep their configurations identical. Under this coupling, it is not hard to see that

$$\begin{aligned}
|g_f(\cdot, \mathbf{k}', \cdot) - g_f(\cdot, \mathbf{k}, \cdot)| &= \left| \int_0^\infty \mathbb{E} \left[f(\bar{\mathbf{K}}^{1:L,1}(t)) - f(\bar{\mathbf{K}}^{1:L,2}(t)) \right] dt \right| \\
&\leq \mathbb{E} \left[\int_0^\infty \left| f(\bar{\mathbf{K}}^{1:L,1}(t)) - f(\bar{\mathbf{K}}^{1:L,2}(t)) \right| dt \right] \\
&\leq \mathbb{E} \left[\int_0^\infty \sum_{\ell'=1}^L \|\bar{\mathbf{K}}^{\ell',1}(t) - \bar{\mathbf{K}}^{\ell',2}(t)\| dt \right] \\
&= \mathbb{E} \left[\int_0^{\tau_{\text{mix}}} \|\bar{\mathbf{K}}^{\ell,1}(t) - \bar{\mathbf{K}}^{\ell,2}(t)\| dt \right], \tag{2.25}
\end{aligned}$$

where in the second inequality we have used the fact that f is 1-Lipschitz continuous under the L^1 norm of the space $\mathcal{K}^L$. For each pair of $\mathbf{k}, \mathbf{k}'$, observe that because $\bar{\sigma}$ is a $\mathbf{k}^0$ -irreducible policy, $\mathbb{E}[\tau_{\text{mix}}]$ is finite; and because $\mathcal{K}$ is a finite set, $\|\bar{\mathbf{K}}^{\ell,1}(t) - \bar{\mathbf{K}}^{\ell,2}(t)\|$ is uniformly bounded. All these finite quantities depend on a single-server system under a policy $\bar{\sigma}$ that is independent of r . As a result, the last expression in (2.25) is of constant order. Moreover, because there are finite pairs of $(\mathbf{k}, \mathbf{k}')$, the supremum $\sup_{\mathbf{k}, \mathbf{k}'} \mathbb{E} \left[\int_0^{\tau_{\text{mix}}} \|\bar{\mathbf{K}}^{\ell,1}(t) - \bar{\mathbf{K}}^{\ell,2}(t)\| dt \right]$ is also of constant order, independent of r . This proves the Stein factor bound in (2.23). Together with the generator comparison, we have proved

$$\sup_{f \in \text{Lip}(1)} \mathbb{E} \left[f(\bar{\mathbf{K}}^{1:L}) \right] - \mathbb{E} \left[f(\widehat{\mathbf{K}}^{1:L} + \boldsymbol{\eta}^{1:L}) \right] = O(\sqrt{r}). \tag{2.12}$$

Because $\sum_{\ell=1}^L \sum_{i \in \mathcal{I}} \eta_i^\ell \leq |\mathcal{I}| \eta_{\max} = O(\sqrt{r})$, for any $f \in \text{Lip}(1)$, we have

$$\left| \mathbb{E} \left[f(\widehat{\mathbf{K}}^{1:L}) \right] - \mathbb{E} \left[f(\widehat{\mathbf{K}}^{1:L} + \boldsymbol{\eta}^{1:L}) \right] \right| \leq \mathbb{E} \left[\sum_{\ell=1}^L \sum_{i \in \mathcal{I}} \eta_i^\ell(t) \right] = O(\sqrt{r}). \quad (2.26)$$

Plugging the above equation to (2.12), we get $\sup_{f \in \text{Lip}(1)} \mathbb{E} \left[f(\overline{\mathbf{K}}^{1:L}) \right] - \mathbb{E} \left[f(\widehat{\mathbf{K}}^{1:L}) \right] = O(\sqrt{r})$. This proves Lemma 2.6. $\square$

2.5.5 Role of tokens and virtual jobs

This section aims to shed light on the role of tokens and virtual jobs in the proposed policy, JRS. We first outline how we devise the token-and-virtual-job mechanism from the perspective of generators. To begin with, consider the scenario where dispatch decisions are solely based on real-job configurations $(\mathbf{K}^\ell)_{\ell=1,2,\dots}$. In this case, the transitions of servers' configurations would be correlated in general due to job arrivals, which are dispatched based on the joint configuration of all servers. To break this correlation, let us introduce tokens but not set an upper limit yet on the number of tokens (thus no virtual jobs). Observe that $\mathbf{K}^{1:L}(t) + \boldsymbol{\eta}^{1:L}(t)$ remains unchanged by job arrivals, so tokens remove the correlation brought about by job arrivals. However, because tokens lack internal phase transitions or departures, the transition dynamics of $\mathbf{K}^{1:L}(t) + \boldsymbol{\eta}^{1:L}(t)$ will diverge from $\overline{\mathbf{K}}^{1:L}(t)$ when $\boldsymbol{\eta}(t)$ is large. In other words, although tokens help decouple the transitions on servers, they cannot keep the transitions of $\mathbf{K}^{1:L}(t) + \boldsymbol{\eta}^{1:L}(t)$ close to $\overline{\mathbf{K}}^{1:L}(t)$. To solve this issue, we finally introduce the mechanism of converting tokens into virtual jobs when the number of tokens is high, where virtual jobs can make internal phase transitions or departures just like real jobs. Now, the sum $\mathbf{K}^{1:L}(t) + \boldsymbol{\zeta}^{1:L}(t) + \boldsymbol{\eta}^{1:L}(t)$ remains unchanged by job arrivals nor the creation of virtual jobs, and the internal phase transitions and job departures are similar to those of $\overline{\mathbf{K}}^{1:L}(t)$. More formally, the generators of $\mathbf{K}^{1:L}(t) + \boldsymbol{\zeta}^{1:L}(t) + \boldsymbol{\eta}^{1:L}(t)$ and $\overline{\mathbf{K}}^{1:L}(t)$ are close to each other – their additive difference can be upper bounded by a quantity proportional to the expected number of tokens, as shown in (2.22). Therefore, by regulating the number of tokens, we can control the difference between the generators of $\mathbf{K}^{1:L}(t) + \boldsymbol{\zeta}^{1:L}(t) + \boldsymbol{\eta}^{1:L}(t)$ and $\overline{\mathbf{K}}^{1:L}(t)$.

Another key design component of JRS is that the subroutine requests jobs based on the observed configurations. This has been used in the proof of

Lemma 2.6 to show that the observed configurations $\mathbf{K}^{1:L} + \boldsymbol{\zeta}^{1:L}$ are close to $\overline{\mathbf{K}}^{1:L}$, which consist of L i.i.d. single-server systems. Recall that each single-server system in $\overline{\mathbf{K}}^{1:L}$ is designed to have a throughput of $\lambda_i r / L$ for each job type $i \in \mathcal{I}$. Therefore, the proximity between $\mathbf{K}^{1:L} + \boldsymbol{\zeta}^{1:L}$ and $\overline{\mathbf{K}}^{1:L}$ ensures that the job request rate mirrors the arrival rate for each job type, regardless of the real-job configurations. The fact that these two rates are approximately equal is important for proving Lemma 2.7 and Lemma 2.8. It guarantees that both the rate of generating virtual jobs (when there are too many tokens) and the rate of dispatching jobs to backup servers (when there are no tokens) are appropriate.

A natural follow-up question is whether the usage of tokens and virtual jobs is fundamental or an artifact of our analysis technique. For example, it is unclear whether removing the upper limit on the number of tokens would still yield an asymptotically optimal policy. This is an interesting question that we do not have a complete answer to. The token-and-virtual-job mechanism emerges as a natural choice under our analysis framework. Nevertheless, it is worth noting that our analysis primarily treats each server's configuration as a generic Markov chain, without utilizing many properties specific to the stochastic bin-packing setting. An exception to this is the proofs in Appendix 2.B.2, where we use the model that each job leaves the system within a constant expected time. It would be interesting to explore more properties of the problem to better understand policy designs without auxiliary state variables like tokens and virtual jobs.

2.6 Conclusion

In this chapter, we study a new setting of stochastic bin-packing in service systems that features time-varying item sizes. Since our formulation is motivated by the problem of virtual-machine scheduling in computing systems, we use the terminology of jobs and servers, where jobs are viewed as items, whose sizes are their resource requirements, and servers as bins. The time-varying item sizes capture the emerging trend in practice that jobs' resource requirements vary over time. Our goal is to design a job dispatch policy to minimize the expected number of active servers in steady state, subject to a constraint on resource contentions. Our main result is the design of a policy that achieves an optimality gap of $O(\sqrt{r})$, where r is the scaling factor of the arrival rate. When specialized to the setting where jobs' resource

requirements remain fixed over time, this result improves upon the state-of-the-art $o(r)$ optimality gap. Our technical approach highlights a novel policy conversion framework, JOIN-REQUESTING-SERVER, that reduces the policy design problem to that in a single-server system.

There are several potential directions that may be worth further exploration. One direction is to strengthen the optimality result within the current setting. Specifically, it is interesting to investigate: (i) whether it is possible to achieve an optimality gap smaller than $\Theta(\sqrt{r})$; and (ii) whether there exist asymptotically optimal policies whose average cost rate of resource contention satisfies the budget strictly instead of asymptotically.

We are also interested in extending our technique to the optimal control of other systems with similar structures. Intuitively, this technique could be applied to systems with many components that evolve mostly independently but are weakly coupled by certain constraints. Viewing each component as a server, we can define a suitable single-server problem and then design a policy for the original system to track the dynamics of the optimal single-server solution. Below we list several variations of our model that can potentially be analyzed using the proposed technique.

- A model where jobs running on each server will be put into a local queue when there are resource contentions. The goal thus becomes finding the optimal trade-offs between the number of active servers and the waiting time of the jobs.
- A model that allows each server to have a Markovian state that affects the dynamics of the jobs running on the server.
- A model that allows jobs to migrate to different servers at the cost of migration delays.
- A closed-system model where jobs re-enter the system after completion.

A third possible direction is to tackle the problem when the arrival rates and the parameters in the job model are unknown, as mentioned in Section 2.4.3. A possible approach is to develop an approximate version of the JRS framework, where the optimal single-server policy and the simulator for the virtual jobs are both learned from data. It is desirable to design such an approximate framework whose performance degrades gracefully as the approximation error increases.

2.A Proof of Theorem 2.2 (Lower Bound)

Proof. It is sufficient to show that given an infinite-server policy σ for $\mathcal{P}((\lambda_i r)_{i \in \mathcal{I}}, \epsilon)$ in (2.2), we have $N(\sigma) \geq \bar{N}^*$. To this end, we will construct a single-server policy $\bar{\sigma}$ such that the resulting system configuration $\bar{\mathbf{K}}(\infty)$ in steady state satisfies:

$$\mathbb{E} [h(\bar{\mathbf{K}}(\infty)) | \bar{\mathbf{K}}(\infty) \neq \mathbf{0}] = C(\sigma) \leq \epsilon, \quad (2.27)$$

$$\bar{\lambda}_i = \frac{\lambda_i r}{N(\sigma)}, \quad \forall i \in \mathcal{I}. \quad (2.28)$$

Let π be the distribution of $\bar{\mathbf{K}}(\infty)$, then $(N(\sigma), \bar{\sigma}, \pi)$ is a feasible solution to the problem $\bar{\mathcal{P}}((\lambda_i r)_{i \in \mathcal{I}}, \epsilon)$ in (2.3). As a result, we have $N(\sigma) \geq \bar{N}^*$. Note that although $\bar{\sigma}$ is actually non-Markovian, i.e., it makes decisions based on not only the current configuration but also the history, as we will show in Appendix 2.C, $\bar{N}^*$ is still a lower bound to the objective value that $\bar{\sigma}$ can achieve in $\bar{\mathcal{P}}((\lambda_i r)_{i \in \mathcal{I}}, \epsilon)$.

The construction of the single-server policy $\bar{\sigma}$ involves simulating an infinite-server system under σ from the empty configuration. At time 0, the policy $\bar{\sigma}$ randomly chooses the ℓ -th server in the infinite-server system with probability p^ℓ , for $\ell = 1, 2, \dots$. It then requests jobs for the single-server system according to a policy $\bar{\sigma}^\ell$. The key to our policy $\bar{\sigma}^\ell$ is to make the single-server system emulate the job assignment at the ℓ -th server of the simulated infinite-server system, but without incurring idleness. We first construct the policy $\bar{\sigma}^\ell$, and then specify the probabilities p^ℓ .

Let us start by introducing some useful notation. Let $\bar{\mathbf{K}}^\ell(t)$ be the single-server system configuration under $\bar{\sigma}^\ell$ at time t and $\mathbf{K}^\ell(t)$ be the configuration of the ℓ -th simulated server in the infinite-server system under σ . We define a stochastic process $\{s^\ell(t), t \geq 0\}$ as follows: $s^\ell(t) = \max_\tau \{ \tau : \int_0^\tau \mathbb{1}_{\{\mathbf{K}^\ell(x) \neq \mathbf{0}\}} dx = t \}$. The “max” is well-defined because the integral is continuous in τ . Intuitively, $s^\ell(t)$ gives the maximum time when the accumulative busy time of the ℓ -th server is t . Note that $\{s^\ell(t), t \geq 0\}$ is only discontinuous when $\mathbf{K}^\ell(\tau)$ reaches 0, thus it is right-differentiable with derivative equal to 1 at any point.

We construct $\bar{\sigma}^\ell$ and the simulation of the infinite-server system under σ in a way such that:

$$\bar{\mathbf{K}}^\ell(t) = \mathbf{K}^\ell(s^\ell(t)) \quad \forall t. \quad (2.29)$$

That is, we want that the single-server system has the same dynamic of the simulated ℓ -th server except skipping the idle period. To this end, we couple the two systems as follows:

1. When the ℓ -th simulated server $\mathbf{K}^\ell(s^\ell(t))$ receives a type i job, we let the single-server system $\bar{\mathbf{K}}^\ell(t)$ request a type i job at time t . For each such job, its phase transition process in the ℓ -th simulated server is the same as that in the single-server system. That is, when we observe any internal transition or departure event in $\bar{\mathbf{K}}^\ell(t)$, we produce a same event on the ℓ -th simulated server $\mathbf{K}^\ell(s^\ell(t))$.
2. The simulations of the rest of the infinite-server system under policy σ are driven by independently generated random seeds.

It is not hard to see that the simulated infinite server-system has the same stochastic behavior as an uncoupled system under σ . Moreover, as we couple all the events that happen in $\bar{\mathbf{K}}^\ell(t)$ and $\mathbf{K}^\ell(s^\ell(t))$, together with the facts that $\bar{\mathbf{K}}^\ell(t)$ and $\mathbf{K}^\ell(s^\ell(t))$ are piecewise constant and $\bar{\mathbf{K}}^\ell(0-) = \mathbf{K}^\ell(s^\ell(0-)) = \mathbf{0}$, we get (2.29).

Next we claim that (2.29) implies the following relationship between the steady-state cost of the single-server system under $\bar{\sigma}^\ell$ and the steady-state cost of the ℓ -th simulated server in the infinite-server system under σ :

$$\mathbb{E} \left[h \left(\bar{\mathbf{K}}^\ell(\infty) \right) \right] = \frac{\mathbb{E} [h(\mathbf{K}^\ell(\infty))]}{\mathbb{P}(\mathbf{K}^\ell(\infty) \neq \mathbf{0})}. \quad (2.30)$$

This is because for all $\mathbf{k} \neq \mathbf{0}$, we have

$$\begin{aligned} \frac{\mathbb{P}(\mathbf{K}^\ell(\infty) = \mathbf{k})}{\mathbb{P}(\mathbf{K}^\ell(\infty) \neq \mathbf{0})} &\stackrel{(a)}{=} \lim_{S \rightarrow \infty} \frac{\int_0^S \mathbb{1}_{\{\mathbf{K}^\ell(s) = \mathbf{k}\}} ds}{\int_0^S \mathbb{1}_{\{\mathbf{K}^\ell(s) \neq \mathbf{0}\}} ds} \stackrel{(b)}{=} \lim_{T \rightarrow \infty} \frac{\int_0^T \mathbb{1}_{\{\mathbf{K}^\ell(s^\ell(t)) = \mathbf{k}\}} dt}{\int_0^T \mathbb{1}_{\{\mathbf{K}^\ell(s^\ell(t)) \neq \mathbf{0}\}} dt} \\ &\stackrel{(c)}{=} \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \mathbb{1}_{\{\mathbf{K}^\ell(s^\ell(t)) = \mathbf{k}\}} dt \stackrel{(d)}{=} \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \mathbb{1}_{\{\bar{\mathbf{K}}^\ell(t) = \mathbf{k}\}} dt \\ &\stackrel{(e)}{=} \mathbb{P}(\bar{\mathbf{K}}^\ell(\infty) = \mathbf{k}), \end{aligned}$$

where (a) and (e) hold because long-run averages converge to steady-state expectations; (b) is due to the fact that $\int_0^T \mathbb{1}_{\{\mathbf{K}^\ell(s^\ell(t)) = \mathbf{k}\}} dt = \int_0^{s^\ell(T)} \mathbb{1}_{\{\mathbf{K}^\ell(s) = \mathbf{k}\}} ds$, for any $\mathbf{k} \neq \mathbf{0}$; (c) is due to the fact that $\mathbb{1}_{\{\mathbf{K}^\ell(s^\ell(t)) \neq \mathbf{0}\}} = 1$; and (d) follows from (2.29).

Let $\bar{\lambda}_i^\ell$ be the long-run request rates of type i jobs in the single-server system under $\bar{\sigma}^\ell$, and λ_i^ℓ be the throughput of type i jobs of ℓ -th simulated server under σ . By the construction of $\bar{\sigma}^\ell$, the single-server system requests jobs based on the arrival events of the ℓ -th simulated server, we have $\bar{\lambda}_i^\ell = \frac{\lambda_i^\ell}{\mathbb{P}(\mathbf{K}^\ell(\infty) \neq \mathbf{0})}$, $\forall i \in \mathcal{I}$.

With the constructed policies $\{\bar{\sigma}^\ell, \ell = 1, 2, \dots\}$, we are ready to define the policy $\bar{\sigma}$. We let $\bar{\sigma}$ choose an index ℓ with probability p^ℓ at time 0, and then follow $\bar{\sigma}^\ell$. We set the probability p^ℓ as

$$p^\ell = \frac{\mathbb{P}(\mathbf{K}^\ell(\infty) \neq \mathbf{0})}{\sum_{\ell'=1}^{\infty} \mathbb{P}(\mathbf{K}^{\ell'}(\infty) \neq \mathbf{0})} = \frac{\mathbb{P}(\mathbf{K}^\ell(\infty) \neq \mathbf{0})}{\sum_{\mathbf{k} \neq \mathbf{0}} \mathbb{E}[X_{\mathbf{k}}(\infty)]}, \quad \forall \ell = 1, 2, \dots \quad (2.31)$$

where the second inequality uses the fact that $\sum_{\ell'=1}^{\infty} \mathbb{P}(\mathbf{K}^{\ell'}(\infty) \neq \mathbf{0}) = \sum_{\mathbf{k} \neq \mathbf{0}} \mathbb{E}[X_{\mathbf{k}}(\infty)]$. Then under $\bar{\sigma}$, we have

$$\begin{aligned} \mathbb{E}[h(\bar{\mathbf{K}}(\infty))] &= \sum_{\ell=1}^{\infty} p^\ell \mathbb{E}[h(\bar{\mathbf{K}}^\ell(\infty))] \\ &\stackrel{(a)}{=} \sum_{\ell=1}^{\infty} \frac{\mathbb{P}(\mathbf{K}^\ell(\infty) \neq \mathbf{0})}{\sum_{\mathbf{k} \neq \mathbf{0}} \mathbb{E}[X_{\mathbf{k}}(\infty)]} \cdot \frac{\mathbb{E}[h(\mathbf{K}^\ell(\infty))]}{\mathbb{P}(\mathbf{K}^\ell(\infty) \neq \mathbf{0})} \\ &= \frac{\sum_{\ell=1}^{\infty} \mathbb{E}[h(\mathbf{K}^\ell(\infty))]}{\sum_{\mathbf{k} \neq \mathbf{0}} \mathbb{E}[X_{\mathbf{k}}(\infty)]} = \frac{\sum_{\mathbf{k} \neq \mathbf{0}} h(\mathbf{k}) \mathbb{E}[X_{\mathbf{k}}(\infty)]}{\sum_{\mathbf{k} \neq \mathbf{0}} \mathbb{E}[X_{\mathbf{k}}(\infty)]} = C(\sigma), \end{aligned}$$

where (a) follows from (2.30) and (2.31). Observe that under $\bar{\sigma}$, $\bar{\mathbf{K}}(\infty) \neq \mathbf{0}$ almost surely, we thus have

$$\mathbb{E}[h(\bar{\mathbf{K}}(\infty)) | \bar{\mathbf{K}}(\infty) \neq \mathbf{0}] = \mathbb{E}[h(\bar{\mathbf{K}}(\infty))] = C(\sigma),$$

which proves (2.27). Moreover, for each $i \in \mathcal{I}$ the request rate $\bar{\lambda}_i$ is given by

$$\begin{aligned} \bar{\lambda}_i &= \sum_{\ell=1}^{\infty} p^\ell \cdot \bar{\lambda}_i^\ell = \sum_{\ell=1}^{\infty} \frac{\mathbb{P}(\mathbf{K}^\ell(\infty) \neq \mathbf{0})}{\sum_{\mathbf{k} \neq \mathbf{0}} \mathbb{E}[X_{\mathbf{k}}(\infty)]} \cdot \frac{\lambda_i^\ell}{\mathbb{P}(\mathbf{K}^\ell(\infty) \neq \mathbf{0})} \\ &= \frac{\sum_{\ell=1}^{\infty} \lambda_i^\ell}{\sum_{\mathbf{k} \neq \mathbf{0}} \mathbb{E}[X_{\mathbf{k}}(\infty)]} = \frac{\lambda_i r}{N(\sigma)}. \end{aligned}$$

This proves (2.28). By the argument presented at the beginning of the proof, we get $N(\sigma) \geq \bar{N}^*$. $\square$

2.B The rest of the proofs needed for Theorem 2.3 (Conversion Theorem)

2.B.1 Proof of Lemma 2.5

Proof. We will show that under the JRS policy, the Markov chain for the system state (represented as $((\mathbf{K}^\ell(t))_{\ell=1,2,\dots}, \boldsymbol{\zeta}^{1:L}(t), \boldsymbol{\eta}^{1:L}(t))$) has a unique stationary distribution by first arguing that it is $\mathbf{k}^0$ -irreducible (here being $\mathbf{k}^0$ -irreducible means the Markov chain has a state that can be reached by all other states through transitions, “ $\mathbf{k}^0$ ” in “ $\mathbf{k}^0$ -irreducible” does not refer to any specific states), and then use Foster-Lyapunov theorem to show the positive recurrence [see, e.g., 205]. Combining $\mathbf{k}^0$ -irreducibility with positive recurrence, we can conclude that the Markov chain under study has a unique stationary distribution.

First, we show that the Markov chain $((\mathbf{K}^\ell(t))_{\ell=1,2,\dots}, \boldsymbol{\zeta}^{1:L}(t), \boldsymbol{\eta}^{1:L}(t))$ is $\mathbf{k}^0$ -irreducible. Specifically, observe that the Markov chain starting from any state $((\mathbf{k}^\ell)_{\ell=1,2,\dots}, \boldsymbol{\zeta}^{1:L}, \boldsymbol{\eta}^{1:L})$ can reach the state $(\mathbf{k}^{1:L}, (\mathbf{0})_{\ell=L+1,\dots}, \mathbf{0}^{1:L}, \mathbf{0}^{1:L})$, after experiencing a sequence of departures and arrivals that clears up all the tokens, virtual jobs and jobs on backup servers. Further, letting $\tilde{\mathbf{k}}$ be the configuration reachable by all other configuration in the single-server system under the policy $\bar{\sigma}$, we argue that starting from any states of the form $(\mathbf{k}^{1:L}, (\mathbf{0})_{\ell=L+1,\dots}, \mathbf{0}^{1:L}, \mathbf{0}^{1:L})$, the Markov chain can reach the state $(\tilde{\mathbf{k}}^{1:L}, (\mathbf{0})_{\ell=L+1,\dots}, \mathbf{0}^{1:L}, \mathbf{0}^{1:L})$. Because for any $\ell \leq L$, there is a transition path from $\mathbf{k}^\ell$ to $\tilde{\mathbf{k}}$, consider the sequence of events where each $\mathbf{K}^\ell(t)$ transitions independently following the path, and the jobs arrive right after $\mathbf{K}^\ell(t)$ making a request, so that the tokens are checked out before $\mathbf{K}^\ell(t)$ has a further transition. In this way, each $\mathbf{K}^\ell(t)$ with $\ell \leq L$ can eventually reach $\tilde{\mathbf{k}}$ from $\mathbf{k}^\ell$. This proves the $\mathbf{k}^0$ -irreducibility of $((\mathbf{K}^\ell(t))_{\ell=1,2,\dots}, \boldsymbol{\zeta}^{1:L}(t), \boldsymbol{\eta}^{1:L}(t))$.

Next, we show that $((\mathbf{K}^\ell(t))_{\ell=1,2,\dots}, \boldsymbol{\zeta}^{1:L}(t), \boldsymbol{\eta}^{1:L}(t))$ satisfies the Foster-Lyapunov criterion, i.e.,

$$\widehat{G}g \leq -1 + b\mathbb{1}_{\{S\}}, \quad (2.32)$$

where g is a non-negative function of the states, S is a finite set, b is a finite number, and $\widehat{G}$ is the infinitesimal generator of the continuous-time Markov chain. Let t_i be the expected remaining time in the system when a job is in

phase i for each $i \in \mathcal{I}$. According to the job model, we have the recurrence relation

$$\left(\mu_{i\perp} + \sum_{i' \in \mathcal{I}: i' \neq i} \mu_{ii'}\right)t_i = \sum_{i' \in \mathcal{I}: i' \neq i} \mu_{ii'}t_{i'} \quad \forall i \in \mathcal{I}. \quad (2.33)$$

We construct a Lyapunov function g as follows:

$$g((\mathbf{k}^\ell)_{\ell=1,2,\dots}, \boldsymbol{\zeta}^{1:L}, \boldsymbol{\eta}^{1:L}) = \sum_{i \in \mathcal{I}} \sum_{\ell=1}^{\infty} t_i k_i^\ell + \sum_{i \in \mathcal{I}} \sum_{\ell=1}^{\infty} t_i \zeta_i^\ell. \quad (2.34)$$

Using the relation (2.33), it can be verified that the drift of g satisfies

$$\begin{aligned} & \widehat{G}g((\mathbf{k}^\ell)_{\ell=1,2,\dots}, \boldsymbol{\zeta}^{1:L}, \boldsymbol{\eta}^{1:L}) \\ & \leq \sum_{i \in \mathcal{I}} \left(\lambda_i t_i r - \sum_{\ell=1}^{\infty} k_i^\ell \right) + \sum_{i \in \mathcal{I}} \left(\sum_{\ell=1}^L \sum_{(\mathbf{k}', \mathbf{a}) \in E(\mathbf{k}^\ell)} \gamma_{\mathbf{k}^\ell, (\mathbf{k}', \mathbf{a})} a_i t_i - \sum_{\ell=1}^{\infty} \zeta_i^\ell \right) \\ & \leq \sum_{i \in \mathcal{I}} \left(\lambda_i t_i r + L \cdot \max_{\mathbf{k} \in \mathcal{K}} \sum_{(\mathbf{k}', \mathbf{a}) \in E(\mathbf{k})} \gamma_{\mathbf{k}, (\mathbf{k}', \mathbf{a})} a_i t_i \right) - \left(\sum_{\ell=1}^{\infty} \sum_{i \in \mathcal{I}} k_i^\ell + \sum_{\ell=1}^{\infty} \sum_{i \in \mathcal{I}} \zeta_i^\ell \right), \end{aligned}$$

where the first inequality uses the fact that virtual jobs are generated at a rate no faster than the total rate of job requests. Then the Foster-Lyapunov criterion in (2.32) is satisfied with b and S given by

$$b = \sum_{i \in \mathcal{I}} \left(\lambda_i t_i r + L \cdot \max_{\mathbf{k} \in \mathcal{K}} \sum_{(\mathbf{k}', \mathbf{a}) \in E(\mathbf{k})} \gamma_{\mathbf{k}, (\mathbf{k}', \mathbf{a})} a_i t_i \right),$$

$$S = \left\{ ((\mathbf{k}^\ell)_{\ell=1,2,\dots}, \boldsymbol{\zeta}^{1:L}, \boldsymbol{\eta}^{1:L}) : g((\mathbf{k}^\ell)_{\ell=1,2,\dots}, \boldsymbol{\zeta}^{1:L}, \boldsymbol{\eta}^{1:L}) \leq b + 1 \right\}.$$

By the Foster-Lyapunov theorem, $((\mathbf{K}^\ell(t))_{\ell=1,2,\dots}, \boldsymbol{\zeta}^{1:L}(t), \boldsymbol{\eta}^{1:L}(t))$ is positive recurrent. $\square$

2.B.2 Proofs of Lemma 2.7 and Lemma 2.8

In this subsection, we prove Lemma 2.7 and Lemma 2.8 together. We begin by introducing some notations. We use $\mathbf{U}^\ell \triangleq (\widehat{\mathbf{K}}^\ell, \boldsymbol{\eta}^\ell)$ to represent the state of the ℓ -th server, and use $\mathbf{U}^{1:L}$ to represent the joint state of the first L servers. We also use the lowercase $\mathbf{u}^\ell, \mathbf{u}^{1:L}$ to represent the realizations of the corresponding random variables. We denote the total number of type i virtual jobs as $V_i \triangleq \sum_{\ell=1}^L \zeta_i^\ell$ for $i \in \mathcal{I}$, and its realization as v_i . We denote the total

number of type i jobs on backup servers as Y_i for $i \in \mathcal{I}$, and its realizations as y_i . We also denote the total number of type i tokens throughout the system as $Z_i \triangleq \sum_{\ell=1}^L \eta_i^\ell$, and its realization as z_i . Our goal can be rewritten as proving $\mathbb{E}[V_i] = O(\sqrt{r})$ and $\mathbb{E}[Y_i] = O(\sqrt{r})$ for each $i \in \mathcal{I}$.

We first give an overview of the proof. Observe that in our model, the expected time that a job stay in the system is fixed. As a result, bounding the number of virtual jobs or jobs on backup servers in the system is equivalent to bounding the rate that they are generated, according to Little's Law. By our construction of the policy, the rate of generating those jobs is closely related to the dynamics of the total number of type i tokens $Z_i(t)$.

To describe the dynamics of $Z_i(t)$, we first introduce two functions dv_i and dy_i :

$$\begin{aligned} dv_i(a_i, z_i) &\triangleq (z_i + a_i - \eta_{\max})^+, \\ dy_i(z_i) &\triangleq (1 - z_i)^+. \end{aligned}$$

The function dv_i represents the increment in the number of type i virtual jobs due to the event that the total number of type i tokens on the first L servers exceeds the token limit $\eta_{\max}$. The function dy_i corresponds to the increment in the total number of type i jobs on backup servers due to the event that a type i job arrives to the system without seeing a type i token. For a function $g: (\mathcal{K} \times \mathcal{K})^L \rightarrow \mathbb{R}$ that only depends on the number of type i tokens z_i , its drift can be written as

$$\begin{aligned} \widehat{G}g(\mathbf{u}^{1:L}) &= \sum_{\ell=1}^L \sum_{(\mathbf{k}', \mathbf{a}) \in E(\mathbf{k}^\ell)} \gamma_{\mathbf{k}^\ell, (\mathbf{k}', \mathbf{a})} (g(z_i + a_i - dv_i(a_i, z_i)) - g(z_i)) \mathbb{1}_{\{\eta^\ell = \mathbf{0}\}} \\ &\quad + \lambda_i r (g(z_i - 1 + dy_i(z_i)) - g(z_i)). \end{aligned} \tag{2.35}$$

We abuse the notation of g here. For ease of exposition, we will simply write dv_i and dy_i to represent $dv_i(a_i, z_i)$ and $dy_i(z_i)$.

By construction, the total number of type i tokens $\{Z_i(t)\}$ is a stochastic process constrained within $[0, \eta_{\max}]$. Note that $Z_i(t)$ increases when some servers request new tokens, and decreases when a real or virtual arrival checks out the token or when some servers have the excessive tokens removed. When $Z_i(t)$ is away from the boundaries, the average rate that it increases is

approximately equal to $\lambda_i r$, and the average rate that $Z_i(t)$ decreases is given by

$$\begin{aligned} \mathbb{E} \left[\sum_{\ell=1}^L \sum_{(\mathbf{k}', \mathbf{a}) \in E(\widehat{\mathbf{K}}^\ell)} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i \mathbb{1}_{\{\eta^\ell = 0\}} \right] &\approx \mathbb{E} \left[\sum_{\ell=1}^L \sum_{(\mathbf{k}', \mathbf{a}) \in E(\overline{\mathbf{K}}^\ell)} \gamma_{\overline{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i \right] \\ &= L \cdot \bar{\lambda}_i \approx \lambda_i r, \end{aligned}$$

where we have used the approximations that $\widehat{\mathbf{K}}^\ell \stackrel{d}{\approx} \overline{\mathbf{K}}^\ell$, $\mathbb{1}_{\{\eta^\ell = 0\}} \approx 1$ and $L = \lceil N \rceil \approx \bar{N}$.

As $\{Z_i(t)\}$ randomly moves up and down with approximately the same rate and reflects on the boundaries of 0 and $\eta_{\max}$, it behaves as a reflected simple symmetric random walk. Intuitively speaking, the steady-state distribution of Z_i is approximately a uniform distribution over $[0, \eta_{\max}]$. Recall that dv_i and dy_i can only be non-zero when $Z_i(t)$ is near the boundaries. Since the length of the interval $\eta_{\max} = \Theta(\sqrt{r})$, we can expect that dv_i and dy_i diminish as $r \rightarrow \infty$.

In the proof, we first establish the relationship between $\mathbb{E}[V_i]$, $\mathbb{E}[Y_i]$ and dv_i , dy_i using Little's Law. Then we derive bounds on dv_i and dy_i by analyzing the drift of several test functions of Z_i . This step is implicitly based on the intuition that Z_i is approximately uniformly distributed over $[0, \eta_{\max}]$, with the tokens being generated and eliminated at similar rates. Finally, we invoke Lemma 2.6 to show that the tokens are indeed generated and eliminated at similar speeds, which leads to bounds on dv_i and dy_i .

Finally, we make some additional remarks on the notations. First, dv_i and dy_i depend on the total number of type i tokens z_i and the number of newly requested jobs a_i , although we omit the dependency expression for ease of exposition. Second, we abuse the notation dv_i and dy_i to denote the corresponding random variables. We also write $\sum_{\mathbf{k}', \mathbf{a}}$ as a shorthand for $\sum_{(\mathbf{k}', \mathbf{a}) \in E(\mathbf{k}^\ell)}$ when the context is clear.

Proofs of Lemma 2.7 and Lemma 2.8.

We are now ready to prove Lemma 2.7 and Lemma 2.8.

Proof. Step 1: Bounding Virtual Jobs and Jobs on Backup Servers using Little's Law. We first apply Little's Law to V_i and Y_i . For each type i , we let the expected time that a type i job stays in the system be t_i . Let

$t_{\max} = \max_{i \in \mathcal{I}} t_i$. Because there are only finitely many types of jobs, and each job spends finite expected time in the system, $t_{\max}$ is a finite constant. By Little's law,

$$\mathbb{E}[V_i] \leq t_{\max} \mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} dv_i \mathbb{1}_{\{\eta^\ell = \mathbf{0}\}} \right], \quad (2.36)$$

$$\mathbb{E}[Y_i] \leq t_{\max} \mathbb{E} [\lambda_i r \cdot dy_i]. \quad (2.37)$$

Step 2: Drift Analysis. The above two equations (2.36) and (2.37) suggest that we can derive upper bounds on $\mathbb{E}[V_i]$ and $\mathbb{E}[Y_i]$ by analyzing the following two terms:

- $\mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} dv_i \mathbb{1}_{\{\eta^\ell = \mathbf{0}\}} \right]$, interpreted as the average rate that $Z_i(t)$ reflects on the boundary at $\eta_{\max}$;
- $\mathbb{E} [\lambda_i r \cdot dy_i]$, interpreted as the average rate that $Z_i(t)$ reflects on the boundary at 0.

We establish the relationships of dv_i , dy_i and Z_i by analyzing the drift of two test functions g .

Letting $g(z_i) = z_i$ and taking steady-state expectation over its drift, by (2.35) and the fact that the drift is zero in steady state, we get

$$\mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} (a_i - dv_i) \mathbb{1}_{\{\eta^\ell = \mathbf{0}\}} + \lambda_i r (-1 + dy_i) \right] = 0. \quad (2.38)$$

Similarly, letting $g(z_i) = z_i^2$ and taking steady-state expectation over its drift, one can verify that

$$\mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} dv_i \mathbb{1}_{\{\eta^\ell = \mathbf{0}\}} \right] \quad (2.39)$$

$$= \frac{1}{\eta_{\max}} \cdot \mathbb{E} \left[\left(\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i \mathbb{1}_{\{\eta^\ell = \mathbf{0}\}} - \lambda_i r \right) \cdot Z_i \right] \quad (2.40)$$

$$+ \frac{1}{2\eta_{\max}} \cdot \mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} (a_i^2 - (dv_i)^2) \mathbb{1}_{\{\eta^\ell = \mathbf{0}\}} + \lambda_i r \cdot (1 - (dy_i)^2) \right]. \quad (2.41)$$

Readers may refer to the complete calculation at the end of this subsection.

Step 3: Estimating the Terms Obtained from Drift Analysis. We will first focus on bounding $\mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} dv_i \mathbb{1}_{\{\eta^\ell = \mathbf{0}\}} \right]$ analyzing the two terms in (2.40) and (2.41) separately. Then we invoke (2.38) to bound $\mathbb{E} [\lambda_i r \cdot dy_i]$.

The term in (2.41) is easy to deal with. Observe that the number of jobs requested each time should be no more than the maximal number of jobs that a server can hold, i.e., $a_i \leq K_{\max}$, so

$$\begin{aligned} (2.41) &\leq \frac{1}{2\eta_{\max}} \cdot \mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} K_{\max}^2 + \lambda_i r \right] \\ &\leq \frac{1}{2\eta_{\max}} \cdot \mathbb{E} \left[\sum_{\ell=1}^L \gamma_{\max} K_{\max}^2 + \lambda_i r \right] \\ &= O(\sqrt{r}), \end{aligned} \tag{2.42}$$

where in the second inequality we have used the fact that the total rate is uniformly bounded by $\gamma_{\max}$, and the last step uses the facts that $L = O(r)$ and $\eta_{\max} = \Theta(\sqrt{r})$.

To bound the term in (2.40), first observe that $Z_i \leq \eta_{\max}$, which implies that

$$(2.40) \leq \mathbb{E} \left[\left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i \mathbb{1}_{\{\eta^\ell = \mathbf{0}\}} - \lambda_i r \right| \right].$$

The term on the RHS of the above equation is the expected absolute difference between the rates of generating and eliminating type i tokens, which can be shown to be small relative to r . Specifically, we claim that

$$\mathbb{E} \left[\left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i \mathbb{1}_{\{\eta^\ell = \mathbf{0}\}} - \lambda_i r \right| \right] = O(\sqrt{r}). \tag{2.43}$$

To show (2.43), first notice that we can remove the indicator $\mathbb{1}_{\{\eta^\ell = \mathbf{0}\}}$ without introducing much error:

$$\mathbb{E} \left[\left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i \mathbb{1}_{\{\eta^\ell = \mathbf{0}\}} - \lambda_i r \right| \right]$$

$$\begin{aligned}
&\leq \mathbb{E} \left[\left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i - \lambda_i r \right| \right] + \mathbb{E} \left[\left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i \mathbb{1}_{\{\eta^\ell \neq \mathbf{0}\}} \right| \right] \\
&\leq \mathbb{E} \left[\left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i - \lambda_i r \right| \right] + \mathbb{E} \left[\left| \sum_{\ell=1}^L \gamma_{\max} K_{\max} \mathbb{1}_{\{\eta^\ell \neq \mathbf{0}\}} \right| \right] \\
&\leq \mathbb{E} \left[\left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i - \lambda_i r \right| \right] + \gamma_{\max} K_{\max} |\mathcal{I}| \eta_{\max},
\end{aligned}$$

where the first inequality is due to triangular inequality, the second inequality is due to the definition of $\gamma_{\max}$, and the last inequality is because $\sum_{\ell=1}^L \mathbb{1}_{\{\eta^\ell \neq \mathbf{0}\}} \leq |\mathcal{I}| \eta_{\max}$. It remains to bound the term $\mathbb{E} \left[\left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i - \lambda_i r \right| \right]$, which can be seen as showing that the rate of generating type i tokens concentrates around the type i jobs' arrival rate $\lambda_i r$. It is natural to think of using some Law of Large Numbers. Unfortunately, $\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i$ is not a sum of i.i.d. random variables due to dependencies among $\widehat{\mathbf{K}}^\ell$ for different ℓ 's. As a result, we want to invoke the Wasserstein distance bound in Lemma 2.6 to replace $\widehat{\mathbf{K}}^\ell$ in the above expression with $\overline{\mathbf{K}}^\ell$. We define the function $f(\mathbf{k}^{1:L})$ as

$$f(\mathbf{k}^{1:L}) = \frac{1}{2\gamma_{\max} K_{\max}} \left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\mathbf{k}^\ell, (\mathbf{k}', \mathbf{a})} a_i - \lambda_i r \right|. \quad (2.44)$$

We claim that $f \in \text{Lip}(1)$. For any two $\mathbf{k}^{1:L,1}, \mathbf{k}^{1:L,2}$,

$$\begin{aligned}
&2\gamma_{\max} K_{\max} \cdot (f(\mathbf{k}^{1:L,1}) - f(\mathbf{k}^{1:L,2})) \\
&= \left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\mathbf{k}^{\ell,1}, (\mathbf{k}', \mathbf{a})} a_i - \lambda_i r \right| - \left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\mathbf{k}^{\ell,2}, (\mathbf{k}', \mathbf{a})} a_i - \lambda_i r \right| \\
&\leq \left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\mathbf{k}^{\ell,1}, (\mathbf{k}', \mathbf{a})} a_i - \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\mathbf{k}^{\ell,2}, (\mathbf{k}', \mathbf{a})} a_i \right| \\
&= \left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} (\gamma_{\mathbf{k}^{\ell,1}, (\mathbf{k}', \mathbf{a})} - \gamma_{\mathbf{k}^{\ell,2}, (\mathbf{k}', \mathbf{a})}) a_i \mathbb{1}_{\{(\mathbf{k}^{\ell,1}) \neq (\mathbf{k}^{\ell,2})\}} \right| \\
&\leq \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} |\gamma_{\mathbf{k}^{\ell,1}, (\mathbf{k}', \mathbf{a})} - \gamma_{\mathbf{k}^{\ell,2}, (\mathbf{k}', \mathbf{a})}| \cdot K_{\max} \cdot \|\mathbf{k}^{\ell,1} - \mathbf{k}^{\ell,2}\|
\end{aligned}$$

$$\begin{aligned}
&\leq \sum_{\ell=1}^L 2\gamma_{\max} K_{\max} \cdot \|\mathbf{k}^{\ell,1} - \mathbf{k}^{\ell,2}\| \\
&= 2\gamma_{\max} K_{\max} \cdot \|\mathbf{k}^{1:L,1} - \mathbf{k}^{1:L,2}\|,
\end{aligned}$$

where the first inequality is due to triangular inequality; the second inequality uses the fact that $a_i \leq K_{\max}$ and $\mathbb{1}_{\{\mathbf{k}^{1,\ell} \neq \mathbf{k}^{2,\ell}\}} \leq \|\mathbf{k}^{1,\ell} - \mathbf{k}^{2,\ell}\|$; the third inequality uses triangular inequality, the fact that the total rate at a configuration $\mathbf{k}$ is bounded by $\gamma_{\max}$ and the property of the L^1 norm $\|\cdot\|$. Therefore, $f \in \text{Lip}(1)$. The Lipschitz continuity of f allows us to invoke Lemma 2.6 and get

$$\begin{aligned}
&\mathbb{E} \left[\left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i - \lambda_i r \right| \right] - \mathbb{E} \left[\left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\overline{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i - \lambda_i r \right| \right] \\
&\leq 2\gamma_{\max} K_{\max} \cdot O(\sqrt{r}).
\end{aligned}$$

Therefore,

$$\mathbb{E} \left[\left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i - \lambda_i r \right| \right] \leq \mathbb{E} \left[\left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\overline{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i - \lambda_i r \right| \right] + O(\sqrt{r}). \quad (2.45)$$

Observe that under a Markovian policy, the request rate of type i jobs can be written as $\bar{\lambda}_i = \mathbb{E}[\sum_{\mathbf{k}', \mathbf{a}} \gamma_{\overline{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i]$, so we have

$$\mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\overline{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i - \lambda_i r \right] = \bar{\lambda}_i \cdot \lceil \bar{N} \rceil - \lambda_i r = O(1). \quad (2.46)$$

Moreover, because $\sum_{\mathbf{k}', \mathbf{a}} \gamma_{\overline{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i$ are i.i.d. for $\ell = 1, \dots, L$, we have

$$\begin{aligned}
&\mathbb{E} \left[\left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\overline{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i - \lambda_i r \right| \right] \\
&\leq \sqrt{\mathbb{E} \left[\left(\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\overline{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i - \lambda_i r \right)^2 \right]} + O(1) \\
&= \sqrt{\sum_{\ell=1}^L \text{Var} \left(\sum_{\mathbf{k}', \mathbf{a}} \gamma_{\overline{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i \right)} + O(1)
\end{aligned}$$

$$= O(\sqrt{r}).$$

Therefore, by combining the arguments above, we get

$$\begin{aligned} & \mathbb{E} \left[\left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i \mathbb{1}_{\{\boldsymbol{\eta}^\ell = \mathbf{0}\}} - \lambda_i r \right| \right] \\ & \leq \mathbb{E} \left[\left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i - \lambda_i r \right| \right] + O(\sqrt{r}) \\ & \leq \mathbb{E} \left[\left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\overline{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i - \lambda_i r \right| \right] + O(\sqrt{r}) \\ & \leq O(\sqrt{r}), \end{aligned} \tag{2.47}$$

which proves (2.43). This implies that the term in (2.40) is also in $O(\sqrt{r})$.

Combining the bounds on the terms in (2.40) and (2.41), we get

$$\mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} dv_i \mathbb{1}_{\{\boldsymbol{\eta}^\ell = \mathbf{0}\}} \right] = O(\sqrt{r}).$$

Finally, we bound $\mathbb{E}[\lambda_i r \cdot dy_i]$. We rearrange the terms in (2.38) and get

$$\begin{aligned} \mathbb{E}[\lambda_i r \cdot dy_i] &= \mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} (-a_i + dv_i) \mathbb{1}_{\{\boldsymbol{\eta}^\ell = \mathbf{0}\}} + \lambda_i r \right] \\ &= \mathbb{E} \left[- \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i \mathbb{1}_{\{\boldsymbol{\eta}^\ell = \mathbf{0}\}} + \lambda_i r \right] + O(\sqrt{r}). \end{aligned}$$

By (2.43), we have $\mathbb{E} \left[- \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i \mathbb{1}_{\{\boldsymbol{\eta}^\ell = \mathbf{0}\}} + \lambda_i r \right] = O(\sqrt{r})$. Therefore,

$$\mathbb{E}[\lambda_i r \cdot dy_i] = O(\sqrt{r}).$$

We invoke the equations (2.36) and (2.37) that we get at the beginning of the proof, and conclude that

$$\begin{aligned} \mathbb{E}[V_i] &\leq t_{\max} \mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} dv_i \mathbb{1}_{\{\boldsymbol{\eta}^\ell = \mathbf{0}\}} \right] = O(\sqrt{r}). \\ \mathbb{E}[Y_i] &\leq t_{\max} \mathbb{E}[\lambda_i r \cdot dy_i] = O(\sqrt{r}). \end{aligned}$$

This finishes the proof. $\square$

Deriving the equality in (2.39)

We show the calculation detail of deriving the following equality.

$$\mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} dv_i \mathbb{1}_{\{\eta^\ell = \mathbf{0}\}} \right] \quad (2.39)$$

$$= \frac{1}{\eta_{\max}} \cdot \mathbb{E} \left[\left(\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i \mathbb{1}_{\{\eta^\ell = \mathbf{0}\}} - \lambda_i r \right) \cdot Z_i \right] \quad (2.40)$$

$$+ \frac{1}{2\eta_{\max}} \cdot \mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} (a_i^2 - (dv_i)^2) \mathbb{1}_{\{\eta^\ell = \mathbf{0}\}} + \lambda_i r \cdot (1 - (dy_i)^2) \right]. \quad (2.41)$$

The equality is obtained by considering the drift of the function $g(z_i) = z_i^2$, which is zero in steady state. Recall that the drift of $g(z_i)$ is given by

$$\begin{aligned} \widehat{G}g(z_i, \mathbf{u}^{1:L}) &= \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\mathbf{k}^\ell, (\mathbf{k}', \mathbf{a})} (g(z_i + a_i - dv_i) - g(z_i)) \mathbb{1}_{\{\eta^\ell = \mathbf{0}\}} \\ &\quad + \lambda_i r (g(z_i - 1 + dy_i) - g(z_i)). \end{aligned} \quad (2.35)$$

We will first calculate $g(z_i + a_i - dv_i) - g(z_i)$, then $g(z_i - 1 + dy_i) - g(z_i)$, and finally plug them into the (2.35).

The calculation of $g(z_i + a_i - dv_i) - g(z_i)$ utilizes the following property of dv_i :

$$(z_i + a_i - dv_i) \cdot dv_i = \eta_{\max} \cdot dv_i. \quad (2.48)$$

This property follows from the definition $dv_i = (z_i + a_i - \eta_{\max})^+$. Intuitively, this is because dv_i is the “force” that pushes z_i back when it hits the boundary at $\eta_{\max}$. Using the property, we have

$$\begin{aligned} &g(z_i + a_i - dv_i) - g(z_i) \\ &= (z_i + a_i - dv_i)^2 - z_i^2 \\ &= (z_i + a_i - dv_i)^2 - (z_i + a_i - dv_i - a_i + dv_i)^2 \\ &= (z_i + a_i - dv_i)^2 \\ &\quad - ((z_i + a_i - dv_i)^2 + 2(-a_i + dv_i) \cdot (z_i + a_i - dv_i) + (-a_i + dv_i)^2) \\ &= 2(a_i - dv_i) \cdot (z_i + a_i - dv_i) - (-a_i + dv_i)^2 \end{aligned}$$

$$\begin{aligned}
&= 2a_i \cdot (z_i + a_i - dv_i) - (-a_i + dv_i)^2 - 2dv_i \cdot \eta_{\max} \\
&= 2a_i \cdot z_i + a_i^2 - (dv_i)^2 - 2dv_i \cdot \eta_{\max}.
\end{aligned}$$

The second last equality is due to (2.48), and the rest are all algebraic manipulations.

We carry out a similar calculation for $g(z_i - 1 + dy_i) - g(z_i)$:

$$\begin{aligned}
&g(z_i - 1 + dy_i) - g(z_i) \\
&= 2(-1 + dy_i) \cdot z_i + (-1 + dy_i)^2 \\
&= -2z_i + 2z_i \cdot dy_i + 1 - 2dy_i + (dy_i)^2 \\
&= -2z_i + 1 + 2(z_i - 1 + dy_i) \cdot dy_i - (dy_i)^2 \\
&= -2z_i + 1 - (dy_i)^2.
\end{aligned}$$

where the last equality is due to the property that

$$(z_i - 1 + dy_i) \cdot dy_i = 0, \quad (2.49)$$

and the rest are all algebraic manipulations.

Putting together,

$$\begin{aligned}
\widehat{G}g(z_i, \mathbf{u}^{1:L}) &= \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\mathbf{k}^\ell, (\mathbf{k}', \mathbf{a})} (g(z_i + a_i - dv_i) - g(z_i)) \mathbb{1}_{\{\boldsymbol{\eta}^\ell = \mathbf{0}\}} \\
&\quad + \lambda_i r (g(z_i - 1 + dy_i) - g(z_i)) \\
&= \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\mathbf{k}^\ell, (\mathbf{k}', \mathbf{a})} (2a_i \cdot z_i + a_i^2 - (dv_i)^2 - 2dv_i \cdot \eta_{\max}) \mathbb{1}_{\{\boldsymbol{\eta}^\ell = \mathbf{0}\}} \\
&\quad + \lambda_i r \cdot (-2z_i + 1 - (dy_i)^2).
\end{aligned}$$

After recombining the terms, we get

$$\begin{aligned}
&\eta_{\max} \cdot \mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} dv_i \mathbb{1}_{\{\boldsymbol{\eta}^\ell = \mathbf{0}\}} \right] \\
&= \mathbb{E} \left[\left(\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i \mathbb{1}_{\{\boldsymbol{\eta}^\ell = \mathbf{0}\}} - \lambda_i r \right) \cdot Z_i \right] \\
&\quad + \frac{1}{2} \mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} (a_i^2 - (dv_i)^2) \mathbb{1}_{\{\boldsymbol{\eta}^\ell = \mathbf{0}\}} + \lambda_i r \cdot (1 - (dy_i)^2) \right]
\end{aligned}$$

This finishes the calculation.

2.B.3 Proof of Conversion Theorem without assuming irreducibility

In this subsection, we prove Theorem 2.3 without assuming $\mathbf{k}^0$ -irreducibility of the subroutine $\bar{\sigma}$. Specifically, suppose that we have a Markovian single-server policy $\bar{\sigma}$ and an initial distribution p_j over its recurrent classes S_j for $j = 1, 2, \dots, J$, we will construct an infinite-server policy σ such that (2.5) and (2.6) still hold. The basic idea is to decompose a general Markovian single-server policy $\bar{\sigma}$ into multiple $\mathbf{k}^0$ -irreducible Markovian policies, each induces one recurrent class and preserves stationary distribution and the throughput of $\bar{\sigma}$ on that recurrent class, as stated in Lemma 2.10 below.

Lemma 2.10 (Decomposing The Reducible Policy). *Let $\bar{\sigma}$ be a general single-server Markovian policy with recurrent classes S_j for $j = 1, 2, \dots, J$. Then for each j exists a Markovian policy $\bar{\sigma}^j$ such that*

- *The induced Markov chain is $\mathbf{k}^0$ -irreducible with the unique recurrent class being S_j ;*
- *The stationary distribution is the same as the stationary distribution under $\bar{\sigma}$ starting from a configuration in S_j .*

Proof. For each $j = 1, 2, \dots, J$, we define the policy $\bar{\sigma}^j$ as follows: when the system has configuration $\mathbf{k} \in S_j$, the policy $\bar{\sigma}^j$ makes the same decisions as $\bar{\sigma}$; when $\mathbf{k} \notin S_j$ and $\mathbf{k} = \mathbf{0}$, the policy starts a timer whose duration follows an exponential distribution with rate 1 and immediately adds $\mathbf{k}^0$ many jobs of each type when the timer ticks for some arbitrary $\mathbf{k}^0 \in S_j$; when $\mathbf{k} \notin S_j$ and $\mathbf{k} \neq \mathbf{0}$, the policy does not request any jobs.

We show that under the new policy $\bar{\sigma}^j$, S_j is also a recurrent class of the induced Markov chain. This is because if the system starts from a configuration $\mathbf{k} \in S_j$, then it will stay in S_j since it makes the same decisions and has the same transitions as under the policy $\bar{\sigma}$. Because S_j is a recurrent class under $\bar{\sigma}$, it is still a recurrent class under $\bar{\sigma}^j$.

To show that the Markov chain induced by $\bar{\sigma}^j$ is $\mathbf{k}^0$ -irreducible, observe that starting from any $\mathbf{k} \notin S_j$, the system state will return to S_j . Specifically,

- If the system starts from a configuration $\mathbf{k}$ such that $\mathbf{k} \notin S_j$ and $\mathbf{k} \neq \mathbf{0}$, then no new jobs will be requested until either $\mathbf{k} \in S_j$ or $\mathbf{k} = \mathbf{0}$. In the latter case, by the construction of the policy, the system jumps to a configuration in S_j after the next transition.
- If the system starts from a configuration $\mathbf{k}$ such that $\mathbf{k} \notin S_j$ and $\mathbf{k} = \mathbf{0}$, the system jumps to a configuration in S_j after the next transition.

The claim that the stationary distribution under $\bar{\sigma}^j$ is the same as the stationary distribution under $\bar{\sigma}$ starting from a configuration in S_j is trivial to show, because when the system initializes from any configuration in S_j , it stays in S_j and the transitions are exactly the same under the two policies. $\square$

The JRS policy with a general Markovian single-server policy as its subroutine $\bar{\sigma}$ is constructed using the $\mathbf{k}^0$ -irreducible policies $\bar{\sigma}^j$'s obtained from the decomposition of $\bar{\sigma}$ and the probabilities p^j 's.

1. We divide all the servers into J server pools, each with infinitely many servers. Let the j -th server pool run the JRS policy with subroutine $\bar{\sigma}^j$ (defined in Section 2.4.2) for each $j = 1, \dots, J$.
2. Whenever we see an arrival of type i , we route the job to the j -th infinite-server system with probability $\frac{p^j \bar{\lambda}_i^j}{\sum_j p^j \bar{\lambda}_i^j}$ for each $j = 1, \dots, J$.

To analyze the policy σ , let π^j and $\bar{\lambda}_i^j$'s be the stationary distribution and throughput of the policy $\bar{\sigma}^j$ for $j = 1, \dots, J$. By Lemma 2.10, we have the following relationships:

$$\pi(\mathbf{k}) = \sum_{j=1}^J p^j \pi^j(\mathbf{k}), \quad \forall \mathbf{k} \in \mathcal{K}, \quad (2.50)$$

$$\bar{\lambda}_i = \sum_{j=1}^J p^j \bar{\lambda}_i^j, \quad \forall i \in \mathcal{I}. \quad (2.51)$$

Based on the above relationships, we can prove the general version of the Conversion Theorem that does not require $\mathbf{k}^0$ -irreducibility.

Proof of Theorem 2.3. For each $j = 1, \dots, J$, Lemma 2.10 implies that the policy $\bar{\sigma}^j$ and stationary distribution π^j form a feasible solution to the single-server problem $\bar{\mathcal{P}}((p^j \bar{\lambda}_i^j \bar{N})_{i \in \mathcal{I}}, \epsilon)$, and the corresponding objective value is $p^j \bar{N}$. Consider the infinite-server system with arrival rates $(p^j \bar{\lambda}_i^j \bar{N})_{i \in \mathcal{I}}$ and budget ϵ . As we have proved Theorem 2.3 for the JRS policy with $\mathbf{k}^0$ -irreducible subroutine $\bar{\sigma}^j$, it follows that

$$\left| \sum_{\mathbf{k} \neq \mathbf{0}} \mathbb{E}[X_{\mathbf{k}}^j] - \lceil p^j \bar{N} \rceil \cdot (1 - \pi^j(\mathbf{0})) \right| = O(\sqrt{r}), \quad (2.52)$$

$$\left| \sum_{\mathbf{k} \neq \mathbf{0}} h(\mathbf{k}) \mathbb{E}[X_{\mathbf{k}}^j] - \lceil p^j \bar{N} \rceil \cdot \sum_{\mathbf{k} \neq \mathbf{0}} h(\mathbf{k}) \pi^j(\mathbf{k}) \right| = O(\sqrt{r}), \quad (2.53)$$

where $X_{\mathbf{k}}^j$ is the random variable representing the steady-state number of servers in configuration $\mathbf{k}$ in the infinite-server system under the JRS policy with subroutine $\bar{\sigma}^j$.

By the construction of σ , the arrival rate to the j -th server pool is equal to $\frac{p^j \bar{\lambda}_i^j}{\sum_j p^j \bar{\lambda}_{i'}^j} \cdot \lambda_i r = \frac{p^j \bar{\lambda}_i^j}{\bar{\lambda}_i} \cdot \lambda_i r = p^j \bar{\lambda}_i^j \bar{N}$, where the first equality is due to (2.51), and the second equality is due to the condition that $\lambda_i r = \bar{\lambda}_i^j \cdot L$. Therefore, (2.52) and (2.53) hold, and we have

$$\begin{aligned} \left| \sum_{\mathbf{k} \neq \mathbf{0}} \mathbb{E}[X_{\mathbf{k}}] - \bar{N} \cdot (1 - \pi(\mathbf{0})) \right| &= \left| \sum_{j=1}^J \sum_{\mathbf{k} \neq \mathbf{0}} \mathbb{E}[X_{\mathbf{k}}^j] - \sum_{j=1}^J p^j \bar{N} \cdot (1 - \pi^j(\mathbf{0})) \right| \\ &\leq \sum_{j=1}^J \left| \sum_{\mathbf{k} \neq \mathbf{0}} \mathbb{E}[X_{\mathbf{k}}^j] - \lceil p^j \bar{N} \rceil \cdot (1 - \pi^j(\mathbf{0})) \right| + O(1) \\ &= O(\sqrt{r}). \end{aligned}$$

Here we use (2.52) and the relationship between $\pi(\mathbf{k})$ and $\pi^j(\mathbf{k})$. Similarly,

$$\begin{aligned} &\left| \sum_{\mathbf{k} \neq \mathbf{0}} h(\mathbf{k}) \mathbb{E}[X_{\mathbf{k}}] - \bar{N} \cdot \sum_{j=1}^J \sum_{\mathbf{k} \neq \mathbf{0}} h(\mathbf{k}) \pi(\mathbf{k}) \right| \\ &= \left| \sum_{j=1}^J \sum_{\mathbf{k} \neq \mathbf{0}} \mathbb{E}[X_{\mathbf{k}}^j] - \sum_{j=1}^J p^j \bar{N} \cdot \sum_{\mathbf{k} \neq \mathbf{0}} h(\mathbf{k}) \pi^j(\mathbf{k}) \right| \\ &= \sum_{j=1}^J \left| \sum_{\mathbf{k} \neq \mathbf{0}} \mathbb{E}[X_{\mathbf{k}}^j] - \lceil p^j \bar{N} \rceil \cdot \sum_{\mathbf{k} \neq \mathbf{0}} h(\mathbf{k}) \pi^j(\mathbf{k}) \right| + O(1) \\ &= O(\sqrt{r}). \end{aligned}$$

After re-indexing the servers, we get (2.5) and (2.6). The bounds on $N(\sigma)$ and $C(\sigma)$, (2.7) and (2.8), follow from (2.5) and (2.6). They can be verified in the same way as that in the proof for the irreducible case, so we omit the argument here. $\square$

2.C Solving the single-server problem

In this section, we show the equivalence of the single-server problem in (2.3) and a linear program (LP) as stated in Theorem 2.4. The equivalence needs to be proved in two directions. In Appendix 2.C.1, we first derive the linear program (2.61) as a relaxation of the single-server problem (2.3) so that the optimal value of the LP is a lower bound to the optimal value of (2.3). Then

in Appendix 2.C.2, we will construct a Markovian single-server policy that achieves the optimal value of the LP, which implies the optimality of the policy.

2.C.1 Lower bound via LP relaxation

In this subsection we derive an LP relaxation of the optimization problem (2.3), restated below:

$$\begin{aligned} & \underset{\bar{N}, \bar{\sigma}, \pi}{\text{minimize}} && \bar{N} \\ & \text{subject to} && \mathbb{E} [h(\bar{\mathbf{K}}(\infty)) | \bar{\mathbf{K}}(\infty) \neq \mathbf{0}] \leq \epsilon, \\ & && \bar{N} \cdot \bar{\lambda}_i = \lambda_i r, \quad \forall i \in \mathcal{I}. \end{aligned} \tag{2.3 revisited}$$

Here we allow $\bar{\sigma}$ to be non-Markovian, i.e., it can make decisions based on the history, but we still require its performance metrics used in the objective and the constraints of (2.3) to have well-defined steady-state distributions.

Observe that both $\bar{\mathbf{K}}(\infty)$ and $\bar{\lambda}_i$ depend on the stationary distribution π , but the constraints in terms of π are implicit. To derive an LP relaxation, we give an explicit characterization of the constraints that must be satisfied by the stationary distribution π induced by any feasible policy $\bar{\sigma}$.

To do this, we derive a version of the stationary equation in terms of a quantity called *transition frequency*. The transition frequency of type i jobs is a function $u_i: \mathcal{K} \rightarrow \mathbb{R}$ describing the steady-state frequency of requesting a type i job when the system has configuration $\mathbf{k}$. To rigorously define transition frequency, we first introduce a concept called *nominal transition*.

Definition 2.11 (Nominal Transition). Consider a single-server system under any policy. When the configuration $\bar{\mathbf{K}}(t)$ transitions from $\mathbf{k}$ to $\mathbf{k}' + \mathbf{a}$ for some $\mathbf{k}, \mathbf{k}' \in \mathcal{K}$ with $\mathbf{a} = (a_i)_{i \in \mathcal{I}}$ new jobs added into service, we decompose the transition by adding intermediate configurations as illustrated below, where $\mathbf{k}$ first goes to $\mathbf{k}'$ if $\mathbf{k}' \neq \mathbf{k}$, then add jobs of each type one by one.

$$\mathbf{k} \rightarrow \mathbf{k}' \rightarrow (\mathbf{k}' + \mathbf{e}_{i_1}) \rightarrow \cdots \rightarrow (\mathbf{k}' + a_{i_1} \mathbf{e}_{i_1}) \rightarrow \cdots \rightarrow (\mathbf{k}' + a_{i_1} \mathbf{e}_{i_1} + \cdots + a_{i_{|\mathcal{I}|}} \mathbf{e}_{i_{|\mathcal{I}|}}),$$

where $(i_1, i_2, \dots, i_{|\mathcal{I}|})$ is a fixed ordering of the set of phases $\mathcal{I}$. We call each short transition in the diagram a *nominal transition*.

For $\mathbf{k}^1, \mathbf{k}^2 \in \mathcal{K}$, we denote $F(\mathbf{k}^1, \mathbf{k}^2, t)$ as the cumulative number of nominal transitions from $\mathbf{k}^1$ to $\mathbf{k}^2$ during the time interval $[0, t]$, which is a

random variable with a distribution depending on the single-server policy and initial distribution of configurations.

Note that for any $i \in \mathcal{I}$ and $\mathbf{k} \in \mathcal{K}$ s.t. $k_i \geq 1$, $F(\mathbf{k}, \mathbf{k} - \mathbf{e}_i, t)$ counts the number of times that a type i job departs when being in configuration $\mathbf{k}$. As a result,

$$F(\mathbf{k}, \mathbf{k} - \mathbf{e}_i, t) = \mathcal{N} \left(\int_0^t k_i \mu_{i\perp} \mathbb{1}_{\{K(s)=\mathbf{k}\}} ds \right),$$

where $\mathcal{N}(t)$ denotes a unit rate Poisson process. If we take expectation, divide both sides by t , and let $t \rightarrow \infty$, we have

$$\lim_{t \rightarrow \infty} \frac{1}{t} \mathbb{E}[F(\mathbf{k}, \mathbf{k} - \mathbf{e}_i, t)] = k_i \mu_{i\perp} \cdot \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \mathbb{P}(\mathbf{K}(s) = \mathbf{k}) ds = k_i \mu_{i\perp} \pi(\mathbf{k}). \quad (2.54)$$

Similarly, $F(\mathbf{k}, \mathbf{k} - \mathbf{e}_i + \mathbf{e}_{i'}, t)$ counts the number of times a job in phase i transitions to phase i' when being in configuration $\mathbf{k}$ for any $i, i' \in \mathcal{I}$, $\mathbf{k} \in \mathcal{K}$ s.t. $i' \neq i$ and $k_i \geq 1$, so

$$\lim_{t \rightarrow \infty} \frac{1}{t} \mathbb{E}[F(\mathbf{k}, \mathbf{k} - \mathbf{e}_i + \mathbf{e}_{i'}, t)] = k_i \mu_{ii'} \pi(\mathbf{k}). \quad (2.55)$$

We define transition frequency as follows.

Definition 2.12 (Transition Frequency). Transition frequency of type i jobs at state $\mathbf{k}$ is the long-run average number of nominal transitions from configuration $\mathbf{k}$ to $\mathbf{k} + \mathbf{e}_i$ per unit time,

$$u_i(\mathbf{k}) \triangleq \lim_{t \rightarrow \infty} \frac{1}{t} \mathbb{E}[F(\mathbf{k}, \mathbf{k} + \mathbf{e}_i, t)]. \quad (2.56)$$

The transition frequencies allow us to derive the following version of the stationary equation.

Lemma 2.13 (Stationary Equation). *Under any policy, the stationary distribution π and the transition frequency u_i satisfy the following equation:*

$$\begin{aligned} & \sum_i u_i(\mathbf{k} - \mathbf{e}_i) \mathbb{1}_{\{k_i \geq 1\}} + \sum_i (k_i + 1) \mu_{i\perp} \pi(\mathbf{k} + \mathbf{e}_i) \\ & + \sum_{i, i': i \neq i'} (k_i + 1) \mu_{ii'} \pi(\mathbf{k} + \mathbf{e}_i - \mathbf{e}_{i'}) \mathbb{1}_{\{k_{i'} \geq 1\}} \\ & = \sum_i u_i(\mathbf{k}) + \left(\sum_i k_i \mu_{i\perp} + \sum_{i, i': i \neq i'} k_i \mu_{ii'} \right) \pi(\mathbf{k}) \end{aligned} \quad (2.57)$$

for any state $\mathbf{k} \in \mathcal{K}$, and $\sum_i, \sum_{i, i'}$ are shorthand for $\sum_{i \in \mathcal{I}}, \sum_{i, i' \in \mathcal{I}}$.

Proof. For each configuration $\mathbf{k} \in \mathcal{K}$, if we look at the difference between the number of nominal transitions into configuration $\mathbf{k}$ and that out of configuration $\mathbf{k}$ by time t , we have the following equation,

$$\begin{aligned}
& \mathbb{1}_{\{\bar{\mathbf{K}}(t)=\mathbf{k}\}} - \mathbb{1}_{\{\bar{\mathbf{K}}(0)=\mathbf{k}\}} \\
&= \sum_i F(\mathbf{k} - \mathbf{e}_i, \mathbf{k}, t) \mathbb{1}_{\{k_i \geq 1\}} + \sum_i F(\mathbf{k} + \mathbf{e}_i, \mathbf{k}, t) \\
&\quad + \sum_{i, i': i \neq i'} F(\mathbf{k} + \mathbf{e}_i - \mathbf{e}_{i'}, \mathbf{k}, t) \mathbb{1}_{\{k_{i'} \geq 1\}} \\
&\quad - \sum_i F(\mathbf{k}, \mathbf{k} + \mathbf{e}_i, t) - \sum_i F(\mathbf{k}, \mathbf{k} - \mathbf{e}_i, t) \mathbb{1}_{\{k_i \geq 1\}} \\
&\quad - \sum_{i, i': i \neq i'} F(\mathbf{k}, \mathbf{k} - \mathbf{e}_i + \mathbf{e}_{i'}, t) \mathbb{1}_{\{k_{i'} \geq 1\}}
\end{aligned} \tag{2.58}$$

By Definition 2.12 and (2.54)-(2.55), if we divide both sides of (2.58) by t and let $t \rightarrow \infty$, we get the stationary equation in (2.57). $\square$

Since the stationary equation in (2.57) is linear in $u_i(\mathbf{k})$ and $\pi(\mathbf{k})$, we can write it in matrix form:

$$A\boldsymbol{\pi} + \sum_{i \in \mathcal{I}} B_i \mathbf{u}_i = 0, \tag{2.59}$$

where $\boldsymbol{\pi}$ and $\mathbf{u}_i$ are column vectors representing $\pi(\cdot)$, $u_i(\cdot)$, and A, B_i are matrices that make (2.59) equivalent to (2.57). Therefore, the following three conditions are necessary for any tuple $(\boldsymbol{\pi}, (\mathbf{u}_i)_{i \in \mathcal{I}})$ to be a possible pair of stationary distribution and transition frequencies for a Markovian policy.

$$\begin{aligned}
& A\boldsymbol{\pi} + \sum_{i \in \mathcal{I}} B_i \mathbf{u}_i = 0 \\
& \sum_{\mathbf{k}} \pi(\mathbf{k}) = 1 \\
& \boldsymbol{\pi}, \mathbf{u}_i \geq 0, \quad \forall i \in \mathcal{I}
\end{aligned} \tag{2.60}$$

Based on the characterization of stationary $\boldsymbol{\pi}$ and $\mathbf{u}_i$'s in (2.60), we can now formulate a linear program $\overline{\mathcal{LP}}((\lambda_i)_{i \in \mathcal{I}}, \epsilon)$. The linear program has decision

variables $\Phi \in \mathbb{R}$, $\boldsymbol{\pi} \in \mathbb{R}^{\mathcal{K}}$, and $\mathbf{u}_i \in \mathbb{R}^{\mathcal{K}}$ for $i \in \mathcal{I}$, where Φ is a factor that scales the throughput of each type of jobs in the direction of $(\lambda_i)_{i \in \mathcal{I}}$.

$$\begin{aligned}
& \underset{\Phi, \boldsymbol{\pi}, (\mathbf{u}_i)_{i \in \mathcal{I}}}{\text{maximize}} && \Phi \\
& \text{subject to} && \mathbf{h}^T \boldsymbol{\pi} \leq \epsilon(1 - \pi(\mathbf{0})) \\
& && \mathbf{1}_o^T \mathbf{u}_i = \Phi \cdot \lambda_i \quad \forall i \\
& && A\boldsymbol{\pi} + \sum_{i \in \mathcal{I}} B_i \mathbf{u}_i = \mathbf{0} \\
& && \mathbf{1}^T \boldsymbol{\pi} = 1 \\
& && \boldsymbol{\pi} \geq 0, \mathbf{u}_i \geq 0 \quad \forall i \in \mathcal{I}
\end{aligned} \tag{2.61}$$

where $\mathbf{h}$ is the vector form of the cost rate function h ; $\mathbf{1}_o^T$ is a $|\mathcal{K}|$ -dimensional vector with one in all entries except those with $\sum_{i \in \mathcal{I}} k_i = K_{\max}$. In addition to the last three constraints on stationarity, the first constraint of $\overline{\mathcal{LP}}((\lambda_i)_{i \in \mathcal{I}}, \epsilon)$ is the resource contention constraint; the second constraint is because the transition frequency from $\mathbf{k}$ to $\mathbf{k} + \mathbf{e}_i$ is equal to the throughput of type i jobs.

By the construction of $\overline{\mathcal{LP}}((\lambda_i)_{i \in \mathcal{I}}, \epsilon)$, any feasible solution of $\overline{\mathcal{P}}((\lambda_i r)_{i \in \mathcal{I}}, \epsilon)$ can be converted to a feasible solution of $\overline{\mathcal{LP}}((\lambda_i)_{i \in \mathcal{I}}, \epsilon)$. Let $\overline{N}^*$ be the optimal value of (2.3) and Φ^* be the optimal value of (2.61). Then we have

$$\overline{N}^* \geq \frac{r}{\Phi^*}. \tag{2.62}$$

2.C.2 Policy construction

In this subsection, we describe a procedure that allows us to construct a policy that achieves the lower bound given by the LP relaxation in (2.61). Specifically, given a feasible solution $(\boldsymbol{\pi}, (\mathbf{u}_i)_{i \in \mathcal{I}})$ to (2.61), we define an *LP-based policy* that requests jobs as follows:

- Case 1: When the system enters a configuration $\mathbf{k}$ with $\pi(\mathbf{k}) \neq 0$, for each $i \in \mathcal{I}$, the policy starts a timer whose duration follows an exponential distribution with rate $\frac{u_i(\mathbf{k})}{\pi(\mathbf{k})}$. The policy requests a type i job when the i -th timer ticks. When the configuration changes, all timers are canceled.
- Case 2: When the system enters a configuration $\mathbf{k}$ with $\pi(\mathbf{k}) = 0$ and $\sum_{i'} u_{i'}(\mathbf{k}) \neq 0$, the policy immediately requests a type i job with probability $\frac{u_i(\mathbf{k})}{\sum_{i'} u_{i'}(\mathbf{k})}$.

- Case 3: When the system enters a configuration $\mathbf{k}$ with $\pi(\mathbf{k}) = 0$ and $\sum_{i'} u_{i'}(\mathbf{k}) = 0$, the policy does not request any jobs.

We denote the LP-based policy based on the solution $(\boldsymbol{\pi}, (\mathbf{u}_i)_{i \in \mathcal{I}})$ as $\bar{\sigma}(\boldsymbol{\pi}, (\mathbf{u}_i)_{i \in \mathcal{I}})$.

Remark 2.14. Note that the definition of the LP-based policy here is stated from a view different from the view in Section 2.4.2: here each request only adds one job to the server, and one request can happen immediately after another; while in Section 2.4.2 each request can add multiple jobs to the server, and there is only one request happening at a time. We refer to the view here as the *impulsive view*, because multiple requests happening at the same time can be thought of as having an infinite request rate. In contrast, we call the view in Section 2.4.2 the *non-impulsive view*.

The LP-based policy can be alternatively described using the non-impulsive view if we see multiple requests happening at the same time as one request that adds multiple jobs to the server. More specifically, each reactive request of the LP-based policy is initiated by an internal transition or departure and consists of one or multiple requests of Case 2; each proactive request of the LP-based policy consists of one request in Case 1 and possibly several requests in Case 2.

The following lemma characterizes the steady-state behavior of a single-server system under the LP-based policy.

Lemma 2.15 (Properties of LP-based Policies). *Consider a single-server system under the LP-based policy $\bar{\sigma}(\boldsymbol{\pi}, (\mathbf{u}_i)_{i \in \mathcal{I}})$, where $(\boldsymbol{\pi}, (\mathbf{u}_i)_{i \in \mathcal{I}})$ is a feasible solution to (2.61). We have that $\boldsymbol{\pi}$ is a stationary distribution under policy $\bar{\sigma}$, and $(\mathbf{u}_i)_{i \in \mathcal{I}}$ are the transition frequencies corresponding to $\bar{\sigma}$.*

The proof of Lemma 2.15 is based on (2.58), following the same argument as the proof of Lemma 2.13, as well as an induction argument.

Proof. Let $(\boldsymbol{\pi}, (\mathbf{u}_i)_{i \in \mathcal{I}})$ be a feasible solution to the LP in (2.61). To show that $\boldsymbol{\pi}$ and $(\mathbf{u}_i)_{i \in \mathcal{I}}$ are also the actual stationary distribution and transition frequencies, it suffices to show that if the initial distribution follows $\boldsymbol{\pi}$, i.e.

$$\mathbb{P}(\bar{\mathbf{K}}(0) = \mathbf{k}) = \pi(\mathbf{k})$$

then under the policy $\bar{\sigma}(\boldsymbol{\pi}, (\mathbf{u}_i)_{i \in \mathcal{I}})$, we have

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \mathbb{P}(\bar{\mathbf{K}}(t) = \mathbf{k}) dt = \pi(\mathbf{k}), \quad (2.63)$$

$$\lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}[F(\mathbf{k}, \mathbf{k} + \mathbf{e}_i, T)] = u_i(\mathbf{k}), \quad (2.64)$$

where F is the cumulative number of nominal transitions under the policy $\bar{\sigma}(\boldsymbol{\pi}, (\mathbf{u}_i)_{i \in \mathcal{I}})$ and the initial distribution $\boldsymbol{\pi}$.

Our proof is based on the following equation:

$$\begin{aligned} & \frac{d}{dt} \mathbb{P}(\bar{\mathbf{K}}(t) = \mathbf{k}) \Big|_{t=0} \\ &= \sum_i \frac{d}{dt} \mathbb{E}[F(\mathbf{k} - \mathbf{e}_i, \mathbf{k}, t)] \Big|_{t=0} - \sum_i \frac{d}{dt} \mathbb{E}[F(\mathbf{k}, \mathbf{k} + \mathbf{e}_i, t)] \Big|_{t=0} \\ & \quad + \sum_i (k_i + 1) \mu_{i\perp} \pi(\mathbf{k} + \mathbf{e}_i) + \sum_{i, i'} (k_i + 1) \mu_{ii'} \pi(\mathbf{k} + \mathbf{e}_i - \mathbf{e}_{i'}) \mathbb{1}_{\{k_{i'} \geq 1\}} \\ & \quad - \sum_i k_i \mu_{i\perp} \pi(\mathbf{k}) - \sum_{i, i'} k_i \mu_{ii'} \pi(\mathbf{k}), \end{aligned} \quad (2.65)$$

The equation is a straightforward consequence of (2.58), following the same argument as the proof of Lemma 2.13.

We prove the following two equations by induction on $\sum_{i \in \mathcal{I}} k_i$.

$$\lim_{t \rightarrow 0} \frac{1}{t} \mathbb{E}[F(\mathbf{k}, \mathbf{k} + \mathbf{e}_i, t)] = u_i(\mathbf{k}), \quad (2.66)$$

$$\frac{d}{dt} \mathbb{P}(\bar{\mathbf{K}}(t) = \mathbf{k}) \Big|_{t=0} = 0. \quad (2.67)$$

We first consider the base case when $\sum_{i \in \mathcal{I}} k_i = 0$. In this case, $\mathbf{k} = \mathbf{0}$ and we have

$$\frac{d}{dt} \mathbb{E}[F(\mathbf{k} - \mathbf{e}_i, \mathbf{k}, t)] \Big|_{t=0} = u_i(\mathbf{k} - \mathbf{e}_i) \mathbb{1}_{\{k_i \geq 1\}} = 0, \quad (2.68)$$

for all i . This reduces (2.65) to

$$\begin{aligned} & \frac{d}{dt} \mathbb{P}(\bar{\mathbf{K}}(t) = \mathbf{k}) \Big|_{t=0} \\ &= \sum_i u(\mathbf{k} - \mathbf{e}_i) \mathbb{1}_{\{k_i \geq 1\}} - \sum_i \frac{d}{dt} \mathbb{E}[F(\mathbf{k}, \mathbf{k} + \mathbf{e}_i, t)] \Big|_{t=0} \\ & \quad + \sum_i (k_i + 1) \mu_{i\perp} \pi(\mathbf{k} + \mathbf{e}_i) + \sum_{i, i'} (k_i + 1) \mu_{ii'} \pi(\mathbf{k} + \mathbf{e}_i - \mathbf{e}_{i'}) \mathbb{1}_{\{k_{i'} \geq 1\}} \\ & \quad - \sum_i k_i \mu_{i\perp} \pi(\mathbf{k}) - \sum_{i, i'} k_i \mu_{ii'} \pi(\mathbf{k}), \end{aligned} \quad (2.69)$$

Now we discuss based on whether $\pi(\mathbf{k}) = 0$. If $\pi(\mathbf{k}) \neq 0$, by the definition of our policy, for all i ,

$$\left. \frac{d}{dt} \mathbb{E} [F(\mathbf{k}, \mathbf{k} + \mathbf{e}_i, t)] \right|_{t=0} = \frac{u_i(\mathbf{k})}{\pi(\mathbf{k})} \cdot \mathbb{P}(\overline{\mathbf{K}}(0) = \mathbf{k}) = u_i(\mathbf{k}), \quad (2.70)$$

which is (2.66). Combining the above equation and the stationary equation (2.57) satisfied by $(\boldsymbol{\pi}, (\mathbf{u}_i)_{i \in \mathcal{I}})$, we conclude that the RHS of (2.69) is zero, i.e.

$$\left. \frac{d}{dt} \mathbb{P}(\overline{\mathbf{K}}(t) = \mathbf{k}) \right|_{t=0} = 0,$$

which is (2.67). For the case when $\pi(\mathbf{k}) = 0$ and $\sum_i u_i(\mathbf{k}) \neq 0$, because the system immediately leave the configuration $\mathbf{k}$ after reaching it through a nominal transition,

$$\left. \frac{d}{dt} \mathbb{P}(\overline{\mathbf{K}}(t) = \mathbf{k}) \right|_{t=0} = 0,$$

i.e., the LHS of (2.69) is 0. Again we compare (2.69) against the stationary equation (2.57) and get

$$\sum_i \left. \frac{d}{dt} \mathbb{E} [F(\mathbf{k}, \mathbf{k} + \mathbf{e}_i, t)] \right|_{t=0} = \sum_i u_i(\mathbf{k}).$$

By the definition of our policy, we have

$$\begin{aligned} \left. \frac{d}{dt} \mathbb{E} [F(\mathbf{k}, \mathbf{k} + \mathbf{e}_i, t)] \right|_{t=0} &= \frac{u_i(\mathbf{k})}{\sum_{i'} u_{i'}(\mathbf{k})} \cdot \sum_{i'} \left. \frac{d}{dt} \mathbb{E} [F(\mathbf{k}, \mathbf{k} + \mathbf{e}_{i'}, t)] \right|_{t=0} \\ &= u_i(\mathbf{k}). \end{aligned} \quad (2.71)$$

which is (2.66). For the case when $\pi(\mathbf{k}) = 0$ and $\sum_i u_i(\mathbf{k}) = 0$, (2.57) implies that $u_i(\mathbf{k} - \mathbf{e}_i) = 0$, $\pi(\mathbf{k} + \mathbf{e}_i) = 0$, $\pi(\mathbf{k} + \mathbf{e}_i - \mathbf{e}_{i'}) = 0$ for any i . Then (2.69) is further reduced to

$$\left. \frac{d}{dt} \mathbb{P}(\overline{\mathbf{K}}(t) = \mathbf{k}) \right|_{t=0} = - \sum_i \left. \frac{d}{dt} \mathbb{E} [F(\mathbf{k}, \mathbf{k} + \mathbf{e}_i, t)] \right|_{t=0}.$$

Because $\mathbb{P}(\overline{\mathbf{K}}(t) = \mathbf{k}) \geq 0$ and $\mathbb{P}(\overline{\mathbf{K}}(0) = \mathbf{k}) = 0$, the LHS of the above expression is non-negative. However, the RHS of the above expression is non-positive. Therefore, both sides are equal to zero, thus we have $\left. \frac{d}{dt} \mathbb{P}(\overline{\mathbf{K}}(t) = \mathbf{k}) \right|_{t=0} = 0$ and $\left. \frac{d}{dt} \mathbb{E} [F(\mathbf{k}, \mathbf{k} + \mathbf{e}_i, t)] \right|_{t=0} = u_i(\mathbf{k})$.

Having proved the base case, we do the induction step. Suppose we have proved (2.66) and (2.67) for all $\mathbf{k}$ such that $\sum_{i \in \mathcal{I}} k_i \leq m - 1$ for some integer $m \geq 1$. We consider $\mathbf{k}$ with $\sum_{i \in \mathcal{I}} k_i = m$. By the induction hypothesis,

$$\left. \frac{d}{dt} \mathbb{E}[F(\mathbf{k} - \mathbf{e}_i, \mathbf{k}, t)] \right|_{t=0} = u_i(\mathbf{k} - \mathbf{e}_i) \mathbf{1}_{\{k_i \geq 1\}}. \quad (2.72)$$

Then we repeat the arguments after (2.68) of the base case verbatim. By induction, we have proved (2.66) and (2.67).

Therefore, given the policy and initial distribution, the distribution of $\overline{\mathbf{K}}(t)$ is stationary, i.e., we always have $\mathbb{P}(\overline{\mathbf{K}}(t) = \mathbf{k}) = \pi(\mathbf{k})$ for all $\mathbf{k} \in \mathcal{K}$. As a result, an analogue of (2.66) holds for all $t \geq 0$: $\mathbb{E}[F(\mathbf{k}, \mathbf{k} + \mathbf{e}_i, t)]$ is differentiable with respect to t and

$$\frac{d}{dt} \mathbb{E}[F(\mathbf{k}, \mathbf{k} + \mathbf{e}_i, t)] = u_i(\mathbf{k}),$$

for all $\mathbf{k} \in \mathcal{K}$ and all $i \in \mathcal{I}$. Therefore,

$$\begin{aligned} \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \mathbb{P}(\overline{\mathbf{K}}(t) = \mathbf{k}) dt &= \lim_{T \rightarrow \infty} \frac{1}{T} \cdot \pi(\mathbf{k}) \cdot T = \pi(\mathbf{k}), \\ \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}[F(\mathbf{k}, \mathbf{k} + \mathbf{e}_i, T)] &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \frac{d}{dt} \mathbb{E}[F(\mathbf{k}, \mathbf{k} + \mathbf{e}_i, t)] dt \\ &= \lim_{T \rightarrow \infty} \frac{1}{T} \cdot u_i(\mathbf{k}) \cdot T = u_i(\mathbf{k}). \end{aligned}$$

This completes the proof. $\square$

2.C.3 Proof of Theorem 2.4

Theorem 2.4 (Optimality of Single-OPT). *Given an optimal solution $(\Phi^*, \boldsymbol{\pi}^*, (\mathbf{u}_i)_{i \in \mathcal{I}})$ to the linear program $\overline{\mathcal{LP}}((\lambda_i)_{i \in \mathcal{I}}, \epsilon)$, we can solve the single-server problem $\overline{\mathcal{P}}((\lambda_i r)_{i \in \mathcal{I}}, \epsilon)$ in (2.3) to achieve an optimal value r/Φ^* , using the optimal policy $\bar{\sigma}(\boldsymbol{\pi}^*, (\mathbf{u}_i^*)_{i \in \mathcal{I}})$ and the optimal stationary distribution $\boldsymbol{\pi}^*$. Moreover, the optimal policy $\bar{\sigma}(\boldsymbol{\pi}^*, (\mathbf{u}_i^*)_{i \in \mathcal{I}})$ is a Markovian policy.*

Proof. By Lemma 2.15, under the policy $\bar{\sigma}(\boldsymbol{\pi}^*, (\mathbf{u}_i^*)_{i \in \mathcal{I}})$, $\boldsymbol{\pi}^*$ is a stationary distribution, and $(\mathbf{u}_i^*)_{i \in \mathcal{I}}$ are the corresponding transition frequencies. Recall the single-server $\bar{\mathcal{P}}((\lambda_i r)_{i \in \mathcal{I}}, \epsilon)$ problem

$$\begin{aligned} & \underset{\bar{N}, \bar{\sigma}, \boldsymbol{\pi}}{\text{minimize}} && \bar{N} \\ & \text{subject to} && \mathbb{E} [h(\bar{\mathbf{K}}(\infty)) | \bar{\mathbf{K}}(\infty) \neq \mathbf{0}] \leq \epsilon, \\ & && \bar{N} \cdot \bar{\lambda}_i = \lambda_i r, \quad \forall i \in \mathcal{I}. \end{aligned} \tag{2.3 revisited}$$

Observe that under the policy $\bar{\sigma}(\boldsymbol{\pi}^*, (\mathbf{u}_i^*)_{i \in \mathcal{I}})$, the cost rate of resource contention is $\mathbf{h}^T \boldsymbol{\pi} \leq \epsilon(1 - \pi(\mathbf{0}))$, the request rate of type i jobs is $\bar{\lambda}_i = \mathbf{1}_o^T \mathbf{u}_i^* = \Phi^* \cdot \lambda_i$, so

$$\begin{aligned} \mathbb{E} [h(\bar{\mathbf{K}}(\infty)) | \bar{\mathbf{K}}(\infty) \neq \mathbf{0}] &= \frac{\mathbf{h}^T \boldsymbol{\pi}}{1 - \pi(\mathbf{0})} \leq \epsilon, \\ \bar{\lambda}_i &= \Phi^* \cdot \lambda_i, \quad \forall i \in \mathcal{I}. \end{aligned}$$

where we have used the fact that $h(\mathbf{0}) = 0$ in the first equality. Therefore, $(\Phi^*/r, \bar{\sigma}(\boldsymbol{\pi}^*, (\mathbf{u}_i^*)_{i \in \mathcal{I}}), \boldsymbol{\pi})$ is a feasible solution to the single-server problem $\bar{\mathcal{P}}((\lambda_i r)_{i \in \mathcal{I}}, \epsilon)$, achieving the objective value of r/Φ^* , which is the optimal value because $r/\Phi^* \leq \bar{N}^*$. $\square$

2.D Performance guarantee of JRS with an estimated model

2.D.1 Assumptions and result

In this section, we consider the performance of JRS when it is based on an estimated model. We will state the performance guarantee in terms of the estimation error, and give a proof sketch by pointing out which part of Theorem 2.3's proof needs to be changed accordingly.

Consider the setting where the maximal jobs on a server $K_{\max}$, the set of job phases $\mathcal{I}$, the cost rate function $h(\cdot)$, and the budget ϵ are all known. However, we only have estimations of the jobs' arrival rates, internal transition rates, and departure rates.

Specifically, for any $i, i' \in \mathcal{I}$, let $\tilde{\mu}_{ii'}$ be the *true* rate of internal transition from phase i to phase i' ; let $\tilde{\mu}_{i\perp}$ be the *true* departure rate of phase i ; let

$\tilde{\lambda}_i r$ be the *true* arrival rate of type i jobs. We let $\mu_{ii'}$'s, $\mu_{i\perp}$'s, and λ_i 's be the *estimated* internal transition rates, departure rates, and arrival rates, respectively. We assume that there exists a small positive constant δ that is independent of the scaling factor r such that the following assumptions hold.

Assumption 2.16 (δ -accurate estimation). For any $i, i' \in \mathcal{I}$,

$$|\mu_{ii'} - \tilde{\mu}_{ii'}| \leq \delta, \quad (2.73)$$

$$|\mu_{i\perp} - \tilde{\mu}_{i\perp}| \leq \delta, \quad (2.74)$$

$$|\lambda_i - \tilde{\lambda}_i| \leq \delta. \quad (2.75)$$

Assumption 2.17 (Scaling of the single-server objective value). Consider the single-server problem in (2.3). Let $(\bar{N}, \bar{\sigma}, \pi)$ be a solution feasible to the single-server problem with the estimated parameters that are δ -accurate, where $\delta \in [0, \delta_{\max})$, for some constant $\delta_{\max} > 0$. We assume that there exist constants $0 < m_1 < m_2$ independent of δ and r such that

$$m_1 r \leq \bar{N} \leq m_2 r.$$

Assumption 2.18 (δ -insensitivity of the optimal value). Consider the single-server problem in (2.3). Let the optimal value of the single-server problem with the estimated parameters be $\bar{N}^*$, where the estimated parameters are δ -accurate; let the optimal value of the single-server problem with true parameters be $\bar{N}_{\text{true}}^*$. We assume that

$$\bar{N}^* \leq \bar{N}_{\text{true}}^* + \delta r.$$

We also assume that JRS can accurately simulate the virtual jobs.

Assumption 2.19 (Accurate simulation). The virtual jobs simulated in JRS follow the true transition dynamics.

Given the above assumptions, we have the following proposition that states the optimality gap of JRS policy with estimated model parameters, which has a similar form as Theorem 2.3 for JRS under true model parameters.

Proposition 2.20 (Optimality gap with model estimation). *Consider a stochastic bin-packing problem in service systems with time-varying job resource requirements. Let the infinite-server policy σ be JRS with subroutine $\bar{\sigma}$. Suppose σ is specified based on an estimated model satisfying Assumption 2.16, 2.17, 2.18, and 2.19, for δ s.t. $\delta \in [0, \delta_{\max})$, where $\delta_{\max}$ is some positive constant independent of r . Let $\bar{N}$ be the objective value achieved by $\bar{\sigma}$ in the*

single-server problem $\overline{\mathcal{P}}((\lambda_i r)_{i \in \mathcal{I}}, \epsilon)$ with estimated parameters. Under σ , for any initial state, we have

$$\left| \sum_{\mathbf{k} \neq \mathbf{0}} \mathbb{E}[X_{\mathbf{k}}] - \lceil N \rceil \cdot \mathbb{P}(\overline{\mathbf{K}} \neq \mathbf{0}) \right| = O(\sqrt{r}) + \delta \cdot O(r), \quad (2.76)$$

$$\left| \sum_{\mathbf{k} \neq \mathbf{0}} h(\mathbf{k}) \mathbb{E}[X_{\mathbf{k}}] - \lceil N \rceil \cdot \mathbb{E}[h(\overline{\mathbf{K}})] \right| = O(\sqrt{r}) + \delta \cdot O(r), \quad (2.77)$$

where $\overline{\mathbf{K}}$ denotes the steady-state configurations of the single-server system under $\overline{\sigma}$ with the estimated parameters. If we let $\overline{\sigma}$ be the optimal policy of the single-server problem with estimated parameters, for any initial state, we have

$$N(\sigma) \leq (1 + B\delta + O(r^{-0.5})) \cdot \overline{N}_{true}^*, \quad (2.78)$$

$$C(\sigma) \leq (1 + B\delta + O(r^{-0.5})) \cdot \epsilon, \quad (2.79)$$

where B is a positive constant independent of r . In other words, σ is $(1 + B\delta + O(r^{-0.5}), 1 + B\delta + O(r^{-0.5}))$ -optimal.

Remark 2.21. Note even if the single-server system with estimated parameters $\mathbf{k}^0$ -irreducible under $\overline{\sigma}$, it is hard to guarantee that the original system is $\mathbf{k}^0$ -irreducible due to the estimation errors. Consequently, the steady-state performance metrics $N(\sigma)$ and $C(\sigma)$ could depend on the system's initial state. Fortunately, our proof for Theorem 2.3 does not rely on the uniqueness of the stationary distribution, so we can adapt the proof to show the inequalities in Proposition 2.20 for any initial state.

Remark 2.22. Note that Assumption 2.19 ensures that the real jobs and virtual jobs on a server are indistinguishable, so that $(\widehat{\mathbf{K}}^{1:L}(t), \boldsymbol{\eta}^{1:L}(t))$ is still a Markov chain. Recall that $\widehat{\mathbf{K}}^\ell$ and $\boldsymbol{\eta}^\ell$ denote the observed configuration and tokens for each normal server ℓ , respectively (see Section 2.5.1). We suspect that Assumption 2.19 can be removed since our proof does not rely too much on $(\widehat{\mathbf{K}}^{1:L}(t), \boldsymbol{\eta}^{1:L}(t))$ being a Markov chain. However, removing the assumption requires a more careful and notationally heavy analysis. We argue that in practice, this assumption is not restrictive, as one can record the traces of the jobs that arrived in the past and resample virtual jobs from those traces.

2.D.2 Lemmas

In this section, we give a proof sketch for Proposition 2.20 when the single-server system under $\bar{\sigma}$ with estimated parameters is $\mathbf{k}^0$ -irreducible. The argument of extending to the general case is essentially the same as Appendix 2.B.3, so we omit it here.

On a high level, the proof of Proposition 2.20 is similar to that of Theorem 2.1. In particular, the key steps of the proof are to show that $d(\mathbf{K}^{1:L}, \bar{\mathbf{K}}^{1:L}) = O(\sqrt{r})$ and $\sum_{\ell=L+1}^{\infty} \sum_{i \in \mathcal{I}} K_i^\ell = O(\sqrt{r})$ as $r \rightarrow \infty$, where $L = \lceil \bar{N} \rceil$, and $\bar{\mathbf{K}}^{1:L}$ are i.i.d. copies of steady-state configurations of the single-server system under $\bar{\sigma}$ with the estimated parameters, and $\mathbf{K}$ is the steady-state configurations of the infinite server system under σ .

Proposition 2.20 is based on the three lemmas stated below. The proof of Proposition 2.20 using the three lemmas is essentially the same as the argument in Section 2.5.3.

Lemma 2.23. *Under the conditions of Proposition 2.20 and the single-server system with estimated parameters under $\bar{\sigma}$ being $\mathbf{k}^0$ -irreducible, for any initial state, we have*

$$d(\widehat{\mathbf{K}}^{1:L}, \bar{\mathbf{K}}^{1:L}) = O(\sqrt{r}) + \delta \cdot O(r)$$

Lemma 2.24. *Under the conditions of Proposition 2.20 and the single-server system with estimated parameters under $\bar{\sigma}$ being $\mathbf{k}^0$ -irreducible, for any initial state and $i \in \mathcal{I}$, the steady-state expected number of virtual jobs of type i s.t.*

$$\mathbb{E} \left[\sum_{\ell=1}^L \zeta_i^\ell \right] = O(\sqrt{r}) + \delta \cdot O(r).$$

Lemma 2.25. *Under the conditions of Proposition 2.20 and the single-server system with estimated parameters under $\bar{\sigma}$ being $\mathbf{k}^0$ -irreducible, for any initial state and $i \in \mathcal{I}$, the steady-state expected number of type i jobs on backup servers s.t.*

$$\mathbb{E} \left[\sum_{\ell=L+1}^{\infty} K_i^\ell \right] = O(\sqrt{r}) + \delta \cdot O(r).$$

These three lemmas are analogous to Lemma 2.6, Lemma 2.7, and Lemma 2.8, respectively. In the rest of the section, we sketch the proofs for the three lemmas above, highlighting the difference from the proofs of their analogues.

2.D.3 Proof sketch for Lemma 2.23

Recall from Section 2.5.4 that the proof of Lemma 2.6 is based on Stein's method, which compares the generator of the i.i.d. copies of the single-server system with the generator of the infinite-server system. To write down the generators with the estimated model, recall that in the single-server system, each transition can be represented by the diagram

$$\mathbf{k} \rightarrow \mathbf{k}' \rightarrow \mathbf{k}' + \mathbf{a},$$

where the arrow $\mathbf{k} \rightarrow \mathbf{k}'$ denotes an internal transition or a departure if $\mathbf{k} \neq \mathbf{k}'$; the arrow $\mathbf{k}' \rightarrow \mathbf{k}' + \mathbf{a}$ denotes a job request that is made right after reaching $\mathbf{k}'$. For any $\mathbf{k}$, let $E(\mathbf{k})$ be the set of $(\mathbf{k}', \mathbf{a}) \in \mathcal{K}^2$ such that $\mathbf{k}' + \mathbf{a} \in \mathcal{K}$. We define two sets of transition rates as below: for any $\mathbf{k} \in \mathcal{K}$ and $(\mathbf{k}', \mathbf{a}) \in E(\mathbf{k})$,

- Under the estimated parameters and the policy $\bar{\sigma}$, we let $\gamma_{\mathbf{k},(\mathbf{k}',\mathbf{a})}$ be the rate of the transition $\mathbf{k} \rightarrow \mathbf{k}' \rightarrow \mathbf{k}' + \mathbf{a}$, and let $\gamma_{\mathbf{k}} \triangleq \sum_{\mathbf{k}',\mathbf{a}} \gamma_{\mathbf{k},(\mathbf{k}',\mathbf{a})}$ be the total transition rate at configuration $\mathbf{k}$.
- Under the *true parameters and the policy* $\bar{\sigma}$, we let $\tilde{\gamma}_{\mathbf{k},(\mathbf{k}',\mathbf{a})}$ be the rate of the transition $\mathbf{k} \rightarrow \mathbf{k}' \rightarrow \mathbf{k}' + \mathbf{a}$, and let $\tilde{\gamma}_{\mathbf{k}} \triangleq \sum_{\mathbf{k}',\mathbf{a}} \tilde{\gamma}_{\mathbf{k},(\mathbf{k}',\mathbf{a})}$ be the total transition rate at configuration $\mathbf{k}$.

Let $\bar{G}$ be the generator of $L = \lceil \bar{N} \rceil$ i.i.d. copies of single-server systems under the estimated parameters. For any $g: \mathcal{K}^L \rightarrow \mathbb{R}$, we have

$$\bar{G}g(\mathbf{k}^{1:L}) = \sum_{\ell=1}^L \sum_{\mathbf{k}',\mathbf{a}} \gamma_{\mathbf{k}^\ell,(\mathbf{k}',\mathbf{a})} (g(\cdot, \mathbf{k}' + \mathbf{a}, \cdot) - g(\cdot, \mathbf{k}^\ell, \cdot)), \quad (2.80)$$

where $\sum_{\mathbf{k}',\mathbf{a}}$ is a shorthand for $\sum_{(\mathbf{k}',\mathbf{a}) \in E(\mathbf{k})}$. Let $\hat{G}$ be the generator of $(\widehat{\mathbf{K}}^{1:L}(t), \boldsymbol{\eta}^{1:L}(t))$ for the infinite-server system. For any function $g: \mathcal{K}^L \rightarrow \mathbb{R}$ and $\psi(\mathbf{k}^{1:L}, \boldsymbol{\eta}^{1:L}) = g(\mathbf{k}^{1:L} + \boldsymbol{\eta}^{1:L})$, we have

$$\begin{aligned} \hat{G}\psi(\mathbf{k}^{1:L}, \boldsymbol{\eta}^{1:L}) &= \sum_{\ell=1}^L \sum_{\mathbf{k}',\mathbf{a}} \tilde{\gamma}_{\mathbf{k}^\ell,(\mathbf{k}',\mathbf{a})} (g(\cdot, \mathbf{k}' + \mathbf{a}, \cdot) - g(\cdot, \mathbf{k}^\ell, \cdot)) \mathbb{1}_{\{\boldsymbol{\eta}^\ell = \mathbf{0}\}} \\ &\quad + \sum_{\ell=1}^L \sum_{\mathbf{k}',\mathbf{a}} \tilde{\gamma}_{\mathbf{k}^\ell,(\mathbf{k}',\mathbf{a})} (g(\cdot, \mathbf{k}' + \boldsymbol{\eta}^\ell, \cdot) - g(\cdot, \mathbf{k}^\ell + \boldsymbol{\eta}^\ell, \cdot)) \mathbb{1}_{\{\boldsymbol{\eta}^\ell \neq \mathbf{0}\}}. \end{aligned} \quad (2.81)$$

To prove Lemma 2.23, we need to show that for any $f \in \text{Lip}(1)$, $\mathbf{k}^{1:L} \in \mathcal{K}^L$, and $\boldsymbol{\eta}^{1:L} \in \mathcal{K}^L$,

$$\overline{G}g_f(\mathbf{k}^{1:L} + \boldsymbol{\eta}^{1:L}) - \widehat{G}\psi_f(\mathbf{k}^{1:L}, \boldsymbol{\eta}^{1:L}) = O(\sqrt{r}) + \delta \cdot O(r), \quad (2.82)$$

where g_f is the solution to

$$\mathbb{E} \left[f(\overline{\mathbf{K}}^{1:L}) \right] - f(\mathbf{k}^{1:L}) = \overline{G}g_f(\mathbf{k}^{1:L}), \quad (2.83)$$

and $\psi_f(\mathbf{k}^{1:L}, \boldsymbol{\eta}^{1:L}) = g_f(\mathbf{k}^{1:L} + \boldsymbol{\eta}^{1:L})$. Same as the proof of Lemma 2.6, we prove (2.82) in two steps: the generator comparison step, and the Stein factor bound step.

In the generator comparison step, we observe that the formula of $\overline{G}g(\mathbf{k}^{1:L})$ and $\widehat{G}\psi(\mathbf{k}^{1:L}, \boldsymbol{\eta}^{1:L})$ in (2.80) and (2.81) look almost the same as (2.15) and (2.17), except that the rates in (2.81) are $\tilde{\gamma}_{\mathbf{k}^\ell, (\mathbf{k}', \mathbf{a})}$ instead of $\gamma_{\mathbf{k}^\ell, (\mathbf{k}', \mathbf{a})}$. As a result, after carrying out similar calculations as in the poof of Lemma 2.6, we get an extra error term involving $\gamma_{\mathbf{k}^\ell, (\mathbf{k}', \mathbf{a})} - \tilde{\gamma}_{\mathbf{k}^\ell, (\mathbf{k}', \mathbf{a})}$, which can be bounded using the lemma below. This error term results in the $\delta \cdot O(r)$ in (2.82).

Lemma 2.26. *Under Assumption 2.16, for any $\mathbf{k} \in \mathcal{K}$ and $(\mathbf{k}', \mathbf{a}) \in E(\mathbf{k})$, we have*

$$|\tilde{\gamma}_{\mathbf{k}, (\mathbf{k}', \mathbf{a})} - \gamma_{\mathbf{k}, (\mathbf{k}', \mathbf{a})}| \leq K_{\max} \delta. \quad (2.84)$$

Proof. When $\mathbf{k} = \mathbf{k}'$, $\gamma_{\mathbf{k}, (\mathbf{k}', \mathbf{a})}$ and $\tilde{\gamma}_{\mathbf{k}, (\mathbf{k}', \mathbf{a})}$ are both the rate of adding $\mathbf{a}$ jobs via a proactive request under the policy $\bar{\sigma}$, so they are identical.

When $\mathbf{k} \neq \mathbf{k}'$ and $\mathbf{a} = \mathbf{0}$, $\gamma_{\mathbf{k}, (\mathbf{k}', \mathbf{a})}$ is equal to the rate of going from $\mathbf{k}$ to $\mathbf{k}'$ via an internal transition or departure under the estimated job model, while $\tilde{\gamma}_{\mathbf{k}, (\mathbf{k}', \mathbf{a})}$ is under the true job model. Because there are at most $K_{\max}$ jobs, and by our assumption the estimation error of each job's transition rates are bounded by δ , thus $\gamma_{\mathbf{k}, (\mathbf{k}', \mathbf{a})}$ and $\tilde{\gamma}_{\mathbf{k}, (\mathbf{k}', \mathbf{a})}$ differ by at most $K_{\max} \delta$.

When $\mathbf{k} \neq \mathbf{k}'$ and $\mathbf{a} \neq \mathbf{0}$, $\gamma_{\mathbf{k}, (\mathbf{k}', \mathbf{a})}$ is equal to the rate of going from $\mathbf{k}$ to $\mathbf{k}'$ via an internal transition or departure, multiplied by the probability of adding $\mathbf{a}$ jobs via a reactive request, under the estimated job model and the policy $\bar{\sigma}$; $\tilde{\gamma}_{\mathbf{k}, (\mathbf{k}', \mathbf{a})}$ is under the true job model instead of the estimated job model, but uses the same policy. Because the rate of going from $\mathbf{k}$ to $\mathbf{k}'$ differs by at most $K_{\max} \delta$ under two different job models, and the probability of adding $\mathbf{a}$ jobs after going from $\mathbf{k}$ to $\mathbf{k}'$ is the same under the same policy, thus $\gamma_{\mathbf{k}, (\mathbf{k}', \mathbf{a})}$ and $\tilde{\gamma}_{\mathbf{k}, (\mathbf{k}', \mathbf{a})}$ differ by at most $K_{\max} \delta$. $\square$

In the Stein factor bound step, we need to show that

$$\sup_{\mathbf{k}, \mathbf{k}' \in \mathcal{K}} |g_f(\cdot, \mathbf{k}', \cdot) - g_f(\cdot, \mathbf{k}, \cdot)| = O(1). \quad (2.85)$$

This involves analyzing the i.i.d. copies of the single-server system, and the fact that the single-server system with estimated parameters is $\mathbf{k}^0$ -irreducible under $\bar{\sigma}$. This part of the proof is identical to the corresponding part of the proof of Lemma 2.6.

2.D.4 Proof sketch for Lemma 2.24 and Lemma 2.25

The proof of Lemma 2.24 and Lemma 2.25 has a similar structure as the proof of Lemma 2.7 and Lemma 2.8. In the first step, we use Little's law to bound expectations of the number of type i virtual jobs, V_i , and the number of type i jobs on backup servers, Y_i . We have two equations almost the same as (2.36) and (2.37) except that the rates $\gamma_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})}$ and $\lambda_i r$ are replaced by $\widetilde{\gamma}_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})}$ and $\widetilde{\lambda}_i r$:

$$\mathbb{E}[V_i] \leq t_{\max} \mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \widetilde{\gamma}_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} dv_i \mathbb{1}_{\{\boldsymbol{\eta}^\ell = \mathbf{0}\}} \right], \quad (2.86)$$

$$\mathbb{E}[Y_i] \leq t_{\max} \mathbb{E} [\widetilde{\lambda}_i r \cdot dy_i], \quad (2.87)$$

where $t_{\max}$ is the maximal expected service time of any type of virtual job or real job. Because $t_{\max} = O(1)$, it suffices to bound $\mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \widetilde{\gamma}_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} dv_i \mathbb{1}_{\{\boldsymbol{\eta}^\ell = \mathbf{0}\}} \right]$ and $\mathbb{E} [\widetilde{\lambda}_i r \cdot dy_i]$.

In the second step, we utilize the fact that the two Lyapunov functions $g(z_i) = z_i$ and $g(z_i) = z_i^2$ have zero drift in steady-state, where z_i is the total number of type i tokens. We get two equalities similar to (2.38) and (2.41):

$$\mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \widetilde{\gamma}_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} (a_i - dv_i) \mathbb{1}_{\{\boldsymbol{\eta}^\ell = \mathbf{0}\}} + \widetilde{\lambda}_i r (-1 + dy_i) \right] = 0. \quad (2.88)$$

$$\mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \widetilde{\gamma}_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} dv_i \mathbb{1}_{\{\boldsymbol{\eta}^\ell = \mathbf{0}\}} \right] \quad (2.89)$$

$$= \frac{1}{\eta_{\max}} \cdot \mathbb{E} \left[\left(\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \tilde{\gamma}_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i \mathbb{1}_{\{\boldsymbol{\eta}^\ell = \mathbf{0}\}} - \tilde{\lambda}_i r \right) \cdot Z_i \right] \quad (2.90)$$

$$+ \frac{1}{2\eta_{\max}} \cdot \mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \tilde{\gamma}_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} (a_i^2 - (dv_i)^2) \mathbb{1}_{\{\boldsymbol{\eta}^\ell = \mathbf{0}\}} + \tilde{\lambda}_i r \cdot (1 - (dy_i)^2) \right]. \quad (2.91)$$

In the third step, we use the above two equalities to bound $\mathbb{E} [\tilde{\lambda}_i r \cdot dy_i]$ and $\mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \tilde{\gamma}_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} dv_i \mathbb{1}_{\{\boldsymbol{\eta}^\ell = \mathbf{0}\}} \right]$. We first use the equality in (2.89) to (2.91) to bound $\mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \tilde{\gamma}_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} dv_i \mathbb{1}_{\{\boldsymbol{\eta}^\ell = \mathbf{0}\}} \right]$. Following the same argument as the proof of Lemma 2.7 and Lemma 2.8 until (2.47), we can show that

$$(2.91) \leq O(\sqrt{r}),$$

$$\begin{aligned} (2.90) &\leq \mathbb{E} \left[\left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \tilde{\gamma}_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i \mathbb{1}_{\{\boldsymbol{\eta}^\ell = \mathbf{0}\}} - \tilde{\lambda}_i r \right| \right] \\ &\leq \mathbb{E} \left[\left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \tilde{\gamma}_{\widehat{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i - \tilde{\lambda}_i r \right| \right] + O(\sqrt{r}) \\ &\leq \mathbb{E} \left[\left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \tilde{\gamma}_{\overline{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i - \tilde{\lambda}_i r \right| \right] + O(\sqrt{r}) + \delta \cdot O(r), \end{aligned} \quad (2.92)$$

where to get (2.92), we apply Lemma 2.23 to replace $\widehat{\mathbf{K}}^\ell$ with $\overline{\mathbf{K}}^\ell$, which causes an $O(\sqrt{r}) + \delta \cdot O(r)$ error. Next, we show that

$$\mathbb{E} \left[\left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \tilde{\gamma}_{\overline{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i - \tilde{\lambda}_i r \right| \right] = O(\sqrt{r}) + \delta \cdot O(r). \quad (2.93)$$

By Assumption 2.16 and Lemma 2.26, we have

$$|\tilde{\lambda}_i r - \lambda_i r| \leq \delta r, \quad (2.94)$$

$$\left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \tilde{\gamma}_{\overline{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i - \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\overline{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i \right| \leq \delta \cdot O(r). \quad (2.95)$$

These two bounds allow us to replace $\tilde{\gamma}_{\bar{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})}$ and $\tilde{\lambda}_i r$ on the LHS of (2.93) with $\gamma_{\bar{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})}$ and $\lambda_i r$ at the cost of introducing $\delta \cdot O(r)$ error. Moreover, because $\{\bar{\mathbf{K}}^\ell\}_{\ell=1, \dots, L}$ are i.i.d., $\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\bar{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i$ concentrates around its mean with $O(\sqrt{r})$ error, where the mean can be shown to be

$$\mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\bar{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i \right] = \bar{\lambda}_i L = \lambda_i r + O(1).$$

Note that the first equality in (2.96) holds because $\mathbb{E}[\sum_{\mathbf{k}', \mathbf{a}} \gamma_{\bar{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i] = \bar{\lambda}_i$ for each ℓ , i.e., $\bar{\lambda}_i$ is the long-run average rate of requesting type i jobs on a single-server system with estimated parameters under $\bar{\sigma}$; the second equality in (2.96) is because $L = \lceil \bar{N} \rceil$, and $\bar{N} \bar{\lambda}_i = \lambda_i r$. As a result,

$$\mathbb{E} \left[\left| \sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \gamma_{\bar{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} a_i - \lambda_i r \right| \right] \leq O(\sqrt{r}). \quad (2.96)$$

Combining (2.94) to (2.96), we get (2.93). Therefore,

$$\begin{aligned} \mathbb{E}[V_i] &\leq t_{\max} \mathbb{E} \left[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \tilde{\gamma}_{\bar{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} dv_i \mathbf{1}_{\{\eta^\ell = \mathbf{0}\}} \right] \\ &\leq O(1) \cdot ((2.90) + (2.91)) \\ &\leq O(\sqrt{r}) + \delta \cdot O(r), \end{aligned}$$

which completes the proof of Lemma 2.24.

Finally, it is straightforward to show that $\mathbb{E}[\tilde{\lambda}_i r \cdot dy_i] = O(\sqrt{r}) + \delta \cdot O(r)$ using (2.88) and the bound on $\mathbb{E}[\sum_{\ell=1}^L \sum_{\mathbf{k}', \mathbf{a}} \tilde{\gamma}_{\bar{\mathbf{K}}^\ell, (\mathbf{k}', \mathbf{a})} dv_i \mathbf{1}_{\{\eta^\ell = \mathbf{0}\}}]$. Therefore, $\mathbb{E}[Y_i] \leq t_{\max} \mathbb{E}[\tilde{\lambda}_i r \cdot dy_i] = O(\sqrt{r}) + \delta \cdot O(r)$. This proves Lemma 2.25.

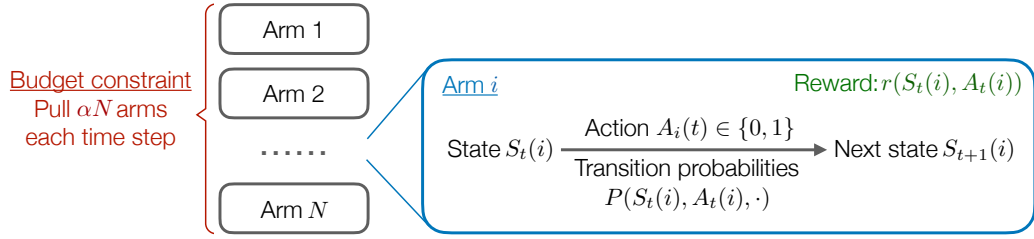
Chapter 3

Restless Bandits with Average Reward: Breaking the Global Attractor Assumption

3.1 Introduction

The restless bandit (RB) problem is a dynamic decision-making problem that involves a number of Markov decision processes (MDPs) coupled by a constraint. Each MDP, referred to as an arm, has a binary action space, {passive, active}. At every decision epoch, the decision maker is constrained to select a fixed number of arms to activate, with the goal of maximizing the expected reward accrued. We illustrate the problem in Figure 3.1. The RB problem has a rich history and finds applications across a spectrum of domains, including wireless communication [3], congestion control [23], inventory control [13], crawling web content [22], machine maintenance [115], clinical trials [269], to name a few. We refer the reader to the recent survey paper [218] for a comprehensive overview of the literature.

Solving for an optimal policy for the RB problem is known to be hard [220]. However, it is possible to find *asymptotically optimal* policies in a computationally efficient manner in the regime where the number of arms, N , grows large. A policy is said to be asymptotically optimal if its *optimality gap* is *diminishing* as $N \rightarrow \infty$, where the optimality gap is the difference between the average reward per arm achieved by an optimal policy and that achieved by this policy. This large N regime, introduced in the seminal

Figure 3.1: The restless bandit problem with N arms.

papers on the renowned Whittle index policy [274, 283], has recently regained significant attention. There has been a growing body of work that proposes new policies and provides refined analysis of their optimality gaps, not only in the infinite-horizon average-reward setting but also in the finite-horizon total-reward setting [56–58, 87, 108, 109, 111, 155, 296, 300] and the infinite-horizon discounted-reward setting [58, 111, 301]. We discuss related work in more detail in Appendix 3.A

Prior work and the global attractor property assumption. Prior work has studied the celebrated Whittle index policy [283] and LP-Priority policies [266] and established sufficient conditions for their asymptotic optimality [107, 108, 266, 274]. One key assumption underpinning all prior work is the *global attractor property* (GAP)—also known as globally asymptotic stability—which pertains to the mean-field/fluid limit of the restless bandit system in the asymptotic limit $N \rightarrow \infty$. GAP stipulates that the system’s state distribution in the mean-field limit converges to the optimal state distribution attaining the maximum reward, from any initial distribution. GAP is a technical condition known to be difficult to verify for a given RB instance and policy, due to the non-linearity of the mean-field dynamics. Moreover, there are documented RB instances where the Whittle index and LP-Priority policies fail to satisfy GAP and are strictly asymptotically suboptimal [105, 274]; see also Figures 3.2 and 3.3. In Section 4.8.2, we further examine several natural classes of randomly generated RB instances; we observe that when the transition kernels and the reward function are generated from some sparse distributions, the percentage of non-GAP instances could be as high as 20%.

More recent work studies the *rate* at which the optimality gap converges to zero. Gast et al. [107, 108] prove a striking $O(\exp(-cN))$ optimality gap for the Whittle index policy and LP-Priority policies, respectively, where c

	Paper	Policy	Optimality Gap	Conditions*
Discrete-time setting	[107]	Whittle Index	$O(\exp(-cN))$	UGAP & Non-singular
	[108]	LP-Priority	$O(\exp(-cN))$	UGAP & Non-degenerate
	Our work	FTVA ($\bar{\pi}^*$)	$O(1/\sqrt{N})$	SA
Continuous-time setting	[274]	Whittle Index	$o(1)$	GAP
	[266]	LP-Priority	$o(1)$	GAP
	[107]	Whittle Index	$O(\exp(-cN))$	UGAP & Non-singular
	[108]	LP-Priority	$O(\exp(-cN))$	UGAP & Non-degenerate
	Our work	FTVA-CT ($\bar{\pi}^*$)	$O(1/\sqrt{N})$	—

Table 3.1: Optimality gap results and conditions. *All papers require the standard unichain assumption.

is a constant. They rely on a stronger version of GAP, called the *uniform global attractor property* (UGAP), which requires the mean-field dynamics to converge to the optimal state distribution *uniformly* over all initial points.¹ In addition to UGAP, these results also rely on a non-singularity or non-degeneracy condition. We are not aware of any rate-of-convergence result in prior work that does not assume UGAP or non-singularity/non-degeneracy. See Table 3.1 for a summary.

Therefore, prior work on average-reward restless bandits leaves two fundamental questions open:

1. Is it possible to achieve asymptotic optimality without GAP?
2. Is it possible to establish a non-trivial convergence rate for the optimality gap in the absence of the non-singular/non-degenerate assumption (and UGAP)?

Our contributions. We consider both the discrete-time and continuous-time settings of the average-reward restless bandit problem. We propose a general, simulation-based framework, **Follow-the-Virtual-Advice** (FTVA) and its continuous-time variant FTVA-CT, which convert any single-armed policy into a policy for the original N -armed problem, with a vanishing

¹In the published conference version of this work [149], we did not distinguish between GAP and UGAP and referred to both as UGAP throughout; the present version corrects this.

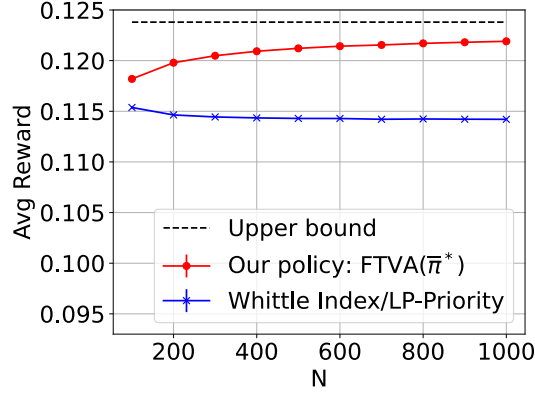


Figure 3.2: A discrete-time RB problem that satisfies SA but not GAP.

performance loss. Our framework can be instantiated to produce a policy with an $O(1/\sqrt{N})$ optimality gap, under the conditions summarized in Table 3.1, which we elaborate on later.

Under our framework, computing an asymptotically optimal policy is efficient since it reduces to deriving an optimal single-armed policy, whose complexity is independent of N . The resultant policy can be implemented with a linear-in- N computational cost and some of its subroutines can be implemented in a distributed fashion over the arms (see Appendix 3.B for more details).

Our results can also be extended to RBs with heterogeneous arms. See Appendix 3.H for details.

We now elaborate on the conditions in Table 3.1. In the discrete-time setting, our result holds under a condition called the Synchronization Assumption (SA), in addition to the standard unichain assumption required by all prior work. The SA condition, which is imposed on the MDP associated with a single arm, admits several intuitive sufficient conditions. While it is unclear whether SA subsumes GAP, we show that there exist problem instances that violate GAP but satisfy SA. Figure 3.2 shows one such instance (constructed by [105], described in Appendix 3.G for ease of reference), in which the Whittle Index and LP-Priority policies coincide and have a non-diminishing optimality gap, whereas our policy, named as **FTVA** ($\bar{\pi}^*$), is asymptotically optimal. In addition, our $O(1/\sqrt{N})$ bound on the optimality gap does not require a non-singularity/non-degeneracy assumption.

More notably, in the continuous-time setting, we *completely eliminate the GAP assumption*. We show that our policy FTVA-CT ($\bar{\pi}^*$) achieves an $O(1/\sqrt{N})$ optimality gap under only the standard unichain assumption (which is required by all prior work).

In both settings, our results are the first asymptotic optimality results that do not require GAP. We point out that GAP is considered necessary for LP-Priority policies [108, 266]. Our policy, FTVA ($\bar{\pi}^*$), is not a priority policy. As such, we hope our results open up new directions for developing new classes of RB algorithms that achieve asymptotic optimality without relying on GAP.

Intuitions. Our algorithm and many existing approaches solve an LP relaxation of the original problem. The solution of the LP induces a policy and gives an “ideal” distribution for the system state. Existing approaches directly apply the LP policy to the current system state, and then perform a simple rounding of the resulting actions so as to satisfy the budget constraint. When the current system state is far from the ideal distribution, the actual actions after rounding may deviate substantially from the LP solution and thus, in the absence of GAP, would fail to drive the system to the optimum.

Our FTVA framework, in contrast, prioritizes constraint satisfaction. We apply the LP solution to the state of a *simulated* system, which is constructed carefully so that the resulting actions satisfy the constraint after a minimal amount of rounding. Consequently, starting from any initial state, our policy steers the system towards the ideal distribution and hence approximates the optimal value.

This method is inspired by the approach in Chapter 2 [150], which also involves a simulated system, but for a stochastic bin-packing problem.

Other related work. Another condition extensively discussed in the RB literature is the indexability condition [283], which is necessary for the Whittle index policy to be well-defined, but not required by LP-Priority policies [266]. However, indexability alone does not guarantee the asymptotic optimality of Whittle index.

So far we have discussed prior work on infinite-horizon RBs with average reward. For the related setting of finite-horizon total reward RBs, a line of recent work has established an $O(1/\sqrt{N})$ optimality gap [56, 111, 155, 296], and an $O(\exp(-cN))$ gap assuming non-degeneracy [108, 300]. For the

infinite-horizon discounted reward setting, [111, 301] propose policies with $O(1/\sqrt{N})$ optimality gap without assuming indexability or GAP. While these results are not directly comparable to ours, it is of future interest to see if our simulation-based framework can be applied to their settings. For a more detailed discussion of prior work, see Appendix 3.A.

Chapter organization. While our continuous-time result is stronger, the discrete-time setting is more accessible. Therefore, we first discuss the discrete-time setting, which includes the problem statement in Section 3.3, our proposed framework, **Follow-the-Virtual-Advice**, in Section 3.4, and our results on the optimality gap in Section 3.5. Results for the continuous-time setting are presented in Section 3.6. We conclude the chapter in Section 3.7. Proofs and additional discussion are given in the appendices.

3.2 Additional related work

3.3 Problem setup

Consider the infinite-horizon, discrete-time restless bandit problem with N arms indexed by $[N] \triangleq \{1, 2, \dots, N\}$. Each arm is associated with an MDP described by the tuple $(\mathbb{S}, \mathbb{A}, P, r)$. Here $\mathbb{S}$ is the state space, assumed to be finite; $\mathbb{A} = \{0, 1\}$ is the action set, and we say the arm is *activated* or *pulled* when action 1 is applied; $P : \mathbb{S} \times \mathbb{A} \times \mathbb{S} \rightarrow [0, 1]$ is the transition kernel, where $P(s, a, s')$ is the probability of transitioning from state s to state s' upon taking action a ; $r = \{r(s, a)\}_{s \in \mathbb{S}, a \in \mathbb{A}}$ is the reward function, where $r(s, a)$ is the reward for taking action a in state s . Throughout the chapter, we assume that the transition kernel P is unichain [230]; that is, under any Markov policy for this single-armed MDP $(\mathbb{S}, \mathbb{A}, P, r)$, the induced Markov chain has a single recurrent class. The unichain assumption is standard in most existing work on restless bandits [107, 108, 266, 274]. We will discuss relaxing the unichain assumption in Appendix 3.D.

In the above setting, we are subject to a *budget constraint* that exactly αN arms must be activated in each time step, where $\alpha \in (0, 1)$ is a given constant and αN is assumed to be an integer for simplicity. This N -armed RB problem can be represented by the tuple $(N, \mathbb{S}^N, \mathbb{A}^N, P, r, \alpha N)$.

A policy π for the N -armed problem specifies the action for each of the N arms based on the history of states and actions. Under policy π , let $S_i^\pi(t) \in \mathbb{S}$

denote the state of the i th arm at time t , and we call $\mathbf{S}^\pi(t) \triangleq (S_i^\pi(t))_{i \in [N]} \in \mathbb{S}^N$ the *system state*. Similarly, let $A_i^\pi(t) \in \mathbb{A}$ denote the action applied to the i th arm at time t , and let $\mathbf{A}^\pi(t) \triangleq (A_i^\pi(t))_{i \in [N]} \in \mathbb{A}^N$ denote the joint action vector.

The controller's goal is to find a policy that maximizes the long-run average of the total expected reward from all N arms, subject to the budget constraint, assuming full knowledge of the model:

$$\begin{aligned} \underset{\text{policy } \pi}{\text{maximize}} \quad & V_N^\pi \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i=1}^N \mathbb{E}[r(S_i^\pi(t), A_i^\pi(t))] \end{aligned} \quad (3.1)$$

$$\text{subject to} \quad \sum_{i=1}^N A_i^\pi(t) = \alpha N, \quad \forall t \geq 0. \quad (3.2)$$

Under the unichain assumption, the value V_N^π of any policy π is independent of the initial state. Let $V_N^* \triangleq \sup_\pi V_N^\pi$ denote the optimal value. The optimality gap of π is defined as $V_N^* - V_N^\pi$. We say a policy π is *asymptotically optimal* if its optimality gap converges to 0 as $N \rightarrow \infty$.

Classical theory guarantees that for a finite-state Markov decision process like an RB, there exists an optimal policy that is Markovian and stationary [230]. Nevertheless, the policies we propose are not Markovian policies; rather, they have internal states. Under such a policy π , the system state $\mathbf{S}^\pi(t)$ together with the internal state form a Markov chain, and the action $\mathbf{A}^\pi(t)$ depends on both the system and internal states. We design a policy such that this Markov chain has a stationary distribution. Let $\mathbf{S}^\pi(\infty)$ and $\mathbf{A}^\pi(\infty)$ denote the random elements that follow the stationary distributions of $\mathbf{S}^\pi(t)$ and $\mathbf{A}^\pi(t)$, respectively. Then the average reward of π is equal to $V_N^\pi = \frac{1}{N} \sum_{i=1}^N \mathbb{E}[r(S_i^\pi(\infty), A_i^\pi(\infty))]$.

In later sections, when the context is clear, we drop the superscript π from S_i^π and A_i^π .

3.4 Follow-the-Virtual-Advice: A simulation-based framework

In this section, we present our framework, **Follow-the-Virtual-Advice (FTVA)**. We first describe a *single-armed problem*, which involves an ‘‘average arm’’ from the original N -armed problem. We then use the optimal single-armed policy to construct the proposed policy FTVA ($\bar{\pi}^*$).

3.4.1 Single-armed problem

The single-armed problem involves the MDP $(\mathbb{S}, \mathbb{A}, P, r)$ associated with a single arm (say arm 1 without loss of generality), where the budget is satisfied *on average*. Formally, consider the problem:

$$\begin{aligned} & \underset{\text{policy } \bar{\pi}}{\text{maximize}} & V_1^{\bar{\pi}} & \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[r(S_1^{\bar{\pi}}(t), A_1^{\bar{\pi}}(t))] \end{aligned} \quad (3.3)$$

$$\text{subject to} \quad \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[A_1^{\bar{\pi}}(t)] = \alpha. \quad (3.4)$$

The constraint (3.4) stipulates that the *average* rate of applying the active action must equal α . Various equivalent forms of this single-armed problem have been considered in prior work [107, 108, 266, 274].

By virtue of the unichain assumption, the single-armed problem can be equivalently rewritten as the following linear program, where each decision variable $y(s, a)$ represents the steady-state probability that the arm is in state s taking action a :

$$\begin{aligned} & \underset{\{y(s, a)\}_{s \in \mathbb{S}, a \in \mathbb{A}}}{\text{maximize}} & \sum_{s \in \mathbb{S}, a \in \mathbb{A}} r(s, a) y(s, a) \end{aligned} \quad (\text{LP})$$

$$\text{subject to} \quad \sum_{s \in \mathbb{S}} y(s, 1) = \alpha \quad (3.5)$$

$$\sum_{s' \in \mathbb{S}, a \in \mathbb{A}} y(s', a) P(s', a, s) = \sum_{a \in \mathbb{A}} y(s, a), \quad \forall s \in \mathbb{S} \quad (3.6)$$

$$\sum_{s \in \mathbb{S}, a \in \mathbb{A}} y(s, a) = 1, \quad y(s, a) \geq 0, \quad \forall s \in \mathbb{S}, a \in \mathbb{A}. \quad (3.7)$$

Here (3.5) corresponds to the relaxed budget constraint, (3.6) is the flow balance equation, and (3.6)–(3.7) guarantee that $y(s, a)$'s are valid steady-state probabilities.

By standard results for average reward MDPs [230], an optimal solution $\{y^*(s, a)\}_{s \in \mathbb{S}, a \in \mathbb{A}}$ to (LP) induces an optimal policy $\bar{\pi}^*$ for the single-armed problem via the following formula:

$$\bar{\pi}^*(a|s) = \begin{cases} y^*(s, a) / (y^*(s, 0) + y^*(s, 1)), & \text{if } y^*(s, 0) + y^*(s, 1) > 0, \\ 1/2, & \text{if } y^*(s, 0) + y^*(s, 1) = 0, \end{cases} \quad (3.8)$$

for $s \in \mathbb{S}$ and $a \in \mathbb{A}$. Let $V_1^{\text{rel}} = V_1^{\bar{\pi}^*}$ be the optimal value of (LP) and the single-armed problem.

(LP) can be viewed as a relaxation of the N -armed problem. To see this, take any N -armed policy π and set $y(s, a)$ to be the fraction of arms in state s taking action a in steady state under π , i.e., $y(s, a) = \mathbb{E}[\frac{1}{N} \sum_{i=1}^N \mathbb{1}_{\{S_i^\pi(\infty)=s, A_i^\pi(\infty)=a\}}]$. Whenever π satisfies the budget constraint (3.2), $\{y(s, a)\}$ satisfies the relaxed constraint (3.5) and the consistency constraints (3.6)–(3.7). Therefore, the optimal value of (LP) is an upper bound of the optimal value of the N -armed problem: $V_1^{\text{rel}} \geq V_N^*$.

3.4.2 Constructing the N -armed policy

We now present **Follow-the-Virtual-Advice**, a simulation-based framework for the N -armed problem. FTVA takes as input *any* single-armed policy $\bar{\pi}$ that satisfies the relaxed budget constraint (3.4) and converts it into a N -armed policy, denoted by $\text{FTVA}(\bar{\pi})$. Of particular interest is when $\bar{\pi}$ is an optimal single-armed policy, which leads to our result on the optimality gap. Below we introduce the general framework of FTVA without imposing any restriction on the input policy $\bar{\pi}$.

The proposed policy $\text{FTVA}(\bar{\pi})$ has two main components:

- *Virtual single-armed processes.* Each arm i simulates a *virtual* single-armed process, whose state is denoted as $\hat{S}_i(t)$, with action $\hat{A}_i(t)$ chosen according to $\bar{\pi}$. To make the distinction conspicuous, we sometimes refer to the state $S_i(t)$ and action $A_i(t)$ in the original N -armed problem as the *real* state/action. The virtual processes associated with different arms are independent.
- *Follow the virtual actions.* At each time step t , we choose the real actions $A_i(t)$'s to best match the virtual actions $\hat{A}_i(t)$'s, to the extent allowed by the budget constraint $\sum_{i=1}^N A_i(t) = \alpha N$.

FTVA is presented in detail in Algorithm 1. Note that we use an appropriate coupling in Algorithm 1 to ensure that the virtual processes $(\hat{S}_i(t), \hat{A}_i(t))$'s are independent and each follows the Markov chain induced by the single-armed policy $\bar{\pi}$. FTVA is designed to steer the real states to be close to the virtual states, thereby ensuring a small *conversion loss* $V_1^{\bar{\pi}} - V_N^{\text{FTVA}(\bar{\pi})}$. Here recall that $V_1^{\bar{\pi}}$ is the average reward achieved by the input policy $\bar{\pi}$ in the single-armed problem, and that $V_N^{\text{FTVA}(\bar{\pi})}$ is the average reward per arm achieved by policy $\text{FTVA}(\bar{\pi})$ in the N -armed problem.

Algorithm 1 Follow-the-Virtual-Advice (FTVA): A simulation-based framework

Input: N -armed problem $(N, \mathbb{S}^N, \mathbb{A}^N, P, r, \alpha N)$, initial states $\mathbf{S}(0)$, single-armed policy $\bar{\pi}$

Initialize: Virtual states $\widehat{\mathbf{S}}(0)$ are N i.i.d. samples following the stationary distribution of $\bar{\pi}$

```

1: for  $t = 0, 1, 2, \dots$  do
     $\triangleright$  Generate virtual actions
2:   Independently sample  $\widehat{A}_i(t) \leftarrow \bar{\pi}(\cdot | \widehat{S}_i(t))$  for each arm  $i \in [N]$ 
     $\triangleright$  Select a set  $\mathcal{A}$  of  $\alpha N$  arms to activate
3:   if  $\sum_{i=1}^N \widehat{A}_i(t) \geq \alpha N$  then
4:      $\mathcal{A} \leftarrow \alpha N$  arms chosen from  $\{i: \widehat{A}_i(t) = 1\}$  (any tie-breaking)
5:   else
6:      $\mathcal{B} \leftarrow \alpha N - \sum_{i=1}^N \widehat{A}_i(t)$  arms from  $\{i: \widehat{A}_i(t) = 0\}$  (any tie-breaking)
7:      $\mathcal{A} \leftarrow \{i: \widehat{A}_i(t) = 1\} \cup \mathcal{B}$ 
8:   Apply  $A_i(t) = 1$  and observe  $S_i(t+1)$  for each arm  $i \in \mathcal{A}$ 
9:   Apply  $A_i(t) = 0$  and observe  $S_i(t+1)$  for each arm  $i \notin \mathcal{A}$ 
     $\triangleright$  Progress virtual processes
10:  for  $i = 1, 2, 3, \dots N$  do
     $\triangleright$  Couple virtual and real states
11:    if  $\widehat{S}_i(t) = S_i(t)$  and  $\widehat{A}_i(t) = A_i(t)$  then
12:       $\widehat{S}_i(t+1) \leftarrow S_i(t+1)$ 
13:    else
14:      Independently sample  $\widehat{S}_i(t+1)$  from  $P(\widehat{S}_i(t), \widehat{A}_i(t), \cdot)$ 

```

3.4.3 Discussion on FTVA and the role of virtual processes

In this subsection, we provide insights into the mechanism of FTVA and explain the crucial role of the virtual processes. In particular, we contrast FTVA with the alternative approach of directly using the real states to choose actions, e.g., by applying the single-armed policy $\bar{\pi}^*$ to each arm's real state. We note that existing policies are essentially real-state-based, so the insights here can also explain why GAP is necessary for existing policies to have asymptotic optimality.

We first observe that the above two approaches are equivalent in the absence of the budget constraint. In particular, even if the initial virtual state and real state of an arm i are different, they will synchronize (i.e., become identical) in finite time *by chance* under mild assumptions (see Section 3.5.1). After this event, *if* there were no constraint, each arm i will consistently follow the virtual actions, i.e., $A_i(t) = \hat{A}_i(t) = \bar{\pi}(\hat{S}_i(t))$ for all t , and the virtual states will remain identical to real states.

In the presence of the budget constraint, the arms may not remain synchronized, so the virtual processes become crucial: they guarantee that the virtual actions $\hat{A}_i(t) = \bar{\pi}(\hat{S}_i(t))$ nearly satisfy budget constraint, which allows the real system to approximately follow the virtual actions to remain synchronized. To see this, note that regardless of the current real states, the N virtual states $\hat{S}_1, \dots, \hat{S}_N$ independently follow the single-armed policy $\bar{\pi}^*$, so, in the long run, each $(\hat{S}_i(\infty), \hat{A}_i(\infty))$ is distributed per $y^*(\cdot, \cdot)$, the optimal solution to (LP). For large N , the sum $\sum_{i=1}^N \hat{A}_i(\infty)$ concentrates around its expectation $N \sum_{s \in \mathbb{S}} y^*(s, 1) = \alpha N$ and thus tightly satisfy the budget constraint. In contrast, the actions generated by applying $\bar{\pi}$ to the real states are likely to significantly violate the constraint, especially when the empirical distribution of the current real states deviates from $y^*(\cdot, \cdot)$.

An example. We provide a concrete example illustrating the above arguments. Suppose the state space for each arm is $\mathbb{S} = \{0, 1, \dots, 7\}$. We label action 1 as the *preferred action* for states 0, 1, 2, 3, and action 0 for the other states. For an arm in state s , applying the preferred action moves the arm to state $(s + 1) \bmod 8$ with probability $p_{s,R}$, and applying the other

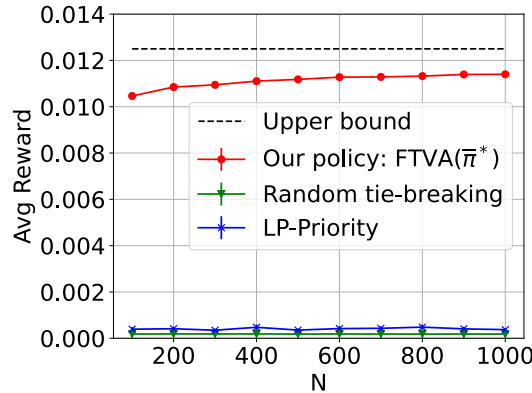


Figure 3.3: Comparing policies based on virtual states and real states.

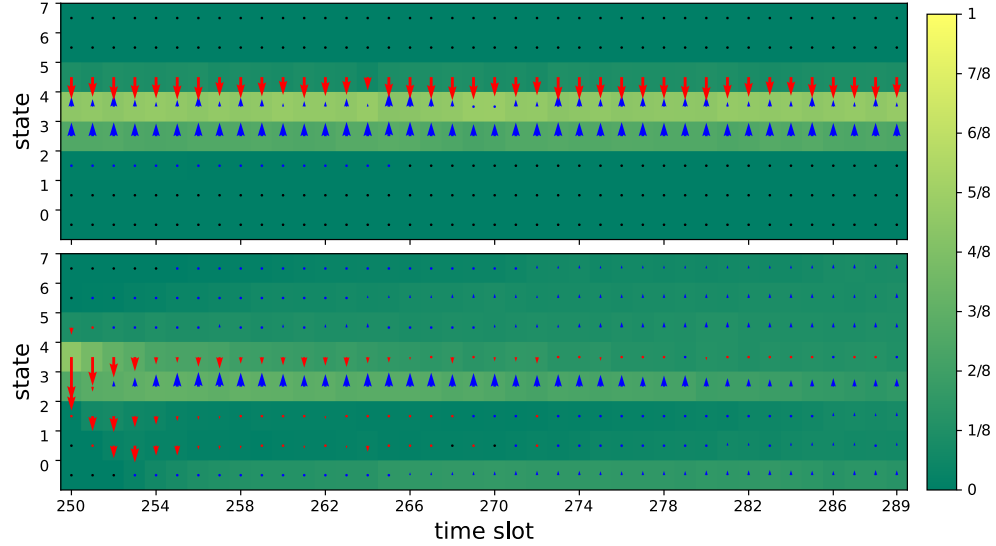


Figure 3.4: Time evolution of the fraction of arms in each state under LP-Priority (upper), or after switching to $\text{FTVA}(\bar{\pi}^*)$ (lower) since time slot 250. The x-axis represents the time slot, which ranges from 250 to 289; the y-axis represents the states; the color represents the fraction of arms in each state at each time slot. The colors and magnitudes of the arrows represent the average directions and rates at which the arms move away from each state.

action moves the arm to state $(s-1)^+$ with probability $p_{s,L}$; the arm stays at state s otherwise. One unit of reward is generated when the arm goes from state 7 to state 0. We assume that the budget is $N/2$ and set $\{p_{s,R}, p_{s,L}\}$ such that the optimal solution of (LP) is $y^*(s, 1) = 1/8$ for $s = 0, 1, 2, 3$ and $y^*(s, 0) = 1/8$ for $s = 4, 5, 6, 7$. That is, the optimal state distribution is $\text{uniform}(\mathbb{S})$, and the optimal single-armed policy $\bar{\pi}^*$ always takes the preferred action so as to traverse state 7 as often as possible.

To see why this MDP makes the corresponding N -armed system tricky to control, consider the situation where $p_{s,L} \gg p_{s,R}$ and most arms are in state 0. Those arms prefer actions 1 to move towards state 7. However, there are only $N/2$ units of budget. If we break ties uniformly at random, each arm is not pulled with probability $1/2$ in each time slot and is likely to return to 0 before they leave $\{0, 1, 2, 3\}$. This phenomenon can be seen from Figure 3.8b in Appendix 3.G.

A similar phenomenon can be observed for LP-Priority policies. Here we consider an LP-Priority in [56, 108, 155] that breaks ties based on the Lagrangian-optimal index. In Figure 3.4, we generate and visualize a sample path under LP-Priority from time 250 to 289, and contrast it with the sample path if the system switches to $\text{FTVA}(\bar{\pi}^*)$ from time 250 onwards. We can see that under LP-Priority, although the arms have a strong tendency to move up from state 4 to state 5, they move back when they reach state 5 and thus get stuck at state 4. In contrast, when switching to $\text{FTVA}(\bar{\pi}^*)$, the arrows gradually change to the correct direction, which helps the arms to escape from state 4 and converge to the uniform distribution over the state space. Intuitively, under $\text{FTVA}(\bar{\pi}^*)$, an increasing number of arms couple their real states with virtual states over time, which allows these arms to consistently apply the preferred actions afterward. In Appendix 3.G.4, we include more visualizations of the policies.

In Figure 3.3, we compare the average reward of $\text{FTVA}(\bar{\pi}^*)$ with those of the random tie-breaking policy and the LP-Priority policy discussed above. We can see that $\text{FTVA}(\bar{\pi}^*)$ is near-optimal, while the other two policies have nearly zero rewards. Further details are provided in Appendix 3.G. Note that while a different tie-breaking rule may solve this particular example with real states, currently there is no known rule that works in general.

3.5 Theoretical result on optimality gap

In this section, we present our main theoretical result, an upper bound on the conversion loss $V_1^{\bar{\pi}} - V_N^{\text{FTVA}(\bar{\pi})}$ for any given single-armed policy $\bar{\pi}$. Setting $\bar{\pi}$ to be an optimal single-armed policy $\bar{\pi}^*$ then leads to an upper bound on the optimality gap of our N-armed policy $\text{FTVA}(\bar{\pi}^*)$. Our result holds under the Synchronization Assumption (SA), which we formally introduce below.

3.5.1 Synchronization Assumption

SA is imposed on a given single-armed policy $\bar{\pi}$. To describe SA, we first define a two-armed system called the *leader-and-follower* system, which consists of a *leader* arm and a *follower* arm. Each arm is associated with the MDP $(\mathbb{S}, \mathbb{A}, P, r)$. At each time step $t \geq 1$, the leader arm is in state $\hat{S}(t)$ and uses the policy $\bar{\pi}$ to choose an action $\hat{A}(t)$ based on $\hat{S}(t)$; the follower arm is in state $S(t)$, and it takes the action $A(t) = \hat{A}(t)$ regardless of $S(t)$. The

state transitions of the two arms are coupled as follows. If $S(t) = \widehat{S}(t)$, then $S(t+1) = \widehat{S}(t+1)$. If $S(t) \neq \widehat{S}(t)$, then $S(t+1)$ and $\widehat{S}(t+1)$ are sampled independently from $P(S(t), A(t), \cdot)$ and $P(\widehat{S}(t), \widehat{A}(t), \cdot)$, respectively. Note that once the states of the two arms become identical, they stay identical indefinitely.

Given the initial states and actions $(S(0), A(0), \widehat{S}(0), \widehat{A}(0)) = (s, a, \widehat{s}, \widehat{a}) \in \mathbb{S} \times \mathbb{A} \times \mathbb{S} \times \mathbb{A}$, we define the *synchronization time* as the first time the two states become identical:

$$\tau^{\text{sync}}(s, a, \widehat{s}, \widehat{a}) \triangleq \min\{t \geq 0: S(t) = \widehat{S}(t)\}. \quad (3.9)$$

Assumption 3.1 (Synchronization Assumption (SA) for a policy $\bar{\pi}$). We say that a single-armed policy $\bar{\pi}$ satisfies the Synchronization Assumption (SA) if for any initial states and actions $(s, a, \widehat{s}, \widehat{a}) \in \mathbb{S} \times \mathbb{A} \times \mathbb{S} \times \mathbb{A}$, the synchronization time $\tau^{\text{sync}}(s, a, \widehat{s}, \widehat{a})$ is a stopping time and satisfies

$$\mathbb{E}[\tau^{\text{sync}}(s, a, \widehat{s}, \widehat{a})] < \infty. \quad (3.10)$$

We view SA as an appealing alternative to GAP for two reasons. First, there are some instances that satisfy SA but not GAP. Two such examples are given in Figures 3.2 and 3.3. [105] provides more examples that violate GAP, all of which can be verified to satisfy SA. Second, SA is easier to verify than GAP. It has been acknowledged in many prior papers that GAP lacks intuitive sufficient conditions and can only be checked numerically. In contrast, SA can be efficiently verified in several ways, which is discussed in Appendix 3.C.

3.5.2 Bounds on conversion loss and optimality gap

We are now ready to state our main theorem.

Theorem 3.2. Consider an N -armed RB problem $(N, \mathbb{S}^N, \mathbb{A}^N, P, r, \alpha N)$ under the single-armed unichain assumption. Let $\bar{\pi}$ be any single-armed policy satisfying SA. For any $N \geq 1$, the conversion loss of FTVA satisfies the upper bound

$$V_1^{\bar{\pi}} - V_N^{\text{FTVA}(\bar{\pi})} \leq \frac{r_{\max} \bar{\tau}_{\max}^{\text{sync}}}{\sqrt{N}}, \quad (3.11)$$

where

$$r_{\max} \triangleq \max_{s \in \mathbb{S}, a \in \mathbb{A}} |r(s, a)|,$$

$$\bar{\tau}_{\max}^{\text{sync}} \triangleq \max_{(s, a, \hat{s}, \hat{a}) \in \mathbb{S} \times \mathbb{A} \times \mathbb{S} \times \mathbb{A}} \mathbb{E} [\tau^{\text{sync}}(s, a, \hat{s}, \hat{a})].$$

Consequently, given any optimal single-armed policy $\bar{\pi}^*$ satisfying SA, for all any $N \geq 1$ the optimality gap of $\text{FTVA}(\bar{\pi}^*)$ is upper bounded as

$$V_N^* - V_N^{\text{FTVA}(\bar{\pi}^*)} \leq V_1^{\bar{\pi}^*} - V_N^{\text{FTVA}(\bar{\pi}^*)} \leq \frac{r_{\max} \bar{\tau}_{\max}^{\text{sync}}}{\sqrt{N}}. \quad (3.12)$$

Proof sketch. The proof of Theorem 3.2 is given in Appendix 3.E. Here we sketch the main ideas of the proof, whose key step involves bounding the conversion loss $V_1^{\bar{\pi}} - V_N^{\text{FTVA}(\bar{\pi})}$ using a fundamental tool from queueing theory, the Little's Law [172]. Specifically, we start with the upper bound

$$\begin{aligned} V_1^{\bar{\pi}} - V_N^{\text{FTVA}(\bar{\pi})} &= \frac{1}{N} \mathbb{E} \left[\sum_{i=1}^N r(\hat{S}_i(\infty), \hat{A}_i(\infty)) - \sum_{i=1}^N r(S_i(\infty), A_i(\infty)) \right] \\ &\leq \frac{2r_{\max}}{N} \mathbb{E} \left[\sum_{i=1}^N \mathbf{1} \left\{ (\hat{S}_i(\infty), \hat{A}_i(\infty)) \neq (S_i(\infty), A_i(\infty)) \right\} \right], \end{aligned} \quad (3.13)$$

which holds since the virtual process $(\hat{S}_i(t), \hat{A}_i(t))$ of each arm i follows the single-armed policy $\bar{\pi}$. We say an arm i is a *bad arm* at time t if $(\hat{S}_i(t), \hat{A}_i(t)) \neq (S_i(t), A_i(t))$, and a *good arm* otherwise. Then $\mathbb{E} \left[\sum_{i=1}^N \mathbf{1} \left\{ (\hat{S}_i(\infty), \hat{A}_i(\infty)) \neq (S_i(\infty), A_i(\infty)) \right\} \right] = \mathbb{E}[\# \text{ bad arms}]$ in steady state.

By Little's Law, we have the following relationship:

$$\begin{aligned} &\mathbb{E}[\# \text{ bad arms}] \\ &= (\text{rate of generating bad arms}) \times \mathbb{E}[\text{time duration of a bad arm}]. \end{aligned}$$

It suffices to bound the two terms on the right hand side. Note that the virtual actions $\hat{A}_i(t)$'s are i.i.d. with mean $\mathbb{E}[\hat{A}_i(t)] = \alpha$; a standard concentration inequality shows that at most $|\sum_{i=1}^N \hat{A}_i(t) - \alpha N| \approx O(\sqrt{N})$ bad arms are generated per time slot. On the other hand, each bad arm stays bad until its real state becomes identical to its virtual state, which occurs in $O(1)$ time by virtue of SA.

3.6 Continuous-time restless bandits

In this section, we consider the continuous-time setting. The setup, policy, and theoretical results for this setting parallel those for the discrete-time setting, except that we no longer require SA. Detailed proofs are provided in Appendix 3.F.

Problem setup. The continuous-time restless bandit (CTRB) problem is similar to its discrete-time counterpart (cf. Section 3.3), except that each single-armed MDP runs in continuous time. In particular, an N -armed CTRB is given by a tuple $(N, \mathbb{S}^N, \mathbb{A}^N, G, r, \alpha N)$, where $\mathbb{S}$ is the finite state space and $\mathbb{A} = \{0, 1\}$ is the action space of a single arm. In continuous time, each arm's dynamics is governed by the transition *rates* (rather than probabilities) $G = \{G(s, a, s')\}_{s, s' \in \mathbb{S}, a \in \mathbb{A}}$, where $G(s, a, s')$ is the rate of transitioning from state s to state $s' \neq s$ upon taking action a . We again assume that the transition kernel G of each arm is unichain. Given the states and actions of all arms, the transitions of different arms are independent from each other. Similarly, $r(s, a)$ is the *instantaneous rate* of accumulating reward while taking action a in state s . The budget constraint now requires that at any moment of time, the total number of arms taking the active action 1 is equal to αN .

The objective is again maximizing the long-run average reward, that is,

$$\underset{\text{policy } \pi}{\text{maximize}} \quad V_N^\pi \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \frac{1}{N} \sum_{i=1}^N \mathbb{E}[r(S_i^\pi(t), A_i^\pi(t))] dt \quad (3.14)$$

$$\text{subject to} \quad \sum_{i=1}^N A_i^\pi(t) = \alpha N \quad \forall t \geq 0. \quad (3.15)$$

Let $V_N^* = \sup_\pi V_N^\pi$ denote the optimal value of the above optimization problem.

Single-armed problem. The single-armed problem for CTRB is defined analogously as its discrete-time counterpart (3.3)–(3.4). This single-armed problem can again be written as a linear program, where the decision variable $y(s, a)$ represents the steady-state probability of being in state s taking action a :

$$\underset{\{y(s, a)\}_{s \in \mathbb{S}, a \in \mathbb{A}}}{\text{maximize}} \quad \sum_{s \in \mathbb{S}, a \in \mathbb{A}} r(s, a) y(s, a) \quad (\text{LP-CT})$$

$$\text{subject to} \quad \sum_{s \in \mathbb{S}} y(s, 1) = \alpha \quad (3.16)$$

$$\sum_{s' \in \mathbb{S}, a \in \mathbb{A}} y(s', a) G(s', a, s) = 0 \quad \forall s \in \mathbb{S} \quad (3.17)$$

$$\sum_{s \in \mathbb{S}, a \in \mathbb{A}} y(s, a) = 1; \quad y(s, a) \geq 0 \quad \forall s \in \mathbb{S}, a \in \mathbb{A}. \quad (3.18)$$

In (3.17), we use the convention that

$$G(s, a, s) = -G(s, a) \triangleq - \sum_{s' \neq s} G(s, a, s').$$

Again, the optimal value of (LP-CT) upper bounds the optimal value for the N -armed problem, i.e., $V_1^{\text{rel}} = V_1^{\bar{\pi}^*} \geq V_N^*$, where $\bar{\pi}^*$ is any optimal single-armed policy and can be computed using the same formula in (3.8). Note that in continuous time, the optimal policy $\bar{\pi}^*$ is carried out through uniformization [230].

Our policy. Our framework for the CTRB, **Follow-the-Virtual-Advice-CT** (FTVA-CT), is presented in Algorithm 2. FTVA-CT works in a similar fashion as its discrete-time counterpart. It takes a single-armed policy $\bar{\pi}$ as input, and each arm i independently simulates a virtual single-armed process $(\hat{S}_i(t), \hat{A}_i(t))$ following $\bar{\pi}$. FTVA-CT then chooses the real actions $A_i(t)$'s to match the virtual actions $\hat{A}_i(t)$'s to the extent allowed by the budget constraint $\sum_{i=1}^N A_i(t) = \alpha N$.

To run FTVA-CT in continuous time for the N -armed problem, we use the following uniformization to set discrete *decision epochs* $\{t_k\}_{k=0,1,\dots}$ for updating the actions and virtual processes. We uniformize at rate $2Ng_{\max}$ with $g_{\max} = \max_{s \in \mathbb{S}, a \in \mathbb{A}} G(s, a)$, which is an upper bound on the total transition rate of the real and virtual states in an N -armed system. We generate a decision epoch either when a real state transitions, or when an independent exponential timer with rate $2Ng_{\max} - \sum_{i=1}^N G(S_i(t), A_i(t))$ ticks.

The definitions of conversion loss, optimality gap, and asymptotic optimality are the same as the discrete-time setting.

Conversion loss and optimality gap. For a given single-armed policy $\bar{\pi}$, we consider a continuous-time version of the leader-and-follower system (cf. Section 3.5.1). For technical reasons, the initial actions are specified

Algorithm 2 Follow-the-Virtual-Advice-CT (FTVA-CT)

Input: $(N, \mathbb{S}^N, \mathbb{A}^N, G, r, \alpha N)$, initial states $\mathbf{S}(0)$, single-armed policy $\bar{\pi}$, max transition rate $g_{\max}$

Initialize: Virtual states $\hat{\mathbf{S}}(0)$ are N i.i.d. samples following the stationary distribution of $\bar{\pi}$; $t_0 = 0$

- 1: **for** $k = 0, 1, 2, \dots$ **do**
 - ▷ *Generate virtual actions*
 - 2: Independently sample $\hat{A}_i(t_k) \leftarrow \bar{\pi}(\cdot | \hat{S}_i(t_k))$ for each arm $i \in [N]$
 - ▷ *Select a set $\mathcal{A}$ of αN arms to activate; arbitrary tie-breaking rule*
 - 3: **if** $\sum_{i=1}^N \hat{A}_i(t_k) \geq \alpha N$ **then**
 - 4: $\mathcal{A} \leftarrow \alpha N$ arms chosen from $\{i: \hat{A}_i(t_k) = 1\}$
 - 5: **else**
 - 6: $\mathcal{B} \leftarrow \alpha N - \sum_{i=1}^N \hat{A}_i(t_k)$ arms from $\{i: \hat{A}_i(t_k) = 0\}$
 - 7: $\mathcal{A} \leftarrow \{i: \hat{A}_i(t_k) = 1\} \cup \mathcal{B}$
 - 8: Apply $A_i(t_k) = 1$ for each $i \in \mathcal{A}$, and apply $A_i(t_k) = 0$ for each $i \notin \mathcal{A}$
 - ▷ *The section below progresses the system to the next decision epoch*
 - 9: $g_k^{\text{real}} \leftarrow \sum_{i=1}^N G(S_i(t_k), A_i(t_k))$
 - 10: $\mathcal{I}_{\text{good}}(t_k) \leftarrow \{i \in [N]: (\hat{S}_i(t_k), \hat{A}_i(t_k)) = (S_i(t_k), A_i(t_k))\}$
 - 11: $\tilde{t}_{k+1} \leftarrow t_k + \text{a sample from the distribution } \text{Exp}(2Ng_{\max} - g_k^{\text{real}})$
 - 12: **if** there is an arm i^* whose real state transitions before $\tilde{t}_{k+1}$ **then**
 - 13: $(t_{k+1}, S_{i^*}(t_{k+1})) \leftarrow (\text{transition time, new real state})$
 - ▷ *Couple virtual and real states*
 - 14: **if** $i^* \in \mathcal{I}_{\text{good}}(t_k)$ **then**
 - 15: $\hat{S}_{i^*}(t_{k+1}) \leftarrow S_{i^*}(t_{k+1})$
 - 16: **else**
 - ▷ *No transitions or transitions of virtual states of uncoupled arms*
 - 17: $t_{k+1} \leftarrow \tilde{t}_{k+1}$
 - 18: $g_k^{\text{virtual}} \leftarrow \sum_{i=1}^N G(\hat{S}_i(t_k), \hat{A}_i(t_k)) \mathbb{1}\{i \notin \mathcal{I}_{\text{good}}(t_k)\}$
 - 19: With probability $1 - \frac{g_k^{\text{virtual}}}{2Ng_{\max} - g_k^{\text{real}}}$, **continue**
 - 20: Sample a pair $(i^*, s') \in [N] \times \mathbb{S}$ with probability

$$(G(\hat{S}_{i^*}(t_k), \hat{A}_{i^*}(t_k), s') \cdot \mathbb{1}\{\hat{S}_{i^*}(t_k) \neq s'\}) / (2Ng_{\max} - g_k^{\text{real}}))$$
 - 21: **if** $i^* \notin \mathcal{I}_{\text{good}}(t_k)$ **then** $\hat{S}_{i^*}(t_{k+1}) \leftarrow s'$
 - 22: Other real and virtual states stay unchanged

differently from the discrete-time setting. Specifically, we assume that the initial action $\hat{A}(0)$ of the leader arm is chosen by $\bar{\pi}$ based on $\hat{S}(0)$, and the follower's initial action $A(0)$ equals $\hat{A}(0)$. As before, the follower arm always takes the same action as the leader arm regardless of its own state. Given initial states $(S(0), \hat{S}(0)) = (s, \hat{s})$, the synchronization time is defined as

$$\tau^{\text{sync}}(s, \hat{s}) \triangleq \inf\{t \geq 0: S(t) = \hat{S}(t)\}. \quad (3.19)$$

We no longer need to impose the Synchronization Assumption, since the unichain assumption automatically implies $\mathbb{E}[\tau^{\text{sync}}(s, \hat{s})] < \infty$ in continuous time—see Lemma 3.18 in Appendix 3.F.

Theorem 3.3. *Consider an N -armed CTRB $(N, \mathbb{S}^N, \mathbb{A}^N, G, r, \alpha N)$ under the single-armed unichain assumption. For any single-armed policy $\bar{\pi}$, the conversion loss of FTVA-CT is upper bounded as*

$$V_1^{\bar{\pi}} - V_N^{\text{FTVA-CT}(\bar{\pi})} \leq \frac{r_{\max}(1 + 2g_{\max}\bar{\tau}_{\max}^{\text{sync}})}{\sqrt{N}}, \quad \forall N \geq 1, \quad (3.20)$$

where $r_{\max} = \max_{s \in \mathbb{S}, a \in \mathbb{A}} |r(s, a)|$, $\bar{\tau}_{\max}^{\text{sync}} = \max_{s \in \mathbb{S}, \hat{s} \in \mathbb{S}} \mathbb{E}[\tau^{\text{sync}}(s, \hat{s})]$.

Consequently, for any optimal single-armed policy $\bar{\pi}^*$, the optimality gap of FTVA-CT($\bar{\pi}^*$) satisfies

$$V_N^* - V_N^{\text{FTVA-CT}(\bar{\pi}^*)} \leq V_1^{\bar{\pi}^*} - V_N^{\text{FTVA-CT}(\bar{\pi}^*)} \leq \frac{r_{\max}(1 + 2g_{\max}\bar{\tau}_{\max}^{\text{sync}})}{\sqrt{N}}, \quad \forall N \geq 1. \quad (3.21)$$

Theorem 3.3 establishes an $O(1/\sqrt{N})$ optimality gap without requiring GAP or any additional assumptions beyond the standard unichain condition.

3.7 Conclusion

In this chapter, we study the average-reward restless bandit problem. We propose a simulation-based framework called **Follow-the-Virtual-Advice** that converts a single-armed optimal policy into a policy in the original N -arm system with an $O(1/\sqrt{N})$ optimality gap. In the discrete-time setting, our results hold under the Synchronization Assumption (SA), a mild and easy-to-verify assumption that covers some problem instances that do not satisfy GAP. In the continuous-time setting, our results do not require any additional assumptions beyond the standard unichain condition. In both settings, our

work is the first to achieve asymptotic optimality without assuming GAP. It will be interesting in future work to explore the possibility of achieving optimality gaps smaller than $O(1/\sqrt{N})$ without relying on GAP or the non-degenerate condition.

Acknowledgement. Y. Hong and W. Wang are supported in part by NSF grants CNS-200773 and ECCS-2145713. Q. Xie is supported in part by NSF grant CNS-1955997 and a J.P. Morgan Faculty Research Award. Y. Chen is supported in part by NSF grants CCF-1704828 and CCF-2233152. This project started when Q. Xie, Y. Chen, and W. Wang were attending the Data-Driven Decision Processes program of the Simons Institute for the Theory of Computing.

3.A More related work

In this section, we provide additional discussion on related work.

Related work on finite-horizon or discounted-reward settings. Apart from the infinite-horizon average-reward setting, there is also a large body of work on other reward settings.

In the finite-horizon total-reward setting, there have been papers [56–58, 87, 108, 109, 111, 155, 296, 300] that readily achieve asymptotic optimality in the $N \rightarrow \infty$ limit without assumptions, with the main focus being on improving the orders of the optimality gaps. Specifically, the optimality gap has been improved from $o(1)$ in [155] to $O(\log N/\sqrt{N})$ in [57, 87, 296] and $O(1/\sqrt{N})$ in [56, 58, 111], without assumptions. Later, Zhang and Frazier [300] propose a policy that achieves an $O(1/N)$ optimality gap under a mild assumption called non-degeneracy; Gast et al. [108, 109] further improve the optimality gap to $O(\exp(-CN))$ under the same non-degeneracy assumption. Although the above-mentioned papers on the finite-horizon total-reward setting are able to prove strong bounds of optimality gap in terms of the scaling of N under minimal conditions, most of them either do not consider the scaling of the time horizon T [56, 155], or have a super-linear dependency on the T (quadratic in [87, 111, 296], and $O(T \log T)$ in [57]). Consequently, most of the bounds in the finite horizon setting are incomparable to the bounds in the infinite-horizon average-reward setting — the latter is analogous to

having a linear dependency on T . There are, however, two exceptions: Brown and Zhang [58] and Gast et al. [109] obtain bounds with linear dependencies on T , under additional assumptions similar to Synchronization Assumption (Assumption 3.1). Nevertheless, there is no direct way for applying the policies in [58, 109] to the infinite-horizon setting, since they both involve solving subproblems whose complexities depend on the time horizon.

The asymptotic optimality in the infinite-horizon discounted-reward setting has also been considered in the prior work [56, 58, 111, 301]. Similar to the finite-horizon setting, asymptotic optimality can be achieved without assumptions. In particular, an $O(N^{\log_2(\sqrt{\gamma})})$ optimality gap is obtained in [56] for a discount factor $\gamma \in (1/2, 1)$, and $O(1/\sqrt{N})$ optimality gaps are achieved in [58, 111, 301]. These results are incomparable to the asymptotic optimality results in the infinite-horizon average-reward setting due to their dependencies on γ . The $O(1/\sqrt{N})$ bounds in [58, 111, 301] scale at least quadratically with the effective horizon $1/(1 - \gamma)$ rather than linearly. Meanwhile, the $O(N^{\log_2(\sqrt{\gamma})})$ bound in [56] has a coefficient that scales linearly with $1/(1 - \gamma)$, but since $N^{\log_2(\sqrt{\gamma})}$ becomes constant in N after taking the limit of $\gamma \rightarrow 1$, this bound is not comparable to the asymptotic optimality results in the average-reward setting, where the bounds must be diminishing in N .

Relationship between the RB problem and other bandit problems.

While the exact optimality of the RB problem is in general intractable, there is a special case that has been solved optimally, known as the rested bandits (also referred to as nonrestless bandits or Markovian bandits), where an arm's state does not change if it is not pulled, and only one arm is pulled at a time. The restful bandit problem has been widely studied, where the celebrated Gittins index policy is proven to be optimal [31, 112, 113, 259, 264, 273, 282]. A more recent reference on this topic is [114].

The RB problem falls within the broader class of bandit problems, for which different formulations exist including stochastic bandits, adversarial bandits, and Bayesian bandits. The common theme in these formulations is to find a reward-maximizing strategy of pulling arms in the presence of uncertainty in the arms' rewards; see the book [177] for a comprehensive overview. Among these formulations, closely related to RBs is the Bayesian bandit problem, where Bayesian posteriors are used to model knowledge of unknown reward distributions. The Bayesian posterior can be seen as a state with known transition probabilities, hence the Bayesian bandit problem

can be analyzed by applying tools from RBs. Examples demonstrating this connection can be found in [56].

3.B Discussion on the computational costs of FTVA

FTVA defined in Algorithm 1 can be computed and implemented efficiently.

First, computing FTVA requires solving (LP) once. Given that (LP) does not depend on N , the computational cost here is a constant.

Next, we examine the computational cost of implementing FTVA.

- Each arm simulates a virtual process. Because the computation required for each simulation does not scale with N , the overall computational cost here is linear in N .
- In each time step, FTVA selects arms from a subset of arms to determine the real actions. The computational cost for this selection process scales linearly with N .

Combining these components, we can see that the computational cost of implementing FTVA is linear in N .

Additionally, we note that the simulation of the virtual processes is independent across the arms, so they can be implemented in a fully distributed manner. However, when taking real actions, the policy needs to know the virtual actions from all arms, so this step cannot be implemented distributively.

3.C Sufficient conditions for Synchronization Assumption (SA)

Recall that for the discrete-time setting, our bound on the conversion loss in Theorem 3.2 holds when the input single-armed policy $\bar{\pi}$ satisfies the Synchronization Assumption (SA). SA stipulates that the two arms in the leader-and-follower system (cf. Section 3.5.1) under the policy $\bar{\pi}$ will reach the same state in finite expected time. All other existing work requires the GAP assumption, which pertains to the global behavior of a non-linear dynamic system and has no known sufficient conditions that are easily checkable. In

comparison, SA is a reachability assumption imposed on a *finite state Markov chain* and thus substantially simpler.

To provide a more intuitive understanding of SA, in this section we present several sufficient conditions for SA to hold. These conditions involve the recurrent classes in the Markov chains induced by certain single-armed policies as well as the existence of self-loops or cycles in the corresponding transition diagrams. Such conditions can be verified, often in a straightforward manner, by inspecting the transition probabilities. As self-loops and cycles are common in many classes of MDPs, these conditions showcase that SA is a relatively mild assumption. We emphasize that the sufficient conditions presented in this section are not exhaustive. They serve as illustrative examples of when SA holds and offer insights on the nature of synchronization.

The discussion in this section pertains to the discrete-time setting. We reiterate that our results for the continuous-time setting do not require SA.

3.C.1 Preliminaries

Before presenting the sufficient conditions, we first prove a few preliminary facts. The first one is a basic property of unichain MDPs.

Proposition 3.4. *Consider two arbitrary Markovian policies $\bar{\pi}$ and $\bar{\pi}'$ for the unichain MDP $(\mathbb{S}, \mathbb{A}, P, r)$. Let the recurrent classes of the two policies $\bar{\pi}$ and $\bar{\pi}'$ be $\mathcal{S}$ and $\mathcal{S}'$, respectively. Then $\mathcal{S} \cap \mathcal{S}' \neq \emptyset$.*

Proof. We prove this proposition by contradiction. Suppose $\mathcal{S} \cap \mathcal{S}' = \emptyset$. We define a new policy $\bar{\pi}''$ as

$$\bar{\pi}''(a|s) = \begin{cases} \bar{\pi}(a|s), & \text{if } s \in \mathcal{S}, \\ \bar{\pi}'(a|s), & \text{otherwise.} \end{cases} \quad \text{for } s \in \mathbb{S}, a \in \mathbb{A}.$$

Then $\mathcal{S}$ and $\mathcal{S}'$ are two distinct recurrent classes under $\bar{\pi}''$, because by definition of $\bar{\pi}''$, an arm with the initial state in $\mathcal{S}$ remains in $\mathcal{S}$, so it cannot reach $\mathcal{S}'$; similarly an arm cannot reach $\mathcal{S}$ from $\mathcal{S}'$. The existence of two recurrent classes under $\bar{\pi}''$ contradicts the unichain condition. Therefore, we must have $\mathcal{S} \cap \mathcal{S}' \neq \emptyset$. $\square$

The next proposition provides a convenient way for verifying SA and is used for establishing other sufficient conditions.

Proposition 3.5. *Consider the leader-and-follower system under the policy $\bar{\pi}$. If there exists some $t < \infty$ such that for all initial states and actions $(s, a, \hat{s}, \hat{a}) \in \mathbb{S} \times \mathbb{A} \times \mathbb{S} \times \mathbb{A}$,*

$$\mathbb{P}(\tau^{\text{sync}}(s, a, \hat{s}, \hat{a}) < t) > 0,$$

then Synchronization Assumption holds for the policy $\bar{\pi}$.

Proof. We claim that for any $(s, a, \hat{s}, \hat{a}) \in \mathbb{S} \times \mathbb{A} \times \mathbb{S} \times \mathbb{A}$,

$$\mathbb{P}(\tau^{\text{sync}}(s, a, \hat{s}, \hat{a}) \geq kt) \leq \left(\max_{s', a', \hat{s}', \hat{a}'} \mathbb{P}(\tau^{\text{sync}}(s', a', \hat{s}', \hat{a}') \geq t) \right)^k, \quad (3.22)$$

where the $\max_{s', a', \hat{s}', \hat{a}'}$ is taken over $\mathbb{S} \times \mathbb{A} \times \mathbb{S} \times \mathbb{A}$. We prove this claim by induction on k . The base case of $k = 1$ is given. Suppose we have proved (3.22) for a certain k , then

$$\begin{aligned} & \mathbb{P}(\tau^{\text{sync}}(s, a, \hat{s}, \hat{a}) \geq (k+1)t) \\ &= \mathbb{P}(\tau^{\text{sync}}(s, a, \hat{s}, \hat{a}) \geq kt) \\ & \quad \cdot \mathbb{P}(\tau^{\text{sync}}(s, a, \hat{s}, \hat{a}) \geq (k+1)t \mid \tau^{\text{sync}}(s, a, \hat{s}, \hat{a}) \geq kt) \\ &= \mathbb{P}(\tau^{\text{sync}}(s, a, \hat{s}, \hat{a}) \geq kt) \\ & \quad \cdot \mathbb{P}\left(\tau^{\text{sync}}(S(kt), A(kt), \hat{S}(kt), \hat{A}(kt)) \geq t \mid S(kt) \neq \hat{S}(kt)\right) \\ &\leq \mathbb{P}(\tau^{\text{sync}}(s, a, \hat{s}, \hat{a}) \geq kt) \cdot \max_{s', a', \hat{s}', \hat{a}'} \mathbb{P}(\tau^{\text{sync}}(s', a', \hat{s}', \hat{a}') \geq t) \\ &\leq \left(\max_{s', a', \hat{s}', \hat{a}'} \mathbb{P}(\tau^{\text{sync}}(s', a', \hat{s}', \hat{a}') \geq t) \right)^{k+1}. \end{aligned}$$

where in the second equality we have used the Markov property of the system, and in the last inequality we apply the induction hypothesis. We have proved (3.22).

The bound on the expectation $\mathbb{E}[\tau^{\text{sync}}(s, a, \hat{s}, \hat{a})]$ follows by summing the tail bound (3.22) over k :

$$\begin{aligned} \mathbb{E}[\tau^{\text{sync}}(s, a, \hat{s}, \hat{a})] &\leq \sum_{k=0}^{\infty} t \mathbb{P}(\tau^{\text{sync}}(s, a, \hat{s}, \hat{a}) \geq kt) \\ &\leq \frac{t}{1 - \max_{s', a', \hat{s}', \hat{a}'} \mathbb{P}(\tau^{\text{sync}}(s', a', \hat{s}', \hat{a}') \geq t)} \\ &= \frac{t}{\min_{s', a', \hat{s}', \hat{a}'} \mathbb{P}(\tau^{\text{sync}}(s', a', \hat{s}', \hat{a}') < t)} < \infty. \end{aligned}$$

We have established SA and thereby finished the proof. $\square$

Remark 3.6. Proposition 3.5 implies that SA only requires that the Markov chain of the leader-and-follower system, $(S(t), \widehat{S}(t))$, can reach the subset of states $\{(s, \widehat{s}) \in \mathbb{S}^2 : s = \widehat{s}\}$ from any initial state. As a result, SA can be efficiently verified using path-finding graph algorithms on the state transition diagram of the Markov chain.

3.C.2 Sufficient conditions based on self-loops

The unichain condition and Proposition 3.4 guarantee that there are some states that the leader arm and the follower arm will both visit for a positive fraction of times.² If we can further show that the two arms have a positive probability to visit one of those states *at the same time*, then Proposition 3.5 can be applied to verify SA. A natural sufficient condition is the existence of self-loops, which guarantees that with a positive probability, the arm reaching those states first will wait for the other arm to come to the same state.

Proposition 3.7. *Consider the single-armed problem with the unichain MDP $(\mathbb{S}, \mathbb{A}, P, r)$ and budget $\alpha \in (0, 1)$. Let $\bar{\pi}$ be a single-armed policy with recurrent class $\mathcal{S}$. If $P(s, 0, s) > 0$ and $P(s, 1, s) > 0$ for all $s \in \mathcal{S}$, then Synchronization Assumption holds for the policy $\bar{\pi}$.*

Proposition 3.7 assumes that all states in the recurrent class of the policy $\bar{\pi}$ have self-loops. We can actually require fewer self-loops if we can characterize the recurrent states of the follower arm. In particular, let $\bar{\pi}_1$ denote the *all-one policy*, the policy that applies action 1 in all states. Observe that if the leader arm applies action 1 for a sufficiently long time, the follower arm will also apply action 1 and effectively follows $\bar{\pi}_1$. The recurrent class of the policy $\bar{\pi}_1$ must intersect the recurrent class of $\bar{\pi}$, so it suffices to have self-loops in the intersection of these two recurrent classes. The idea is formalized in Proposition 3.8.

Proposition 3.8. *Consider the single-armed problem with the unichain MDP $(\mathbb{S}, \mathbb{A}, P, r)$ and budget $\alpha \in (0, 1)$. Let $\bar{\pi}$ be a single-armed policy with recurrent class $\mathcal{S}$. Let the recurrent class of the all-one policy $\bar{\pi}_1$ be denoted as $\mathcal{S}_1$. If the following conditions hold:*

1. *There exists $s^a \in \mathcal{S}$ such that $\bar{\pi}(1|s^a) > 0$ and $P(s^a, 1, s^a) > 0$;*

²Although the follower arm does not follow a Markovian policy since it copies actions from the leader arm, by standard results on average reward MDP, the set of states in which the follower arm spends a positive fraction of time is the same as the recurrent class of some Markovian policy (cf. Chapter 8.9 of [230]). This class of states must intersect with $\mathcal{S}$ by Proposition 3.4 and the unichain condition.

2. There exists $s^b \in \mathcal{S} \cap \mathcal{S}_1$ such that $P(s^b, 0, s^b) > 0$ and $P(s^b, 1, s^b) > 0$, then Synchronization Assumption holds for the policy $\bar{\pi}$.

Proposition 3.9 below gives another sufficient condition, which only requires one self-loop. Note that this condition does not subsume the one in Proposition 3.8.

Proposition 3.9. *Consider the single-armed problem with the unichain MDP $(\mathbb{S}, \mathbb{A}, P, r)$ and budget $\alpha \in (0, 1)$. Let $\bar{\pi}$ be a single-armed policy with recurrent class $\mathcal{S}$. Let the recurrent class of the all-one policy $\bar{\pi}_1$ be denoted as $\mathcal{S}_1$. If there exists $s^* \in \mathcal{S} \cap \mathcal{S}_1$ such that $\bar{\pi}(1|s^*) > 0$ and $P(s^*, 1, s^*) > 0$, then Synchronization Assumption holds for the policy $\bar{\pi}$.*

We remark that Propositions 3.8 and 3.9 have analogous versions that are stated in terms of the all-zero policy $\bar{\pi}_0$, the policy that applies action 0 in all states. We omit the details.

Now we prove the above propositions. Note that Proposition 3.7 is strictly weaker than Proposition 3.8: because the single-armed policy $\bar{\pi}$ satisfies the budget constraint with $\alpha \in (0, 1)$, there must be a state $s^a \in \mathcal{S}$ such that $\bar{\pi}(1|s^a) > 0$. Therefore, we only need to prove Propositions 3.8 and 3.9.

Proof of Proposition 3.8. Given any initial states and actions $(s, a, \hat{s}, \hat{a}) \in \mathbb{S} \times \mathbb{A} \times \mathbb{S} \times \mathbb{A}$, we construct a sequence of positive probability events that leads to $S(t) = \hat{S}(t)$:

1. The leader arm reaches the state $s^a \in \mathcal{S}$ after t_1 time slots;
2. The leader arm stays in state s^a and keeps applying action 1 for t_2 times slots; meanwhile, the follower arm also applies action 1 and reaches the state $s^b \in \mathcal{S} \cap \mathcal{S}_1$;
3. The leader arm reaches the state s^b in another t_3 time slots; meanwhile, the follower arm stays in the state s^b , so the two arms synchronize.

The transitions of the two arms during the above sequence of events are illustrated in Figure 3.5. We argue that there exists suitably large t_1, t_2, t_3 such that the above three events happen with a positive probability. The first event can happen for a suitably large t_1 because the leader arm follows the policy $\bar{\pi}$, and s^a is in the recurrent class of $\bar{\pi}$. In the second event, the leader arm can stay in state s^a and keep applying action 1 because $\bar{\pi}(1|s^a) > 0$ and $P(s^a, 1, s^a) > 0$. The follower arm can reach s^b after a suitably large t_2 time slots because it keeps applying action 1 and s^b is in the recurrent class of the all-one policy $\bar{\pi}_1$. In the third event, the leader arm can reach s^b after a suitably large t_3 time slots because s^b is also in the recurrent class of

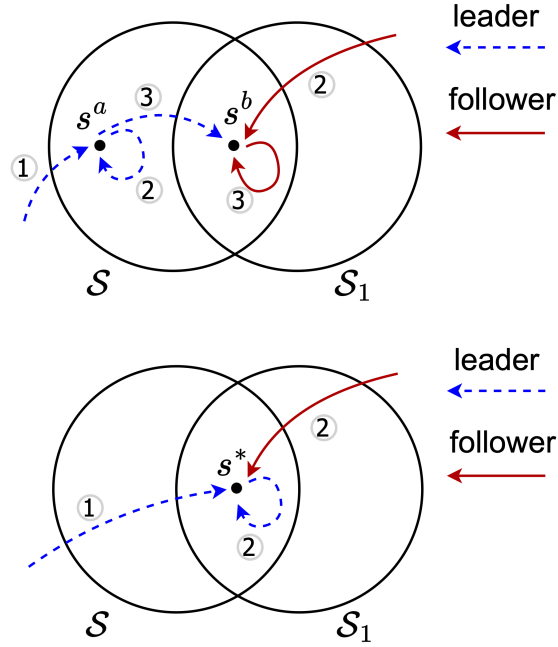


Figure 3.5: The illustration of the positive probability sample paths that lead to synchronization in Proposition 3.8 (top) and Proposition 3.9 (bottom). In each figure, the dotted arrows correspond to the sample path of the leader arm's state, while the solid arrows correspond to the sample path of the follower arm's state. The numbers near the arrows denote the temporal order of the transition events.

$\bar{\pi}$. Meanwhile, the follower arm can stay in s^b because $P(s^b, 0, s^b) > 0$ and $P(s^b, 1, s^b) > 0$. Therefore, we have proved that for any $(s, a, \hat{s}, \hat{a})$,

$$\mathbb{P}(\tau^{\text{sync}}(s, a, \hat{s}, \hat{a}) \leq t_1 + t_2 + t_3) > 0.$$

By Proposition 3.5, we establish SA. $\square$

Proof of Proposition 3.9. Given any initial states and actions $(s, a, \hat{s}, \hat{a}) \in \mathbb{S} \times \mathbb{A} \times \mathbb{S} \times \mathbb{A}$, we construct a sequence of positive probability events that leads to $S(t) = \hat{S}(t)$:

1. The leader arm reaches the state s^* after t_1 time slots;
2. The leader arm stays in the state s^* and keeps applying action 1 for t_2 time slots; meanwhile, the follower arm also applies action 1 and reaches the state s^* , so the two arms synchronize.

The transitions of the two arms during the above sequence of events are illustrated in Figure 3.5. We argue that there exists suitable t_1 and t_2 such that the above two events happen with a positive probability. The first event can happen for a suitably large t_1 because s^* is in the recurrent class of $\bar{\pi}$. In the second event, the leader arm can stay in the state s^* and keeps applying action 1 because $\bar{\pi}(1|s^*) > 0$ and $P(s^*, 1, s^*) > 0$. Meanwhile, the follower arm can reach s^* after a suitably large t_2 time slots because s^* is also in the recurrent class of the all-one policy $\bar{\pi}_1$. Therefore, we have proved that for any $(s, a, \hat{s}, \hat{a})$,

$$\mathbb{P}(\tau^{\text{sync}}(s, a, \hat{s}, \hat{a}) \leq t_1 + t_2) > 0.$$

By Proposition 3.5, we establish SA. $\square$

3.C.3 Sufficient conditions based on cycles

The above sufficient conditions are based on self-loops, which can be viewed as cycles of minimal length. These conditions can be generalized to longer cycles, formally defined below.

Definition 3.10 (Cycle). Consider the MDP $(\mathbb{S}, \mathbb{A}, P, r)$. We call an ordered set of states $C = (s_0, s_1, \dots, s_L)$ a *cycle under the Markovian policy $\bar{\pi}$* if there is a positive probability of transitioning from s_j to s_{j+1} for all j under the policy $\bar{\pi}$ (we identify s_{L+1} with s_0). We call the ordered set C a *cycle under any policy* if there is a positive probability of transitioning from s_j to s_{j+1} for all j under any policy. In both cases, we call L the *length of the cycle C* .

We give two sufficient conditions based on cycles. These conditions are relaxations of the conditions in Proposition 3.8 and 3.9.

Proposition 3.11. *Consider the single-armed problem with the unichain MDP $(\mathbb{S}, \mathbb{A}, P, r)$ and budget $\alpha \in (0, 1)$. Let $\bar{\pi}$ be a single-armed policy with recurrent class $\mathcal{S}$. Let the recurrent class of the all-one policy $\bar{\pi}_1$ be denoted as $\mathcal{S}_1$. If the following conditions hold:*

1. *There exists a cycle C^a under the policy $\bar{\pi}$ in $\mathcal{S}$, and $\bar{\pi}(1|s) > 0$ for all $s \in C^a$;*
2. *There exists cycle C^b under any policy in $\mathcal{S} \cap \mathcal{S}_1$;*
3. *The lengths of the two cycles C^a and C^b are relatively prime.*

then Synchronization Assumption holds for the policy $\bar{\pi}$.

Proposition 3.12. *Consider the single-armed problem with the unichain MDP $(\mathbb{S}, \mathbb{A}, P, r)$ and budget $\alpha \in (0, 1)$. Let $\bar{\pi}$ be a single-armed policy with*

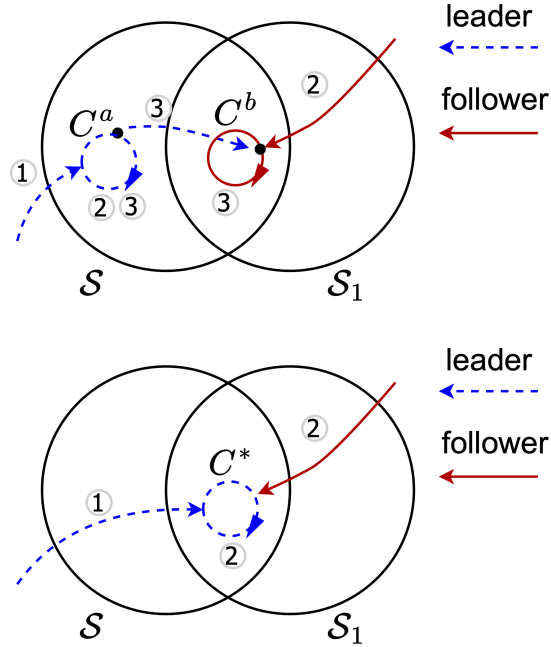


Figure 3.6: The illustration of the positive probability sample paths that lead to synchronization in Proposition 3.11 (top) and Proposition 3.12 (bottom). In each figure, the dotted arrows correspond to the sample path of the leader arm's state, while the solid arrows correspond to the sample path of the follower arm's state. The numbers near the arrows denote the temporal order of the transition events.

recurrent class $\mathcal{S}$. Let the recurrent class of the all-one policy $\bar{\pi}_1$ be denoted as $\mathcal{S}_1$. If the following conditions hold:

1. There exists a cycle C^* under the policy $\bar{\pi}$ in $\mathcal{S} \cap \mathcal{S}_1$, and $\bar{\pi}(1|s) > 0$ for all $s \in C^*$;
2. The policy $\bar{\pi}_1$ is aperiodic,

then Synchronization Assumption holds for the policy $\bar{\pi}$.

Similarly to Propositions 3.8 and 3.9, Propositions 3.11 and 3.12 have analogous version that are stated in terms of the all-zero policy $\bar{\pi}_0$. We omit the details.

Now we prove Propositions 3.11 and 3.12.

Proof of Proposition 3.11. Given any initial states and actions $(s, a, \hat{s}, \hat{a}) \in \mathbb{S} \times \mathbb{A} \times \mathbb{S} \times \mathbb{A}$, we construct a sequence of positive probability events that leads to $S(t) = \hat{S}(t)$:

1. The leader arm reaches the cycle C^a after t_1 time slots;
2. The leader arm transitions along the cycle and keeps applying action 1; meanwhile, the follower arm also applies action 1 and reaches the cycle C^b ; this happens in t_2 times slots; let the states of the leader and follower arm at this moment be s^a and s^b ;
3. The leader arm transitions along the cycle C^a for another c laps, for some positive integer c , and then spends another t_3 time slots to go from s^a to s^b ; meanwhile, the follower arm transitions along the cycle C^b for another d laps, for some positive integer d .

The transition of the arms during the sequence of events are as illustrated in Figure 3.6. We argue that there exists some suitable t_1, t_2, t_3, c , and d such that the above four events happen with a positive probability, and the two arms synchronize after the four events.

The first event happens with a positive probability for suitably large t_1 because C^a is in $\mathcal{S}$. The second event can happen for suitably large t_2 because $\pi(1|s) > 0$ for all $s \in C^a$, and C^b is in $\mathcal{S}_1$.

After the first two events, suppose the state of the leader arm is s^a , and the state of the follower arm is s^b , then $s^a \in C^a$ and $s^b \in C^b$. Because both cycles are in $\mathcal{S}$, there exists a positive probability path from s^a to s^b under $\bar{\pi}$. Let t_3 be the length of the path. Let L^a and L^b be the length of cycles C^a and C^b . Because L^a and L^b are relatively prime, we can take c and d such that $dL^b - cL^a = t_3$. With this choice of c, d , and t_3 , the two arms will synchronize after the third and fourth events.

Now we argue that the third and fourth events happen with a positive probability. Because C^a is a cycle under the policy $\bar{\pi}$, the leader arm can transition along the cycle for an arbitrary number of time slots. Similarly, because C^b is a cycle under any policy, the follower arm can transition along the cycle for an arbitrary number of time slots; the leader arm can reach s^b from s^a because of the positive probability path described in the last paragraph.

Therefore, we have proved that for any $(s, a, \hat{s}, \hat{a})$,

$$\mathbb{P}(\tau^{\text{sync}}(s, a, \hat{s}, \hat{a}) \leq t_1 + t_2 + cL^a + t_3) > 0.$$

By Proposition 3.5, we establish SA. □

Proof of Proposition 3.12. Given any initial states and actions $(s, a, \hat{s}, \hat{a}) \in \mathbb{S} \times \mathbb{A} \times \mathbb{S} \times \mathbb{A}$, we construct a sequence of positive probability events that leads to $S(t) = \hat{S}(t)$:

1. The leader arm reaches the cycle C^* after t_1 time slots;
2. The leader arm transitions along the cycle and keeps applying action 1; meanwhile, the follower arm also applies action 1 and reaches the same state as the leader arm after t_2 time slots.

The transition of the arms during the sequence of events are as illustrated in Figure 3.6. We argue that the above two events happen with a positive probability for suitable t_1 and t_2 . The first event can happen for suitably large t_1 because the leader arm follows the policy $\bar{\pi}$ the cycle C^* is in the recurrent class of $\bar{\pi}$. After reaching C^* , the leader arm can remain on C^* because C^* is a cycle under the policy $\bar{\pi}$, and $\bar{\pi}(1|s) > 0$ for all $s \in C^*$. The follower arm can reach the same state as the leader arm because the cycle C^* is also in the recurrent class of $\bar{\pi}_1$, and $\bar{\pi}_1$ is aperiodic. Therefore, we have proved that for any $(s, a, \hat{s}, \hat{a})$,

$$\mathbb{P}(\tau^{\text{sync}}(s, a, \hat{s}, \hat{a}) \leq t_1 + t_2) > 0.$$

By Proposition 3.5, we establish SA. □

3.D On relaxing the unichain condition

Throughout this chapter, we have imposed the (all-policy) unichain condition, i.e., all Markovian single-armed policies have a single recurrent class. Making this blanket assumption simplifies our presentation, but it can be substantially relaxed for many of our results, as we discuss below.

The main change caused by dropping the unichain assumption is the definition of the single-armed problem, because there could be multiple recurrent classes under a single-armed policy, each corresponding to a different long-run average reward. To adapt to this change, we can let both the single-armed policy $\bar{\pi}$ and the initial state distribution μ be optimization variables of the single-armed problem. We denote the average reward given $\bar{\pi}$ and μ as $V_1^{\bar{\pi}, \mu}$. We still denote the optimal value of the single-armed problem as V_1^{rel} . Note that the single-armed problem can still be solved by the same linear program in (LP) or (LP-CT): The optimal initial state distribution μ^* can

be taken as the optimal stationary state distribution $(y^*(s, 1) + y^*(s, 0))_{s \in \mathbb{S}}$, given the optimal solution of LP, y^* .

The definitions of FTVA and FTVA-CT do not require any change, considering that they are already using the stationary distribution of $\bar{\pi}$ to initialize the virtual states. The conversion losses, now denoted as $V_1^{\bar{\pi}, \mu} - V_N^{\text{FTVA}(\bar{\pi}, \mu)}$ and $V_1^{\bar{\pi}, \mu} - V_N^{\text{FTVA-CT}(\bar{\pi}, \mu)}$, have the same bound as what we currently have in equation (3.11) in Theorem 3.2 for discrete-time RBs and equation (3.20) in Theorem 3.3 for continuous-time RBs. These bounds hold as long as the given input policy $\bar{\pi}$ has *finite expected synchronization times*, which can be seen by inspecting the proofs of the theorems given in Appendix 3.E and Appendix 3.F.

The optimality gap bounds in (3.12) in Theorem 3.2 and (3.21) in Theorem 3.3 still follow from the conversion loss bounds when the unichain assumption is dropped. To see this, for each of the discrete-time and continuous-time settings, consider an optimal single-armed policy $\bar{\pi}^*$ and optimal initial state distribution μ^* . Note that $V_1^{\bar{\pi}^*, \mu^*} = V_1^{\text{rel}} \geq V_N^*$. If $V_1^{\bar{\pi}^*, \mu^*} - V_N^{\text{FTVA}(\bar{\pi}^*, \mu^*)} = O(1/\sqrt{N})$, then $V_N^* - V_N^{\text{FTVA}(\bar{\pi}^*, \mu^*)} = O(1/\sqrt{N})$; if $V_1^{\bar{\pi}^*, \mu^*} - V_N^{\text{FTVA-CT}(\bar{\pi}^*, \mu^*)} = O(1/\sqrt{N})$, then $V_N^* - V_N^{\text{FTVA-CT}(\bar{\pi}^*, \mu^*)} = O(1/\sqrt{N})$.

In the discrete-time setting, our sufficient conditions for finite expected synchronization times, presented in Propositions 3.8, 3.9, 3.11 and 3.12, are valid as long as each of the two policies $\bar{\pi}$ and $\bar{\pi}_1$ has a single recurrent class and the two classes intersect. Similarly, Lemma 3.18 in Appendix 3.F, which establishes finite expected synchronization times for the continuous-time setting, holds under the same condition on $\bar{\pi}$ and $\bar{\pi}_1$.

3.E Proofs for discrete-time RBs

3.E.1 Overview

In this section, we focus on proving the bound (3.11) on the conversion loss $V_1^{\bar{\pi}} - V_N^{\text{FTVA}(\bar{\pi})}$ for any single-armed policy $\bar{\pi}$. The asymptotic optimality result (3.12) in Theorem 3.2 will be a direct consequence of (3.11) if we take $\bar{\pi}$ to be any optimal single-armed policy $\bar{\pi}^*$.

We first show that under FTVA ($\bar{\pi}$) defined in Algorithm 1, each virtual process is an independent Markov chain induced by applying $\bar{\pi}$ to the single-armed system $(\mathbb{S}, \mathbb{A}, P, r)$, as stated in Lemma 3.13 below. We will rigorously prove it in the next subsection.

Lemma 3.13. *Under FTVA ($\bar{\pi}$) given in Algorithm 1, for each $t = 0, 1, 2, \dots$, $i \in [N]$, $\hat{\mathbf{s}} \in \mathbb{S}^N$, and $\hat{\mathbf{a}} \in \mathbb{A}^N$,*

$$\mathbb{P}\left(\hat{\mathbf{A}}(t) = \hat{\mathbf{a}} \mid \hat{\mathbf{S}}(t), \dots, \hat{\mathbf{S}}(0)\right) = \prod_{i \in [N]} \bar{\pi}(\hat{a}_i \mid \hat{S}_i(t)) \quad (3.23)$$

$$\mathbb{P}\left(\hat{\mathbf{S}}(t+1) = \hat{\mathbf{s}} \mid \hat{\mathbf{S}}(t), \hat{\mathbf{A}}(t), \dots, \hat{\mathbf{S}}(0), \hat{\mathbf{A}}(0)\right) = \prod_{i \in [N]} P(\hat{S}_i(t), \hat{A}_i(t), \hat{s}_i). \quad (3.24)$$

Let $(y(s, a))_{s \in \mathbb{S}, a \in \mathbb{A}}$ be the steady-state state-action distribution of the single-armed system under $\bar{\pi}$. By (3.23)(3.24), for each $t \geq 0$, $(\hat{S}_i(t), \hat{A}_i(t))$ for $i \in [N]$ are i.i.d. with the distribution $(y(s, a))_{s \in \mathbb{S}, a \in \mathbb{A}}$.

Next, as sketched in Section 3.5.2, we bound $V_1^{\bar{\pi}} - V_N^{\text{FTVA}(\bar{\pi})}$ as

$$\begin{aligned} V_1^{\bar{\pi}} - V_N^{\text{FTVA}(\bar{\pi})} &= \frac{1}{N} \mathbb{E} \left[\sum_{i=1}^N r(\hat{S}_i(\infty), \hat{A}_i(\infty)) - \sum_{i=1}^N r(S_i(\infty), A_i(\infty)) \right] \\ &\leq \frac{2r_{\max}}{N} \mathbb{E} \left[\sum_{i=1}^N \mathbb{1} \left\{ (\hat{S}_i(\infty), \hat{A}_i(\infty)) \neq (S_i(\infty), A_i(\infty)) \right\} \right]. \end{aligned} \quad (3.13)$$

We say an arm i is a *bad arm* at time t if $(\hat{S}_i(t), \hat{A}_i(t)) \neq (S_i(t), A_i(t))$, and a *good arm* otherwise. Then

$$\mathbb{E} \left[\sum_{i=1}^N \mathbb{1} \left\{ (\hat{S}_i(\infty), \hat{A}_i(\infty)) \neq (S_i(\infty), A_i(\infty)) \right\} \right] = \mathbb{E}[\# \text{ bad arms}]$$

in steady state.

By Little's Law, we have the following relationship:

$$\begin{aligned} &\mathbb{E}[\# \text{ bad arms}] \\ &= (\text{rate of generating bad arms}) \times \mathbb{E}[\text{time duration of a bad arm}]. \end{aligned}$$

To make the quantities in the above expression precise, we make the following definitions.

We say there is a *disagreement event of the arm i* happening at time t if $\hat{A}_i(t) \neq A_i(t)$. The disagreement events are the only cause that turns good

arms into bad arms because otherwise by the construction of our algorithm, $\widehat{S}_i(t)$ will remain the same as $S_i(t)$. The number of disagreement events in each time slot is determined by how much the budget required by virtual actions violates the constraint. We call the expected number of disagreement events when the virtual states are $\widehat{\mathbf{s}}$ the *instantaneous disagreement rate*, denoted as $d(\widehat{\mathbf{s}})$.

We also define *disagreement period of the arm i* as a continuous period of time when arm i is a bad arm, separated by disagreement events. Formally, **Definition 3.14** (Disagreement period). Given a sample path of the arm i 's real states and virtual states $(S_i(t), \widehat{S}_i(t))_{t \geq 0}$, we define the disagreement period of the arm i as a time interval $[t_{\text{begin}}, t_{\text{end}} - 1]$ such that

$$\begin{aligned} A_i(t_{\text{begin}}) &\neq \widehat{A}_i(t_{\text{begin}}); \\ A_i(t) &= \widehat{A}_i(t) \text{ and } S_i(t) \neq \widehat{S}_i(t) \quad \forall t \in [t_{\text{begin}} + 1, t_{\text{end}} - 1]; \\ A_i(t_{\text{end}}) &\neq \widehat{A}_i(t_{\text{end}}) \text{ or } S_i(t_{\text{end}}) = \widehat{S}_i(t_{\text{end}}). \end{aligned}$$

We use D_{avg} to denote the long-run average length of the disagreement periods, i.e., the average of $t_{\text{end}} - t_{\text{begin}}$ of all disagreement periods.

With the definitions, we can state Little's Law in the context of disagreement periods. We omit its proof since it is a direct consequence of Little's Law.

Lemma 3.15 (Little's Law for disagreement periods).

$$\mathbb{E} \left[\sum_{i=1}^N \mathbf{1} \left\{ (\widehat{S}_i(\infty), \widehat{A}_i(\infty)) \neq (S_i(\infty), A_i(\infty)) \right\} \right] = \mathbb{E} [d(\widehat{\mathbf{S}}(\infty))] \cdot D_{\text{avg}}, \quad (3.25)$$

where $d(\widehat{\mathbf{s}})$ denotes the instantaneous disagreement rate when the virtual states are $\widehat{\mathbf{S}}(t) = \widehat{\mathbf{s}}$; D_{avg} denotes the long-run average length of the disagreement periods.

3.E.2 Lemmas and the proof of Theorem 3.2

In this section, we will prove Theorem 3.2. We first rigorously prove Lemma 3.13.

Proof of Lemma 3.13. The first equation (3.23) is obvious because each virtual action $\widehat{A}_i(t)$ is sampled from $\pi(\cdot | \widehat{S}_i(t))$ independent of anything else.

We only need to show (3.24). Recall that by the definition of FTVA ($\bar{\pi}$), $\widehat{S}_i(t+1)$ could either be equal to $S_i(t+1)$ if $(\widehat{S}_i(t), \widehat{A}_i(t)) = (S_i(t), A_i(t))$, or generated afresh independent of anything else if $(\widehat{S}_i(t), \widehat{A}_i(t)) \neq (S_i(t), A_i(t))$. Let $\mathcal{I}_{\text{good}}(t) = \{i: (\widehat{S}_i(t), \widehat{A}_i(t)) = (S_i(t), A_i(t))\}$. Then

$$\begin{aligned}
& \mathbb{P}\left(\widehat{\mathbf{S}}(t+1) = \widehat{\mathbf{s}} \mid \widehat{\mathbf{S}}(t), \widehat{\mathbf{A}}(t), \dots, \widehat{\mathbf{S}}(0), \widehat{\mathbf{A}}(0)\right) \\
&= \prod_{i \in \mathcal{I}_{\text{good}}(t)} \mathbb{P}\left(S_i(t+1) = \widehat{s}_i \mid S_i(t), A_i(t)\right) \\
&\quad \cdot \prod_{i \notin \mathcal{I}_{\text{good}}(t)} \mathbb{P}\left(\widehat{S}_i(t+1) = \widehat{s}_i \mid \widehat{S}_i(t), \widehat{A}_i(t)\right) \\
&= \prod_{i \in \mathcal{I}_{\text{good}}(t)} P(S_i(t), A_i(t), \widehat{s}_i) \prod_{i \notin \mathcal{I}_{\text{good}}(t)} P(\widehat{S}_i(t), \widehat{A}_i(t), \widehat{s}_i) \\
&= \prod_{i \in [N]} P(\widehat{S}_i(t), \widehat{A}_i(t), \widehat{s}_i),
\end{aligned}$$

where the last equality is by the definition of $\mathcal{I}_{\text{good}}(t)$.

Combining (3.23) and (3.24), we have confirmed that each virtual process is an independent single-armed process with transition kernel P and policy $\bar{\pi}$. Because the initial virtual states $\{\widehat{S}_i(0)\}_{i \in [N]}$ are sampled i.i.d. from the stationary distribution of $\bar{\pi}$, $\{\widehat{S}_i(t)\}_{i \in [N]}$ are also i.i.d. with the same distribution. Therefore, $\{(\widehat{S}_i(t), \widehat{A}_i(t))\}_{i \in [N]}$ are i.i.d. with distribution $(y(s, a))_{s \in \mathbb{S}, a \in \mathbb{A}}$. $\square$

The next two lemmas bound the two quantities in Lemma 3.15, the long-run average disagreement rate and the long-run average length of the disagreement periods.

Lemma 3.16 (Average disagreement rates). *The instantaneous disagreement rate is equal to*

$$d(\widehat{\mathbf{s}}) = \mathbb{E} \left[\left| \sum_{i=1}^N \widehat{A}_i(t) - \alpha N \right| \mid \widehat{\mathbf{S}}(t) = \widehat{\mathbf{s}} \right]. \quad (3.26)$$

The long-run average disagreement rate is bounded as

$$\mathbb{E} \left[d(\widehat{\mathbf{S}}(\infty)) \right] \leq \frac{1}{2} \sqrt{N}. \quad (3.27)$$

Proof. We first prove the expression of instantaneous disagreement rate in (3.26). By definition, the instantaneous disagreement rate is equal to the expected number of arms such that $A_i(t) \neq \hat{A}_i(t)$ conditioning on the virtual states being $\hat{\mathbf{s}}$. Because FTVA tries to match as many $A_i(t)$'s to $\hat{A}_i(t)$'s as possible, there are exactly $|\sum_{i=1}^N \hat{A}_i(t) - \alpha N|$ arms such that $A_i(t) \neq \hat{A}_i(t)$. This proves (3.26).

To bound the average disagreement rate, letting $\hat{\mathbf{s}} = \hat{\mathbf{S}}(\infty)$ in (3.26) and taking expectation, we have

$$\begin{aligned} \mathbb{E} [d(\hat{\mathbf{S}}(\infty))] &= \mathbb{E} \left[\left| \sum_{i=1}^N \hat{A}_i(\infty) - \alpha N \right| \right] \\ &\leq \left(\mathbb{E} \left[\left(\sum_{i=1}^N \hat{A}_i(\infty) - \alpha N \right)^2 \right] \right)^{1/2}. \end{aligned} \quad (3.28)$$

By Lemma 3.13, the virtual actions $\hat{A}_i(\infty)$'s are i.i.d. binomial random variables such that $\mathbb{P}(\hat{A}_i(\infty) = 1) = \sum_{s \in \mathcal{S}} y(s, 1) = \alpha$, $\sum_{i=1}^N \hat{A}_i(\infty)$ has distribution $\text{Binomial}(N, \alpha)$, whose mean and variance are αN and $N\alpha(1 - \alpha) \leq N/4$. Therefore the expression in (3.28) is equal to $\text{Var} \left[\sum_{i=1}^N \hat{A}_i(\infty) \right]^{1/2} = \sqrt{N}/2$. $\square$

Lemma 3.17 (Average length of disagreement periods).

$$D_{\text{avg}} \leq \bar{\tau}_{\text{max}}^{\text{sync}}. \quad (3.29)$$

Proof. To bound D_{avg} , it suffices to bound the expected length of a disagreement period with arbitrary initial states. Without loss of generality, consider a disagreement period on arm i that starts at time $t_{\text{begin}} = 0$, with initial states $(S_i(0), A_i(0), \hat{S}_i(0), \hat{A}_i(0)) = (s, a, \hat{s}, \hat{a})$. During the disagreement period, the i -th arm can be seen as a leader-and-follower system, where the real state-action pair $(S_i(t), A_i(t))$ corresponds to the follower arm and the virtual state-action pair $(\hat{S}_i(t), \hat{A}_i(t))$ corresponds to the leader arm. By Assumption 3.1, the leader arm and the follower arm will synchronize in a finite expected time. When they synchronize, the disagreement period stops, which means $t_{\text{end}} \leq \tau^{\text{sync}}(s, a, \hat{s}, \hat{a})$. Therefore, the expected length of the disagreement period satisfies

$$\mathbb{E} [t_{\text{end}} - t_{\text{begin}}] = \mathbb{E} [t_{\text{end}}] \leq \mathbb{E} [\tau^{\text{sync}}(s, a, \hat{s}, \hat{a})] \leq \bar{\tau}_{\text{max}}^{\text{sync}}.$$

This holds for arbitrary initial states $(s, a, \hat{s}, \hat{a})$, so $D_{\text{avg}} \leq \bar{\tau}_{\text{max}}^{\text{sync}}$. $\square$

Given the three lemmas above, we can prove Theorem 3.2.

Proof of Theorem 3.2. Combining Lemma 3.15, 3.17 and 3.16, we have

$$\mathbb{E} \left[\sum_{i=1}^N \mathbb{1} \left\{ (\hat{S}_i(\infty), \hat{A}_i(\infty)) \neq (S_i(\infty), A_i(\infty)) \right\} \right] \leq \frac{1}{2} \bar{\tau}_{\text{max}}^{\text{sync}} \sqrt{N}.$$

Plugging the above inequality into the bound on conversion loss in (3.13), we get

$$\begin{aligned} V_1^{\bar{\pi}} - V_N^{\text{FTVA}(\bar{\pi})} &= \frac{1}{N} \mathbb{E} \left[\sum_{i=1}^N r(\hat{S}_i(\infty), \hat{A}_i(\infty)) - \sum_{i=1}^N r(S_i(\infty), A_i(\infty)) \right] \\ &\leq \frac{2r_{\text{max}}}{N} \mathbb{E} \left[\sum_{i=1}^N \mathbb{1} \left\{ (\hat{S}_i(\infty), \hat{A}_i(\infty)) \neq (S_i(\infty), A_i(\infty)) \right\} \right] \\ &\leq \frac{r_{\text{max}} \bar{\tau}_{\text{max}}^{\text{sync}}}{\sqrt{N}}. \end{aligned}$$

This finishes the proof. $\square$

3.F Proofs for continuous-time RBs

3.F.1 Preliminary: finite synchronization time

In this section, we prove that the synchronization time has a finite expectation in the continuous-time setting. Unlike the discrete-time setting, we do not need to make any additional assumptions other than the standard unichain condition. Our proof is based on the observation that the holding time distribution of a continuous-time Markov chain has support on the whole positive real line, so an arbitrary number of transitions can happen in any time interval, which, from the uniformization perspective, implies self-loops in all states. We can thus prove that synchronization time has a finite expectation using similar logic as in the proof of Proposition 3.7 and 3.8.

Lemma 3.18 (Synchronization in continuous time). *Consider the single-armed policy $\bar{\pi}$ and the corresponding leader-and-follower system. Given the initial states $(S(0), \hat{S}(0)) = (s, \hat{s})$, the synchronization time is bounded as*

$$\mathbb{E} [\tau^{\text{sync}}(s, \hat{s})] < \infty.$$

Proof of Lemma 3.18. Consider the always-1 policy $\bar{\pi}_1$ given by

$$\bar{\pi}_1(a|s) = \begin{cases} 1, & \text{if } a = 1, \\ 0, & \text{if } a = 0. \end{cases} \quad \text{for } s \in \mathbb{S}, a \in \mathbb{A}.$$

By the unichain property, $\bar{\pi}_1$ only has a single recurrent class, which we denote as $\mathcal{S}_1$. We denote the recurrent class under $\bar{\pi}$ as $\mathcal{S}$.

We first prove by contradiction that $\mathcal{S} \cap \mathcal{S}_1 \neq \emptyset$. Suppose $\mathcal{S} \cap \mathcal{S}_1 = \emptyset$. Then we define a new policy $\bar{\pi}'$ as

$$\bar{\pi}'(a|s) = \begin{cases} \bar{\pi}(a|s), & \text{if } s \in \mathcal{S}, \\ \bar{\pi}_1(a|s), & \text{otherwise.} \end{cases} \quad \text{for } s \in \mathbb{S}, a \in \mathbb{A}.$$

Then $\mathcal{S}$ and $\mathcal{S}_1$ are two distinct recurrent classes under $\bar{\pi}'$. This is because by definition of $\bar{\pi}'$, an arm with the initial state in $\mathcal{S}$ remains in $\mathcal{S}$, so it cannot reach $\mathcal{S}_1$; similarly an arm cannot reach $\mathcal{S}$ from $\mathcal{S}_1$. The existence of two recurrent classes contradicts the unichain condition. Therefore, we must have $\mathcal{S} \cap \mathcal{S}_1 \neq \emptyset$.

We show that given any pair of initial states $s, \hat{s} \in \mathbb{S}$, the probability that $\tau^{\text{sync}}(s, \hat{s}) < 3$ is positive. We construct a sequence of positive probability events that leads to $S(t) = \hat{S}(t)$ before time 3:

1. The leader arm reaches a state s^a such that $\bar{\pi}(1|s^a) > 0$ by time 1 and chooses action 1;
2. The leader arm has no transition during $[1, 2]$, so $\hat{S}(t) = s^a$ and $\hat{A}(t) = 1$ for all $t \in [1, 2]$;
3. The follower arm applies the same action as the leader arm during $[1, 2]$, so $A(t) = 1$ for all $t \in [1, 2]$; and reaches the state s^b by time 2 for some $s^b \in \mathcal{S} \cap \mathcal{S}_1$;
4. The follower arm stays at s^b during $[2, 3]$, and the leader arm reaches s^b by time 3, so the two arms synchronize.

The first two events have positive probabilities because the policy $\bar{\pi}$ applies action 1 with a positive fraction of time. To see why the third event has a positive probability, observe that the follower arm applies action 1 for all $t \in [1, 2]$ regardless of its state, so it is effectively under policy $\bar{\pi}_1$ and could traverse all states in $\mathcal{S}_1$. The fourth event has a positive probability because the leader arm under $\bar{\pi}$ can traverse all states in $\mathcal{S}$. Note that in the above arguments, we are implicitly assuming that the two arms do not synchronize

in the middle of the four events, because otherwise we are done. Also, we use the fact that in a continuous-time Markov chain, there can be an arbitrary number of transitions during any time interval. Therefore, we have proved that for all $s, \hat{s} \in \mathbb{S}$,

$$\mathbb{P}(\tau^{\text{sync}}(s, \hat{s}) < 3) > 0.$$

Now we prove by induction that for all $s, \hat{s} \in \mathbb{S}$ and $k = 0, 1, 2, \dots$,

$$\mathbb{P}(\tau^{\text{sync}}(s, \hat{s}) \geq 3k) \leq \left(\max_{s', \hat{s}' \in \mathbb{S}} \mathbb{P}(\tau^{\text{sync}}(s', \hat{s}') \geq 3) \right)^k. \quad (3.30)$$

The base case of $k = 1$ is already known. Suppose we have proved (3.30) for a certain k , then

$$\begin{aligned} & \mathbb{P}(\tau^{\text{sync}}(s, \hat{s}) \geq 3(k+1)) \\ &= \mathbb{P}(\tau^{\text{sync}}(s, \hat{s}) \geq 3k) \cdot \mathbb{P}(\tau^{\text{sync}}(s, \hat{s}) \geq 3(k+1) \mid \tau^{\text{sync}}(s, \hat{s}) \geq 3k) \\ &= \mathbb{P}(\tau^{\text{sync}}(s, \hat{s}) \geq 3k) \cdot \mathbb{P}(\tau^{\text{sync}}(S(3k), \hat{S}(3k)) \geq 3 \mid S(3k) \neq \hat{S}(3k)) \\ &\leq \mathbb{P}(\tau^{\text{sync}}(s, \hat{s}) \geq 3k) \cdot \max_{s', \hat{s}' \in \mathbb{S}} \mathbb{P}(\tau^{\text{sync}}(s', \hat{s}') \geq 3) \\ &\leq \left(\max_{s', \hat{s}' \in \mathbb{S}} \mathbb{P}(\tau^{\text{sync}}(s', \hat{s}') \geq 3) \right)^{k+1}. \end{aligned}$$

where in the second equality we have used the Markov property of the system, and in the last inequality we apply the induction hypothesis. This proves (3.30).

The bound on the expectation $\mathbb{E}[\tau^{\text{sync}}(s, \hat{s})]$ follows summing the tail bound (3.30) over $k = 0, 1, 2, \dots$:

$$\begin{aligned} \mathbb{E}[\tau^{\text{sync}}(s, \hat{s})] &\leq \sum_{k=0}^{\infty} 3\mathbb{P}(\tau^{\text{sync}}(s, \hat{s}) \geq 3k) \\ &\leq \frac{3}{1 - \max_{s', \hat{s}' \in \mathbb{S}} \mathbb{P}(\tau^{\text{sync}}(s', \hat{s}') \geq 3)} \\ &< \infty, \end{aligned}$$

where the second inequality is due to (3.30) and the fact that

$$\max_{s', \hat{s}' \in \mathbb{S}} \mathbb{P}(\tau^{\text{sync}}(s', \hat{s}') \geq 3) \leq 1 - \min_{s', \hat{s}' \in \mathbb{S}} \mathbb{P}(\tau^{\text{sync}}(s', \hat{s}') < 3) < 1.$$

This finishes the proof. $\square$

3.F.2 Overview of the proof of Theorem 3.3

In this section, we focus on proving the bound (3.20) on the conversion loss $V_1^{\bar{\pi}} - V_N^{\text{FTVA-CT}(\bar{\pi})}$ for any single-armed policy $\bar{\pi}$. The asymptotic optimality result (3.21) in Theorem 3.3 will be a direct consequence of (3.20) if we take $\bar{\pi}$ to be any optimal single-armed policy $\bar{\pi}^*$.

We first show that under FTVA-CT ($\bar{\pi}$) defined in Algorithm 2, $\{(\mathbf{S}(t), \hat{\mathbf{S}}(t))\}_{t \geq 0}$ is a continuous-time Markov chain and each virtual process is an independent Markov chain induced by applying $\bar{\pi}$ to the single-armed system, as stated in Lemma 3.19 below. We will prove it in the next subsection.

Lemma 3.19. *Under the FTVA-CT ($\bar{\pi}$) defined in Algorithm 2, we have the following:*

- (1) $\{(\mathbf{S}(t), \hat{\mathbf{S}}(t))\}_{t \geq 0}$ is a continuous-time Markov chain.
- (2) $\{\hat{S}_i(t)\}_{t \geq 0}$ for $i \in [N]$ are N independent Markov chains, whose transition rate from state s to s' is $\sum_{a \in \mathbb{A}} G(s, a, s') \bar{\pi}(a|s)$, for $s, s' \in \mathbb{S}$ s.t. $s \neq s'$.

Let $(y(s, a))_{s \in \mathbb{S}, a \in \mathbb{A}}$ be the steady-state state-action distribution of the single-armed system under $\bar{\pi}$. Then the second result above implies that for each $t \geq 0$, $(\hat{S}_i(t), \hat{A}_i(t))$ for $i \in [N]$ are i.i.d. with the distribution $(y(s, a))_{s \in \mathbb{S}, a \in \mathbb{A}}$.

Next, we can upper bound the conversion loss $V_1^{\bar{\pi}} - V_N^{\text{FTVA-CT}(\bar{\pi})}$ as

$$\begin{aligned}
 & V_1^{\bar{\pi}} - V_N^{\text{FTVA-CT}(\bar{\pi})} \\
 &= \frac{1}{N} \mathbb{E} \left[\sum_{i=1}^N r(\hat{S}_i(\infty), \hat{A}_i(\infty)) - \sum_{i=1}^N r(S_i(\infty), A_i(\infty)) \right] \\
 &\leq \frac{2r_{\max}}{N} \mathbb{E} \left[\sum_{i=1}^N \mathbb{1} \left\{ (\hat{S}_i(\infty), \hat{A}_i(\infty)) \neq (S_i(\infty), A_i(\infty)) \right\} \right] \\
 &\leq \frac{2r_{\max}}{N} \mathbb{E} \left[\sum_{i=1}^N \mathbb{1} \left\{ \hat{S}_i(\infty) \neq S_i(\infty) \right\} + \sum_{i=1}^N \mathbb{1} \left\{ \hat{A}_i(\infty) \neq A_i(\infty) \right\} \right]. \quad (3.31)
 \end{aligned}$$

The bound is slightly different from the (3.13) since we want to separately deal with the number of arms whose real and virtual states do not agree and those whose real and virtual actions do not agree, as shown in the lemma below.

Lemma 3.20. *It holds that*

$$\mathbb{E} \left[\sum_{i=1}^N \mathbb{1} \left\{ \hat{S}_i(\infty) \neq S_i(\infty) \right\} \right] \leq 2g_{\max} \bar{\tau}_{\max}^{\text{sync}} \mathbb{E} \left[\left| \sum_{i=1}^N \hat{A}_i(\infty) - \alpha N \right| \right] \quad (3.32)$$

$$\mathbb{E} \left[\sum_{i=1}^N \mathbb{1} \left\{ \hat{A}_i(\infty) \neq A_i(\infty) \right\} \right] = \mathbb{E} \left[\left| \sum_{i=1}^N \hat{A}_i(\infty) - \alpha N \right| \right] \quad (3.33)$$

Before showing the proof of Lemma 3.20, we will first use Lemma 3.20 to prove Theorem 3.3.

Proof of Theorem 3.3. Combining (3.31) and Lemma 3.20, we have

$$V_1^{\bar{\pi}} - V_N^{\text{FTVA-CT}(\bar{\pi})} \leq \frac{2r_{\max}}{N} (1 + 2g_{\max} \bar{\tau}_{\max}^{\text{sync}}) \mathbb{E} \left[\left| \sum_{i=1}^N \hat{A}_i(\infty) - \alpha N \right| \right].$$

By Lemma 3.19, the virtual actions $\hat{A}_i(\infty)$'s are i.i.d. binomial random variables such that $\mathbb{P}(\hat{A}_i(\infty) = 1) = \sum_{s \in \mathbb{S}} y(s, a) = \alpha$, so the distribution of $\sum_{i=1}^N \hat{A}_i(\infty)$ is Binomial(N, α), whose mean and variance are αN and $N\alpha(1 - \alpha) \leq N/4$. Therefore,

$$\begin{aligned} \mathbb{E} \left[\left| \sum_{i=1}^N \hat{A}_i(\infty) - \alpha N \right| \right] &\leq \left(\mathbb{E} \left[\left(\sum_{i=1}^N \hat{A}_i(\infty) - \alpha N \right)^2 \right] \right)^{1/2} \\ &= \text{Var} \left[\sum_{i=1}^N \hat{A}_i(\infty) \right]^{1/2} \\ &\leq \frac{1}{2} \sqrt{N}. \end{aligned}$$

Combining the above calculations, we get

$$V_1^{\bar{\pi}} - V_N^{\text{FTVA-CT}(\bar{\pi})} \leq \frac{r_{\max} (1 + 2g_{\max} \bar{\tau}_{\max}^{\text{sync}})}{\sqrt{N}}.$$

□

3.F.3 Proof of Lemma 3.19

Proof. We first show that $\{(\mathbf{S}(t), \hat{\mathbf{S}}(t))\}_{t \geq 0}$ is a continuous-time Markov chain. Observe that $\{(\mathbf{S}(t), \hat{\mathbf{S}}(t))\}_{t \geq 0}$ is piecewise constant between decision epochs $\{t_k\}_{k=0,1,\dots}$, so it suffices to consider the time between the decision epochs $t_{k+1} - t_k$ and the states at the decision epochs $(\mathbf{S}(t_k), \hat{\mathbf{S}}(t_k))$ for $k = 0, 1, 2, \dots$

We claim that at each decision epoch t_k , the time until the next decision epoch $t_{k+1} - t_k$ is exponentially distributed conditioned on $(\mathbf{S}(t_k), \hat{\mathbf{S}}(t_k))$. Observe from the pseudo-code in Algorithm 2 that at the decision epoch t_k , conditioned on $(\mathbf{S}(t_k), \mathbf{A}(t_k), \hat{\mathbf{S}}(t_k))$, the time until the next decision epoch t_{k+1} is the minimum of two independent exponential random variables, corresponding to the exponential timer and the transition of real states. The rates of the two exponential random variables are $2Ng_{\max} - g_k^{\text{real}}$ and g_k^{real} , where $g_k^{\text{real}} = \sum_{i=1}^N G(S_i(t_k), A_i(t_k))$. Therefore, conditioned on $(\mathbf{S}(t_k), \mathbf{A}(t_k), \hat{\mathbf{S}}(t_k))$, $t_{k+1} - t_k$ has an exponential distribution with rate $2Ng_{\max} - g_k^{\text{real}} + g_k^{\text{real}} = 2Ng_{\max}$. Because the rate $2Ng_{\max}$ is a constant, if we take expectation over $\mathbf{A}(t_k)$ and only conditioned on $(\mathbf{S}(t_k), \hat{\mathbf{S}}(t_k))$, the time until the next decision epoch $t_{k+1} - t_k$ is still exponentially distributed with rate $2Ng_{\max}$.

Also, it is obvious from the pseudo-code in Algorithm 2 that conditioned on $(\mathbf{S}(t_k), \hat{\mathbf{S}}(t_k))$, the distribution of $(\mathbf{S}(t_{k+1}), \hat{\mathbf{S}}(t_{k+1}))$ is independent of $t_{k+1} - t_k$ and $(\mathbf{S}(t), \hat{\mathbf{S}}(t))$ for $t < t_k$. Therefore, we conclude that $\{(\mathbf{S}(t), \hat{\mathbf{S}}(t))\}_{t \geq 0}$ is a continuous-time Markov chain.

Next, we show that the $\{\hat{S}_i(t)\}_{t \geq 0}$ for $i \in [N]$ are N independent Markov chains, whose transition rate from state s to s' is $\sum_{a \in \mathbf{A}} G(s, a, s') \bar{\pi}(a|s)$ for $s' \neq s$. Because we have shown that $\{(\mathbf{S}(t), \hat{\mathbf{S}}(t))\}_{t \geq 0}$ is a continuous-time Markov chain with a constant transition rate, it suffices to focus on the embedded chain $\{(\mathbf{S}(t_k), \hat{\mathbf{S}}(t_k))\}_{k=0,1,\dots}$ and examine the probability of the transitions that change the virtual states $\hat{\mathbf{S}}(t_k)$. Note that between two decision epochs t_k and t_{k+1} , there is at most one i such that $\hat{S}_i(t)$ changes. Let $\mathcal{I}_{\text{good}}(t_k) = \{i \in [N] : \hat{S}_i(t_k) = S_i(t_k), \hat{A}_i(t_k) = A_i(t_k)\}$. For any arm $i \in \mathcal{I}_{\text{good}}(t_k)$, the transitions of its virtual and real states are coupled, which implies that the probability that $\hat{S}_i(t_{k+1}) = s'$ is the same as the probability that $S_i(t_{k+1}) = s'$, for any $s' \neq \hat{S}_i(t_k)$, conditioned on the states and actions at t_k . Formally,

$$\mathbb{P}(\hat{S}_i(t_{k+1}) = s' \mid \mathbf{S}(t_k), \mathbf{A}(t_k), \hat{\mathbf{S}}(t_k), \hat{\mathbf{A}}(t_k)) \mathbf{1}\{i \in \mathcal{I}_{\text{good}}(t_k)\}$$

$$\begin{aligned}
&= \mathbb{P}\left(S_i(t_{k+1}) = s' \middle| \mathbf{S}(t_k), \mathbf{A}(t_k), \widehat{\mathbf{S}}(t_k), \widehat{\mathbf{A}}(t_k)\right) \mathbb{1}\{i \in \mathcal{I}_{\text{good}}(t_k)\} \\
&= \frac{G(S_i(t_k), A_i(t_k), s')}{2Ng_{\max}} \mathbb{1}\{i \in \mathcal{I}_{\text{good}}(t_k)\} \\
&= \frac{G(\widehat{S}_i(t_k), \widehat{A}_i(t_k), s')}{2Ng_{\max}} \mathbb{1}\{i \in \mathcal{I}_{\text{good}}(t_k)\}, \tag{3.34}
\end{aligned}$$

where the second equality uses the fact that the transition probability after uniformization is equal to the ratio between the transition rate and the uniformization rate; the last equality is by the definition of $\mathcal{I}_{\text{good}}(t_k)$. For $i \notin \mathcal{I}_{\text{good}}(t_k)$ and $s' \neq \widehat{S}_i(t_k)$, we have $\widehat{S}_i(t_{k+1}) = s'$ only when the exponential clock ticks and the pair (i, s') is sampled, so

$$\begin{aligned}
&\mathbb{P}\left(\widehat{S}_i(t_{k+1}) = s' \middle| \mathbf{S}(t_k), \mathbf{A}(t_k), \widehat{\mathbf{S}}(t_k), \widehat{\mathbf{A}}(t_k)\right) \mathbb{1}\{i \notin \mathcal{I}_{\text{good}}(t_k)\} \\
&= \mathbb{E} \left[\frac{2Ng_{\max} - g_k^{\text{real}}}{2Ng_{\max}} \frac{G(\widehat{S}_i(t_k), \widehat{A}_i(t_k), s')}{2Ng_{\max} - g_k^{\text{real}}} \middle| \mathbf{S}(t_k), \mathbf{A}(t_k), \widehat{\mathbf{S}}(t_k), \widehat{\mathbf{A}}(t_k) \right] \\
&\quad \cdot \mathbb{1}\{i \notin \mathcal{I}_{\text{good}}(t_k)\} \\
&= \frac{G(\widehat{S}_i(t_k), \widehat{A}_i(t_k), s')}{2Ng_{\max}} \mathbb{1}\{i \notin \mathcal{I}_{\text{good}}(t_k)\}. \tag{3.35}
\end{aligned}$$

Summing up (3.34)(3.35), we have that for any $i \in [N]$ and $s' \neq \widehat{S}_i(t_k)$,

$$\mathbb{P}\left(\widehat{S}_i(t_{k+1}) = s' \middle| \mathbf{S}(t_k), \mathbf{A}(t_k), \widehat{\mathbf{S}}(t_k), \widehat{\mathbf{A}}(t_k)\right) = \frac{G(\widehat{S}_i(t_k), \widehat{A}_i(t_k), s')}{2Ng_{\max}}. \tag{3.36}$$

Because $\widehat{A}_i(t_k)$ is independently sampled from the distribution $\bar{\pi}(\cdot | \widehat{S}_i(t_k))$, taking expectation over $(\mathbf{A}(t_k), \widehat{\mathbf{A}}(t_k))$ in the above equation, we get

$$\mathbb{P}\left(\widehat{S}_i(t_{k+1}) = s' \middle| \mathbf{S}(t_k), \widehat{\mathbf{S}}(t_k)\right) = \frac{\sum_{a \in \mathbb{A}} G(\widehat{S}_i(t_k), a, s') \bar{\pi}(a | \widehat{S}_i(t_k))}{2Ng_{\max}}, \tag{3.37}$$

for any $i \in [N]$ and $s' \neq \widehat{S}_i(t_k)$. Because the uniformization rate is $2Ng_{\max}$, the rate for $\widehat{S}_i(t)$ to transition to s' is equal to $\sum_{a \in \mathbb{A}} G(\widehat{S}_i(t), a, s') \bar{\pi}(a | \widehat{S}_i(t))$ conditioned on $(\mathbf{S}(t), \widehat{\mathbf{S}}(t))$. Observe that this transition rate only depends on $\widehat{S}_i(t)$, which implies that $\widehat{S}_i(t)$'s for $i \in [N]$ are N i.i.d. Markov chains.

Finally, recall that we initialize the virtual states $\widehat{S}_i(0)$'s as N i.i.d. samples from the stationary distribution of the single-armed system under the policy

$\bar{\pi}$. Since we have proved that $\{\widehat{S}_i(t)\}_{t \geq 0}$'s are N i.i.d. Markov chains induced by applying $\bar{\pi}$ to the single-armed systems, for each $t \geq 0$, $\widehat{S}_i(t)$'s remain stationary and i.i.d. Therefore, for each $t \geq 0$, $(\widehat{S}_i(t), \widehat{A}_i(t))$'s are i.i.d., and for each i , the distribution of $(\widehat{S}_i(t), \widehat{A}_i(t))$ is equal to the steady-state state-action distribution $(y(s, a))_{s \in \mathbb{S}, a \in \mathbb{A}}$. $\square$

3.F.4 Proof of Lemma 3.20

In this section, we prove the key intermediate result, Lemma 3.20. The second equation (3.33) in Lemma 3.20 is obvious from the definition of the policy. We can therefore focus on proving (3.32), i.e., bounding the long-run average number of arms whose real states are not equal to their virtual states. We call such arms *bad arms*, and the rest of the arms *good arms*.

We will use a similar approach as Appendix 3.E: we invoke Little's Law to write the average number of bad arms as a product of the average *disagreement rate* and the average length of *disagreement periods*, based on a suitable definition of the *disagreement events*.

We define the disagreement event in a different way than in the discrete-time RB setting, because unlike in the discrete-time RB setting where $A_i(t) \neq \widehat{A}_i(t)$ can immediately cause $S_i(t) \neq \widehat{S}_i(t)$, in the continuous-time RB setting, actions that last for only a short time may not have an effect on the states. Therefore, we say a disagreement event for arm i happens at time t only when arm i behaves differently than it would have behaved if $A_i(t) = \widehat{A}_i(t)$.

Specifically, the disagreement event of a good arm can be defined as having a state transition that turns it into a bad arm, since a good arm is supposed to remain good if we always have $A_i(t) = \widehat{A}_i(t)$; for bad arms, how the arm "would have behaved" need to be defined with the help of an extra structure, namely, exponential timers, as described below.

Definition 3.21 (Exponential timers for simulating the leader-and-follower system). For each $i \in [N]$, we run a timer for each arm i that ticks every random amount of time. When the timer of arm i starts, it decides the time of its next tick based on the i -th arm $(S_i(t), A_i(t), \widehat{S}_i(t), \widehat{A}_i(t))$: if $S_i(t) \neq \widehat{S}_i(t)$ and $A_i(t) \neq \widehat{A}_i(t)$, the timer ticks after a time that is exponentially distributed with rate $G(S_i(t), \widehat{A}_i(t))$; otherwise, the timer pauses. The timer restarts when it ticks or when there is any event happening in the system.

The purpose of the exponential timer is to simulate the transition time of the leader-and-follower system, which characterizes how the arm would have

behaved if $A_i(t) = \widehat{A}_i(t)$. Specifically, when arm i has $S_i(t) \neq \widehat{S}_i(t)$ (so it is a bad arm) and $A_i(t) \neq \widehat{A}_i(t)$, we construct an imaginary leader-and-follower system whose state-action pairs are $(S_i(t), \widehat{A}_i(t), \widehat{S}_i(t), \widehat{A}_i(t))$. We let the transition times of the virtual state $\widehat{S}_i(t)$ in the two systems be identical, and the transition times of the real state $S_i(t)$ in the two systems be independent. Then the transition time of $S_i(t)$ in the leader-and-follower system has an exponential distribution with rate $G(S_i(t), \widehat{A}_i(t))$, which is equal to the time that the exponential timer ticks.

Therefore, there are two ways that the transition of arm i with $S_i(t) \neq \widehat{S}_i(t)$ and $A_i(t) \neq \widehat{A}_i(t)$ can deviate from the leader-and-follower system described above: either arm i itself has a transition in the real state $S_i(t)$, or when the exponential timer ticks. This statement is actually also true if $S_i(t) = \widehat{S}_i(t)$ and $A_i(t) \neq \widehat{A}_i(t)$ because in that case, the exponential timer pauses.

We can thus formally define the *disagreement event of arm i* as below.

Definition 3.22 (Disagreement event for continuous-time RBs). For each $i \in [N]$, a disagreement event of arm i happens when its real state $S_i(t)$ transitions or its exponential timer ticks, while $A_i(t) \neq \widehat{A}_i(t)$.

After defining the disagreement events, we can define the *disagreement period of arm i* as the period of time when $S_i(t) \neq \widehat{S}_i(t)$, separated by disagreement events, formally stated below.

Definition 3.23 (Disagreement period for continuous-time RBs). Given a sample path of the arm i 's real states and virtual states $(S_i(t), \widehat{S}_i(t))$, and an exponential timer, we define the disagreement period of the arm i as a time interval $[t_{\text{begin}}, t_{\text{end}})$ such that

There is a disagreement event at t_{begin} ;

There is no disagreement event during $(t_{\text{begin}}, t_{\text{end}})$,

and $S_i(t) \neq \widehat{S}_i(t)$ for $t \in (t_{\text{begin}}, t_{\text{end}})$;

There is a disagreement event at t_{end} or $S_i(t_{\text{end}}) = \widehat{S}_i(t_{\text{end}})$.

We let $d(\mathbf{s}, \widehat{\mathbf{s}})$ be the *instantaneous rate of disagreement events* (instantaneous disagreement rate) when the system has real and virtual states $(\mathbf{S}(t), \widehat{\mathbf{S}}(t)) = (\mathbf{s}, \widehat{\mathbf{s}})$. Let D_{avg} be the long-run average length of the disagreement periods.

Observe that the number of bad arms (the arms such that $S_i(t) \neq \widehat{S}_i(t)$) is the number of arms in disagreement periods, so we can apply Little's Law to the number of bad arms.

Lemma 3.24 (Little's Law for disagreement periods, continuous-time version). *It holds that*

$$\mathbb{E} \left[\sum_{i=1}^N \mathbb{1} \left\{ \widehat{S}_i(\infty) \neq S_i(\infty) \right\} \right] = \mathbb{E} \left[d(\mathbf{S}(\infty), \widehat{\mathbf{S}}(\infty)) \right] \cdot D_{avg}, \quad (3.38)$$

where $d(\mathbf{s}, \widehat{\mathbf{s}})$ denotes the instantaneous disagreement rate when the virtual states are $\widehat{\mathbf{S}}(t) = \widehat{\mathbf{s}}$; D_{avg} denotes the long-run average length of the disagreement periods.

Lemma 3.25 (Average disagreement rates). *The instantaneous disagreement rate is equal to*

$$d(\mathbf{s}, \widehat{\mathbf{s}}) \leq 2g_{\max} \mathbb{E} \left[\left| \sum_{i=1}^N \widehat{A}_i(t) - \alpha N \right| \mid \widehat{\mathbf{S}}(t) = \widehat{\mathbf{s}} \right]. \quad (3.39)$$

Proof. For each arm i such that $A_i(t) \neq \widehat{A}_i(t)$, a disagreement event happens if its real state transitions, which happens at the rate $G(S_i(t), A_i(t))$, or if its exponential timer ticks, which happens at the rate 0 (if it is a good arm) or $G(S_i(t), \widehat{A}_i(t))$ (if it is a bad arm). Therefore, the rate that disagreement events happen at arm i is no more than

$$G(S_i(t), A_i(t)) + G(S_i(t), \widehat{A}_i(t)) \leq 2g_{\max}.$$

By the definition of the policy, the expected number of arms with $A_i(t) \neq \widehat{A}_i(t)$ is

$$\mathbb{E} \left[\left| \sum_{i=1}^N \widehat{A}_i(t) - \alpha N \right| \mid \widehat{\mathbf{S}}(t) = \widehat{\mathbf{s}} \right],$$

so the instantaneous disagreement rate of the system is as given in (3.39). $\square$

Lemma 3.26 (Average length of disagreement periods). *It holds that*

$$D_{avg} \leq \bar{\tau}_{\max}^{\text{sync}}. \quad (3.40)$$

Proof. To bound D_{avg} , it suffices to bound the expected length of a disagreement period with arbitrary initial states. Without loss of generality, consider a disagreement period on arm i that starts at time $t_{\text{begin}} = 0$, with initial states $(S_i(0), \widehat{S}_i(0)) = (s, \widehat{s})$. During the disagreement period, there is no disagreement event, so as argued in the paragraph after Definition 3.21, the

transitions of the i -th arm is identical to a leader-and-follower system. Therefore, we either have $S_i(t) = \widehat{S}_i(t)$ after $\tau^{\text{sync}}(s, \widehat{s})$ amount of time, or have a disagreement event before that. In either case, $t_{\text{end}} \leq \tau^{\text{sync}}(s, \widehat{s})$. Therefore, the expected length of the disagreement period satisfies

$$\mathbb{E}[t_{\text{end}} - t_{\text{begin}}] = \mathbb{E}[t_{\text{end}}] \leq \mathbb{E}[\tau^{\text{sync}}(s, \widehat{s})] \leq \bar{\tau}_{\text{max}}^{\text{sync}}.$$

This holds for arbitrary initial states $(s, \widehat{s})$, so $D_{\text{avg}} \leq \bar{\tau}_{\text{max}}^{\text{sync}}$. $\square$

Proof of Lemma 3.20. Combining Lemma 3.24, 3.25, and 3.26, we have

$$\begin{aligned} \mathbb{E} \left[\sum_{i=1}^N \mathbb{1} \left\{ \widehat{S}_i(\infty) \neq S_i(\infty) \right\} \right] &= \mathbb{E} \left[d(\mathbf{S}(\infty), \widehat{\mathbf{S}}(\infty)) \right] \cdot D_{\text{avg}} \\ &\leq 2g_{\text{max}} \bar{\tau}_{\text{max}}^{\text{sync}} \mathbb{E} \left[\left| \sum_{i=1}^N \widehat{A}_i(\infty) - \alpha N \right| \right]. \end{aligned}$$

This proves (3.32).

Observe that by the definition of our algorithm, at any time t ,

$$\sum_{i=1}^N \mathbb{1} \left\{ \widehat{A}_i(t) \neq A_i(t) \right\} = \left| \sum_{i=1}^N \widehat{A}_i(t) - \alpha N \right| \quad a.s.$$

Taking the steady-state expectation, we get (3.33). $\square$

3.G Experiment details and additional experiments

In this section, we include the details of the experiments in the main body, as well as some additional experiments. In Appendix 3.G.1 and 3.G.2, we describe the details of the experiments of Figure 3.2 and 3.3. We conduct the simulations of these two experiments with different initial points and present the results in Appendix 3.G.3. To give an intuitive understanding of the difference between these policies, we also display more visualization of the sample paths under different policies in Appendix 3.G.4. ³

³Our simulation code can be found in the link <https://github.com/YigeHong/rb-break-ugap-ftva>.

3.G.1 Experiment details of Figure 3.2

Restless bandits setting In Figure 3.2, we consider the discrete-time N -arm restless bandits represented by the tuple $(N, \mathbb{S}^N, \mathbb{A}^N, P, r, \alpha N)$. We vary the number of arms N , and keep the rest of the parameters fixed. The state space of each arm is $\mathbb{S} = \{1, 2, 3\}$. The action space is $\mathbb{A} = \{0, 1\}$. The transition kernel P is given by

$$\begin{aligned} P(\cdot, 0, \cdot) &= \begin{bmatrix} 0.02232142 & 0.10229283 & 0.87538575 \\ 0.03426605 & 0.17175704 & 0.79397691 \\ 0.52324756 & 0.45523298 & 0.02151947 \end{bmatrix}, \\ P(\cdot, 1, \cdot) &= \begin{bmatrix} 0.14874601 & 0.30435809 & 0.54689589 \\ 0.56845754 & 0.41117331 & 0.02036915 \\ 0.25265570 & 0.27310439 & 0.4742399 \end{bmatrix}, \end{aligned}$$

where the number in s -th row and s' -th column in each matrix represents $P(s, 0, s')$ or $P(s, 1, s')$, for $s, s' \in \{1, 2, 3\}$. The reward function r is given by

$$\begin{aligned} r(\cdot, 0) &= [0 \quad 0 \quad 0], \\ r(\cdot, 1) &= [0.37401552 \quad 0.11740814 \quad 0.07866135], \end{aligned}$$

where the s -th entry of each vector represents $r(s, 0)$ or $r(s, 1)$, for $s \in \{1, 2, 3\}$. The budget parameter $\alpha = 0.4$, so $0.4N$ arms are pulled in each time slot. This setting is taken from the Appendix E of [105] as a counterexample to the GAP assumption. We note that this RB problem obviously satisfies Synchronization Assumption for any policy: observe that $P(s, a, s') > 0$ for all $s, s' \in \mathbb{S}, a \in \mathbb{A}$, so two arms with any initial states have a positive probability of synchronizing in the next time slot, which implies finite synchronization time by Proposition 3.5.

Simulation setting We plot the long-run average reward against the number of arms for different policies. The number of arms N varies from 100, 200, $\dots$, 1000. Three policies are considered: our policy **FTVA** ($\bar{\pi}^*$), the Whittle's index policy [283], and an LP-Priority policy [266]. For each data point, we obtain the long-run average reward and its confidence interval by simulating 50 independent trajectories, each with a length of 1000 time slots. The initial states of all arms are simply chosen to be 1, because simulation results do not quite depend on the initial states (see Appendix 3.G.3).

Detail of policies We implement Whittle index policy and LP-Priority in the standard way. The resulting priorities of both policies turn out to be $1 > 2 > 3$ (state 1 has the highest priority). This is not a coincidence given that the solution of the single-armed problem (LP) is:

$$y^*(\cdot, \cdot) = \begin{bmatrix} 0 & 0.29943 \\ 0.23768 & 0.10057 \\ 0.36232 & 0 \end{bmatrix},$$

where the s -th row represents $y^*(s, 0)$ and $y^*(s, 1)$, for $s \in \{1, 2, 3\}$. This solution implies that a reasonable policy should almost always pull arms in state 1, pull arms in state 2 for a certain fraction of time, and almost never pull arms in state 3. Therefore the only reasonable priority in this RB setting is $1 > 2 > 3$.

Our policy FTVA ($\bar{\pi}^*$) is implemented according to the pseudocode in Section 3.4. The non-trivial detail here is the tie-breaking rule for selecting the set of arms to activate based on the virtual actions. Our tie-breaking rule is to select $A_i(t)$'s such that the number of good arms, i.e., the arms such that $S_i(t) = \hat{S}_i(t)$ and $A_i(t) = \hat{A}_i(t)$, is maximized. We have experimented with alternative tie-breaking rules, whose performance matches our theoretical results, though sometimes their asymptotic optimality requires larger values of N to be observed. The relationship between various tie-breaking rules and their impact on finite- N performances remains unclear in the present analysis, leaving it as a topic for future investigation.

3.G.2 Experiment details of Figure 3.3

Restless bandits setting In Figure 3.3, we consider the discrete-time N -arm bandits constructed as below. Suppose the state space for each arm is $\mathbb{S} = \{0, 1, \dots, 7\}$. Each state has a *preferred action*, which is action 1 for states 0, 1, 2, 3, and action 0 otherwise. For an arm in state s , applying the preferred action moves the arm to state $(s + 1) \bmod 8$ with probability $p_{s,R}$, and applying the other action moves the arm to state $(s - 1)^+$ with probability $p_{s,L}$.⁴ The probabilities $p_{\cdot,R}$ and $p_{\cdot,L}$ are given by

$$p_{\cdot,R} = [0.1 \quad 0.1 \quad 0.1 \quad 0.1 \quad 0.1 \quad 0.1 \quad 0.1 \quad 0.1],$$

⁴Here the subscript L means “left”, and R means “right”. We are imagining the arms being lined up in a row from state 0 to state 7, and the preferred action moves an arm to the right.

$$p_{\cdot, \mathbb{L}} = [1.0 \quad 1.0 \quad 0.48 \quad 0.47 \quad 0.46 \quad 0.45 \quad 0.44 \quad 0.43].$$

When the arm transitions from 7 to 0, one unit of reward is generated. Equivalently, if we consider the expected reward of applying an action at a certain state, we can define the reward function as $r(7, 0) = p_{7, \mathbb{R}}$, and $r(s, a) = 0$ for all other $s \in \mathbb{S}, a \in \mathbb{A}$. The parameter $\alpha = 1/2$, so $N/2$ arms are activated in each time slot.

Simulation setting We plot the long-run average reward against the number of arms for different policies. The number of arms N varies from 100, 200, $\dots$, 1000. Three policies are considered: our policy FTVA ($\bar{\pi}^*$), a random tie-breaking policy, and a particular LP-Priority policy that prioritizes arms with larger Lagrangian optimal indices (see the definition of Lagrangian optimal indices in [56, 108, 155]). This particular LP-Priority policy is also referred to as the LP-Index policy in the literature. But for simplicity, we just refer to it as LP-Priority in this chapter. For each data point, we obtain the long-run average reward and its confidence interval by simulating 50 independent trajectories, each with a length of 1000 time slots. The initial states for the simulations are fixed: $N/3$ arms are in state 1 and $2N/3$ arms are in state 2. Note that we fix this initial point in the simulation of Figure 3.3 in order to demonstrate that both the random tie-breaking and LP-Priority can get nearly zero rewards in this example. For other choices of fixed points, there is also a strong separation between the performance of FTVA ($\bar{\pi}^*$) and random tie-breaking or LP-Priority, which we will show in the next section.

Details of policies The optimal solution of (LP) is $y^*(s, 1) = 1/8$ for $s = 0, 1, 2, 3$, $y^*(s, 0) = 1/8$ for $s = 4, 5, 6, 7$, and $y^*(s, a) = 0$ for other $s \in \mathbb{S}$ and $a \in \mathbb{A}$. Note that the same y^* remains optimal even if we remove the budget constraint. The optimal LP solution suggests that a reasonable policy should almost always pull arms in states 0, 1, 2, 3, and almost never pull arms in states 4, 5, 6, 7.

The random tie-breaking policy that we consider prioritizes arms whose states are in $\{0, 1, 2, 3\}$ over arms whose states are in $\{4, 5, 6, 7\}$, and it breaks ties uniformly at random when there are more than $N/2$ arms in either of the two sets.

The LP-Priority policy we consider prioritizes arms with larger Lagrangian optimal indices [56, 108]. Specifically, it involves solving the Lagrangian relaxation of the original LP, which replaces the budget constraint with a

penalty term determined by the optimal Lagrange multiplier. In our setting, because the optimal solution y^* remains optimal even without the budget constraint, we can simply remove the budget constraint to get the Lagrangian relaxation.⁵ The resulting Lagrangian optimal indices are given by:

$$[0.0125 \quad 0.1375 \quad 0.0725 \quad 0.07125 \quad -0.07 - 0.06875 \quad -0.0675 \quad -0.06625],$$

and the priority is $1 > 2 > 3 > 0 > 7 > 6 > 5 > 4$.

The implementation of our policy FTVA ($\bar{\pi}^*$) is the same as in the last experiment.

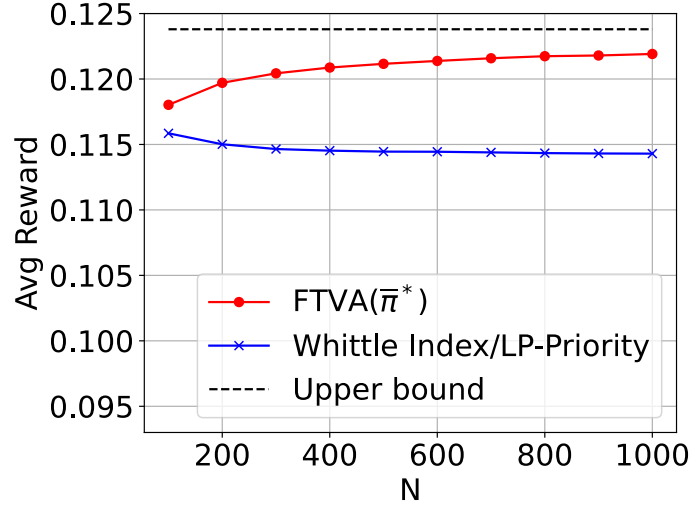
Further discussions of the policies As shown in Figure 3.3, the random tie-breaking policy and the LP-Priority policy get nearly zero rewards. We have discussed why this happens for the random tie-breaking policy in Section 3.4.3. For the LP-Priority policy based on the Lagrangian optimal index, the reason why it does not work is less obvious. Some sample paths suggest that under this policy, the arms will concentrate on $\{3, 4, 5\}$ most of the time and thus cannot get a reward. Here is a possible explanation for why arms cannot easily escape from $\{3, 4, 5\}$: when some of the arms transition to state 6, by the priority $6 > 5 > 4$, those arms in state 6 are likely to be activated. Because states 6 do not prefer action 1, those arms will transition back to state 5 and thus get trapped.

In addition to the numerical results, we also note that our policy FTVA ($\bar{\pi}^*$) is provably asymptotically optimal in this RB problem even though its transition kernel is not unichain. This stems from the fact that the optimal single-armed policy $\bar{\pi}^*$ has a single recurrent class $\mathbb{S}$, and satisfies Synchronization Assumption. See Appendix 3.D for the discussion on the conditions under which Synchronization Assumption and Theorem 3.2 hold when unichain is not assumed.

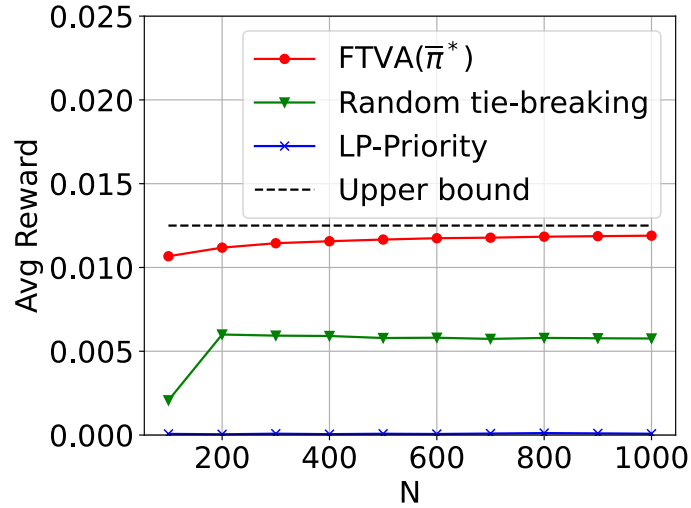
3.G.3 Varying the initial points of the Figure 3.2 and 3.3 examples

In Figure 3.2 and Figure 3.3, we have fixed one initial point for each simulation, for the ease of presentation. In this section, we rerun the simulations for these

⁵A nuance is that the optimal Lagrange multiplier for the budget constraint is not unique in this setting, so there can be different Lagrange relaxations. We focus on the simplest one.



(a) The experiment of Figure 3.2



(b) The experiment of Figure 3.3

Figure 3.7: We re-run the two experiments of Figure 3.2 and 3.3 with 20 sets of simulation with randomly generated initial points. For FTVA, we calculate the *minimum* average reward over the 20 sets; for the other policies, we calculate the *maximum* average reward over the 20 sets.

two settings with more initial points. For the simulations of each setting, we generate initial states in the following way.

- We first choose a probability distribution on the state space $\mathbb{S}$ from all the possible probability distributions on $\mathbb{S}$ uniformly at random. Let $\pi(s)$ denote the probability of state s under the chosen distribution. This distribution is chosen independently across the 20 sets of simulations.
- For each N -armed problem, we set the initial state such that $N \cdot \pi(s)$ arms are in state s for each s , with proper rounding.
- For each N -armed problem, we run all the policies with this initial state.

We illustrate the results of the 20 sets of simulations in the Figure 3.7a and 3.7b in the following way: For $\text{FTVA}(\bar{\pi}^*)$, for each N , we take the minimum average reward over the 20 sets of simulations and plot them in the figure; For other policies, for each N , we take the maximum average reward over the 20 sets of simulations. We can see that the min curve for $\text{FTVA}(\bar{\pi}^*)$ still approaches the optimal value as N increases, whereas the max curves for other policies are strictly separated from the optimal value.

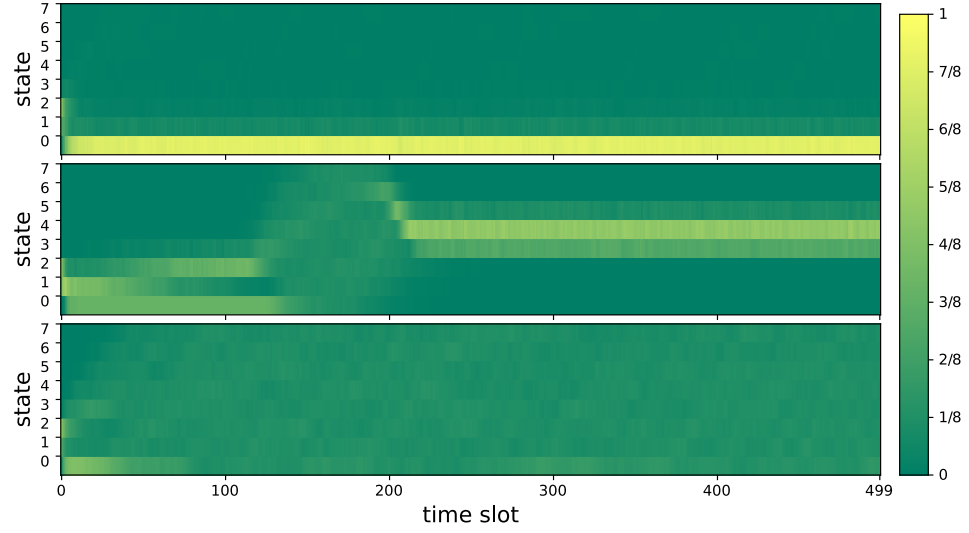
3.G.4 Visualization of the policies

In Figure 3.4 of Section 3.4.3, we illustrate what LP-Priority and $\text{FTVA}(\bar{\pi}^*)$ does differently by visualizing their sample paths. Here we include more such figures.

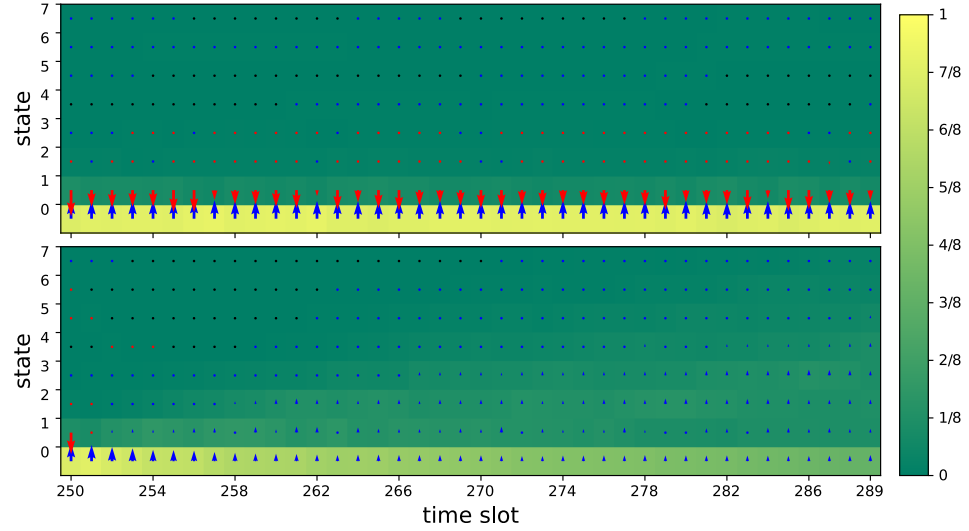
We still use the same example defined in Section 3.4.3. To recap, this example has 8 states, numbered as $\{0, 1, 2, 3, 4, 5, 6, 7\}$. The optimal single-armed policy always chooses the preferred action in each state, which is 1 for states 0, 1, 2, 3, and 0 for states 4, 5, 6, 7. Under the optimal single-armed policy, each arm moves from state s to $(s + 1) \bmod 8$ with probability 0.1 or stays at s with probability 0.9. The optimal distribution for the single-armed system is uniform over the 8 states.

When simulating the policies in the N -armed system, we can tell that a policy is good if the fraction of arms in each state is roughly uniform, and bad if the arms concentrate on a small number of states. We can also see why this happens by looking at whether the preferred action is chosen, and whether the arms are moving from s to $(s + 1) \bmod 8$.

The three heatmaps in Figure 3.8a illustrate how the fractions of arms in each state change over time under the three policies: random tie-breaking,



(a) Time evolution of the fractions of arms in each state under the three policies during the first 500 time slots. Top to bottom: Random Tie-Breaking, LP-Priority, and FTVA($\bar{\pi}^*$).



(b) Time evolution of the fraction of arms in each state under Random Tie-Breaking (upper), or after switching to FTVA($\bar{\pi}^*$) (lower) since time slot 250. The color and magnitude of the arrows represent the average movement of the arms in each state.

Figure 3.8: Visualization of the sample paths under different policies for the example in Section 3.4.3.

LP-Priority, and $\text{FTVA}(\bar{\pi}^*)$. The x-axis represents the time slot, which ranges from 0 to 499; the y-axis represents the states; the brighter color represents a larger fraction of arms in this state at this time, and the specific value represented by each color can be found in the color bar on the right.

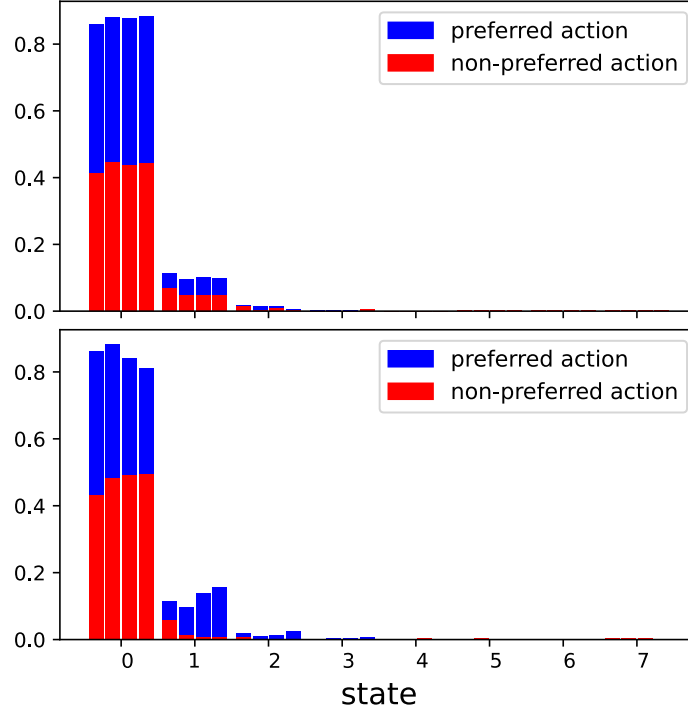
We can see that under the random tie-breaking, the arms concentrate around state 0; under LP-Priority, the arms concentrate around states $\{3, 4, 5\}$ after a burn-in period; under $\text{FTVA}(\bar{\pi}^*)$, the arms uniformly spread out over the 8 states.

In Figure 3.8b, we take a closer look at the sample path of the random tie-breaking policy from time 250 to 289 and contrast it with the sample path if the system uses $\text{FTVA}(\bar{\pi}^*)$ from time 250 onwards. We also add some arrows indicating the *drift* of the arms, i.e., the average direction that the arms in this state are moving into. The colors of the arrows represent the direction: the blue arrow implies that the arm takes the preferred action and moves in the correct direction, while the red arrow implies that the arm takes the non-preferred action and moves in the wrong direction. The magnitudes of the arrows represent how fast they are moving.

We can see that under random tie-breaking, although the arms have a strong tendency to move up from state 0 to state 1, they move back when they reach state 1 and thus get stuck at state 0. In contrast, when switching to $\text{FTVA}(\bar{\pi}^*)$, most arrows point in the right direction, which implies that the arms consistently apply the preferred actions. As a result, $\text{FTVA}(\bar{\pi}^*)$ helps the arms to escape from state 0 and converge to the uniform distribution over the state space.

Figure 3.9a illustrates the fraction of arms in state s taking action a in a few time slots after 250 in the same sample path of the random tie-breaking policy in Figure 3.8a. For each s and a . The upper bar plot is under the random tie-breaking policy, while the lower bar plot is after switching to $\text{FTVA}(\bar{\pi}^*)$. The x-axis represents the state; the y-axis represents the fraction; there are four bars in each state, corresponding to 4 time-steps. Each bar has two segments, where the length of the blue segment indicates the fraction of arms taking the preferred action in this state, and the length of the red segment indicates the fraction of arms taking the non-preferred action in this state; the lower segment of the bar corresponds to action 0, and the higher segment of the bar corresponds to action 1.

We can see that under $\text{FTVA}(\bar{\pi}^*)$, more arms choose the preferred actions for state 1 than under the random tie-breaking policy, which prevents the

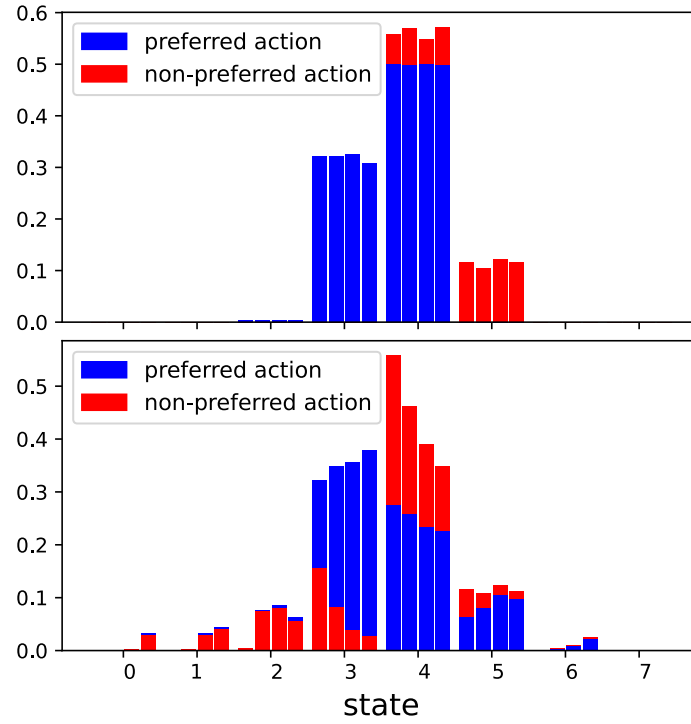


(a) The fraction of arms in each state taking each action in the 4 time steps following time 250, under Random Tie-Breaking (upper plot) or after switching to $\text{FTVA}(\bar{\pi}^*)$ (lower plot).

Figure 3.9: Actions taken by different policies for the example in Section 3.4.3. In each bar plot, the X-axis represents the states; the Y-axis represents the fractions. Each state has bars for the 4 time steps. The blue segment of each bar is the fraction of arms taking the preferred action; the red segment corresponds to the non-preferred action. The lower segment of each bar corresponds to action 0; the higher segment corresponds to action 1.

arms in state 1 from moving back to state 0 as we see in Figure 3.8b under the random tie-breaking policy.

We can see that under $\text{FTVA}(\bar{\pi}^*)$, more arms choose the preferred actions for state 1 than under the random tie-breaking policy, which prevents the arms in state 1 from moving back to state 0 as we see in Figure 3.8b under the random tie-breaking policy.



(b) The fraction of arms in each state taking each action in the 4 time steps following time 250, under LP-Priority (upper plot) or after switching to $\text{FTVA}(\bar{\pi}^*)$ (lower plot).

Figure 3.9: Actions taken by different policies (continued).

Figure 3.9b is similar to Figures 3.9a, except that the random tie-breaking policy is replaced by the LP-Priority policy. We can again see that under LP-Priority there is a force that causes the arms to concentrate on a state, whereas switching to $\text{FTVA}(\bar{\pi}^*)$ helps the system to escape from that state.

The common failure mode for random tie-breaking and LP-Priority in this setting is that the arms concentrate on a bad state s , which prevents the arms on state $(s + 1)$ from applying the preferred action due to the budget constraint. This causes a *livelock* where the arms move back and forth between state s and $(s + 1)$, and fail to gain a reward by moving past state 7. $\text{FTVA}(\bar{\pi}^*)$ solves this issue by letting more arms in state $(s + 1)$ follow the preferred action, which breaks the livelock and helps all arms converge to the optimal distribution.

3.H Generalization to heterogeneous arms

In this section, we show how FTVA and its analysis can be extended to the case with heterogenous arms. We focus on the discrete-time case for simplicity. The continuous-time case can be analyzed in a similar fashion.

3.H.1 Setting and algorithm

Suppose the arms are divided into K types, with $\beta_k N$ arms in each type $k \in \{1, 2, \dots, K\} \triangleq [K]$, and each type is associated with an MDP. The transition kernels and reward functions can be different across types. Suppose each arm is indexed by $i \in [N]$, and we denote the type of the i -th arm as $k(i)$. For any type k arm, we let $P_k(s, a, s')$ be its probability of transitioning from state s to s' upon taking action a , and let $r_k(s, a)$ be its reward for taking action a in state s .

Consider the linear program below, whose variables $y_k(s, a)$ represents the steady-state probability that a type k arm is in state s taking action a , for $k \in [K], s \in \mathbb{S}, a \in \mathbb{A}$.

$$\begin{aligned} & \underset{\{y_k(s, a)\}_{k \in [K], s \in \mathbb{S}, a \in \mathbb{A}}}{\text{maximize}} && \sum_k \sum_{s, a} \beta_k r_k(s, a) y_k(s, a) \end{aligned} \quad (\text{LP-Het})$$

$$\text{subject to} \quad \sum_k \sum_s \beta_k y_k(s, 1) = \alpha, \quad (3.41)$$

$$\sum_{s', a} y_k(s', a) P_k(s', a, s) = \sum_a y_k(s, a), \quad \forall k \in [K], s \in \mathbb{S}, \quad (3.42)$$

$$\sum_{s, a} y_k(s, a) = 1 \quad \forall k \in [K]; \quad (3.43)$$

$$y_k(s, a) \geq 0, \quad \forall k \in [K], s \in \mathbb{S}, a \in \mathbb{A}.$$

where the three constraints (3.41), (3.42), and (3.43) correspond to (3.5), (3.6), and (3.7) in (LP); when writing summations, we omit $\in [K]$, $\in \mathbb{S}$ and $\in \mathbb{A}$ for simplicity.

(LP-Het) can be viewed as a relaxation of the N -armed problem. To see this, take any N -armed policy π and set $y_k(s, a)$ to be the fraction of arms in state s taking action a among type k arms in steady state under π , i.e.,

$$y_k(s, a) = \frac{1}{\beta_k N} \mathbb{E} \left[\sum_{k(i)=k} \mathbb{1}_{\{S_i^\pi(\infty)=s, A_i^\pi(\infty)=a\}} \right].$$

Whenever π satisfies the budget constraint (3.2), $\{y_k(s, a)\}_{k \in [K], s \in \mathbb{S}, a \in \mathbb{A}}$ satisfies (3.41)–(3.43). Therefore, the optimal value of (LP-Het), which we denote as V_1^{rel} , is an upper bound of the optimal value of the N -armed problem, i.e., $V_1^{\text{rel}} \geq V_N^*$.

The optimal solution to (LP-Het), $\{y_k^*(s, a)\}_{k \in [K], s \in \mathbb{S}, a \in \mathbb{A}}$, induces a optimal single-armed policy $\bar{\pi}_k^*$ for each type $k \in [K]$:

$$\bar{\pi}_k^*(a|s) = \begin{cases} y_k^*(s, a)/(y_k^*(s, 0) + y_k^*(s, 1)), & \text{if } y_k^*(s, 0) + y_k^*(s, 1) > 0, \\ 1/2, & \text{if } y_k^*(s, 0) + y_k^*(s, 1) = 0, \end{cases} \quad (3.44)$$

for $s \in \mathbb{S}$ and $a \in \mathbb{A}$. Standard argument in [230] show that if a type k arm runs $\bar{\pi}_k^*$, it achieves the steady-state expected reward $\sum_{s,a} r_k(s, a)y_k^*(s, a)$, and requires $\sum_s y_k^*(s, 1)$ unit of budget in steady-state. If each type k arm could independently run the optimal single-armed policy of its type, since the fraction of type k arms in the N -armed system is β_k , in steady state, the expected reward per arm would be V_1^{rel} , and the budget requirement per arm would be $\sum_{k=1}^K \sum_s y_k^*(s, 1) = \alpha$, which is analogous to the homogeneous case where the optimal policy of the relaxed problem in (3.3) achieves V_1^{rel} reward and requires α expected budget in steady-state.

To convert from the single-armed policies to an N -armed policy, FTVA lets each arm of type k independently simulate a virtual single-armed process following $\bar{\pi}_k^*$. Lines 3-14 of Algorithm 1 stay the same. We denote the resulting policy as $\text{FTVA}(\{\bar{\pi}_k^*\}_{k \in [K]})$. The resulting policy can be proved to achieve $O(1/\sqrt{N})$ optimality gap, under the assumption that Synchronization Assumption is satisfied for each $\bar{\pi}_k^*$, as stated below:

Theorem 3.27. *In restless bandits with heterogeneous arms, let $\{\bar{\pi}_k^*\}_{k \in [K]}$ be the optimal single-armed policies induced by (LP-Het). Assume that for each $k \in [K]$, $\bar{\pi}_k^*$ satisfies Synchronization Assumption with the synchronization times $\{\tau_k^{\text{sync}}(s, a, \hat{s}, \hat{a})\}_{(s,a,\hat{s},\hat{a}) \in \mathbb{S} \times \mathbb{A} \times \mathbb{S} \times \mathbb{A}}$. For any $N \geq 1$, the optimality gap of $\text{FTVA}(\{\bar{\pi}_k^*\}_{k \in [K]})$ is upper bounded as*

$$V_N^* - V_N^{\text{FTVA}(\{\bar{\pi}_k^*\}_{k \in [K]})} \leq \frac{r_{\max} \bar{\tau}_{\max}^{\text{sync}}}{\sqrt{N}}, \quad (3.45)$$

where

$$r_{\max} \triangleq \max_{s \in \mathbb{S}, a \in \mathbb{A}} |r(s, a)|,$$

$$\bar{\tau}_{\max}^{\text{sync}} \triangleq \max_{k \in [K], (s,a,\hat{s},\hat{a}) \in \mathbb{S} \times \mathbb{A} \times \mathbb{S} \times \mathbb{A}} \mathbb{E} [\tau_k^{\text{sync}}(s, a, \hat{s}, \hat{a})].$$

3.H.2 Proof for Theorem 3.27

In this section, we prove Theorem 3.27.

Since $V_1^{\text{rel}} \geq V_N^*$, it suffices to show that

$$V_1^{\text{rel}} - V_N^{\text{FTVA}(\{\bar{\pi}_k^*\}_{k \in [K]})} \leq \frac{r_{\max} \bar{r}_{\max}^{\text{sync}}}{\sqrt{N}}. \quad (3.46)$$

We start with an inequality that has the same form as (3.13) in the homogeneous case:

$$\begin{aligned} & V_1^{\text{rel}} - V_N^{\text{FTVA}(\{\bar{\pi}_k^*\}_{k \in [K]})} \\ &= \frac{1}{N} \mathbb{E} \left[\sum_{i=1}^N r(\hat{S}_i(\infty), \hat{A}_i(\infty)) - \sum_{i=1}^N r(S_i(\infty), A_i(\infty)) \right] \\ &\leq \frac{2r_{\max}}{N} \mathbb{E} \left[\sum_{i=1}^N \mathbb{1} \left\{ (\hat{S}_i(\infty), \hat{A}_i(\infty)) \neq (S_i(\infty), A_i(\infty)) \right\} \right], \end{aligned} \quad (3.47)$$

where in the first equality, we used the fact that the virtual processes are independently running the single-armed policy of the corresponding types, so they achieve the average reward V_1^{rel} .

Defining the disagreement events, disagreement periods, and disagreement rates in the same way as Appendix 3.E and applying Little's law, we have

$$\mathbb{E} \left[\sum_{i=1}^N \mathbb{1} \left\{ (\hat{S}_i(\infty), \hat{A}_i(\infty)) \neq (S_i(\infty), A_i(\infty)) \right\} \right] = \mathbb{E} [d(\hat{\mathbf{S}}(\infty))] \cdot D_{\text{avg}}, \quad (3.48)$$

where $d(\hat{\mathbf{s}})$ is the instantaneous disagreement rate, and D_{avg} is the long-run average length of disagreement periods. With (3.47) and (3.48), it remains to bound $\mathbb{E} [d(\hat{\mathbf{S}}(\infty))]$ and D_{avg} .

By the definition of the policy, we have

$$\mathbb{E} [d(\hat{\mathbf{S}}(\infty))] = \mathbb{E} \left[\left| \sum_{i=1}^N \hat{A}_i(\infty) - \alpha N \right| \right].$$

Moreover, $\hat{A}_i(\infty)$'s are independent Bernoulli random variables, and $\mathbb{E} [\hat{A}_i(\infty)] = y_k^*(s, 1)$ if arm i is of type k . As a result,

$$\mathbb{E} \left[\sum_{i=1}^N \hat{A}_i(\infty) \right] = \sum_k \beta_k N \sum_{s,a} y_k^*(s, 1) = \alpha N.$$

Using the same Cauchy-Schwartz argument as in the proof of the homogeneous case in Appendix 3.E, it is not hard to show that $\mathbb{E} \left[d(\hat{\mathbf{S}}(\infty)) \right] \leq \sqrt{N}/2$.

As for D_{avg} , it is not hard to see from the definition that the length of each disagreement period of a type k arm is stochastically dominated by $\tau_k^{\text{sync}}(s, a, \hat{s}, \hat{a})$, where $(s, a, \hat{s}, \hat{a})$ are the initial states and actions of that disagreement period. Therefore, the average length of disagreement period is bounded by

$$D_{\text{avg}} \leq \max_{k \in [K], (s, a, \hat{s}, \hat{a}) \in \mathbb{S} \times \mathbb{A} \times \mathbb{S} \times \mathbb{A}} \mathbb{E} [\tau_k^{\text{sync}}(s, a, \hat{s}, \hat{a})] = \bar{\tau}_{\text{max}}^{\text{sync}}.$$

Combining the above calculations, we get (3.46), which finishes the proof.

Chapter 4

Unichain and Aperiodicity are Sufficient for Asymptotic Optimality of Average-Reward Restless Bandits

4.1 Introduction

In the previous chapter, we have taken a first step towards relaxing the long-standing GAP assumption. Recall that we have proposed a policy named Follow-The-Virtual-Advice (FTVA) for the discrete-time RB problem and a variant, FTVA-CT, for the continuous-time setting. We have “removed” GAP in both settings, but to different extent: For the continuous-time setting, FTVA-CT has completely removed the GAP assumption, achieving asymptotic optimality under the unichain property; for the discrete-time setting, however, FTVA still requires an alternative condition named Synchronization Assumption (SA). While SA is more intuitive and easier-to-verify than GAP, the reliance on SA is still unsatisfactory. In particular, there exist RB instances where SA is not satisfied and FTVA performs suboptimally. In Section 4.8.3 and Appendix 4.A, we provide two counterexamples to SA, and discuss ways to construct more such examples.

The need for additional assumptions like GAP and SA reveals a fundamental gap in our understanding of the restless bandit problem. As such, the literature on RBs leaves open the following question: *Is it possible to*

efficiently find a policy that achieves asymptotic optimality in infinite-horizon, average-reward RBs under only unichain and aperiodicity type of conditions, without imposing any additional conditions?

4.1.1 Our contributions

Answer to the question. In this chapter, we focus on the discrete-time RB problem and give a definitive, affirmative answer to this long-standing question. We propose a novel class of policies named *focus-set policies*, and construct two concrete instances of focus-set policies that are asymptotically optimal with an $O(1/\sqrt{N})$ optimality gap under a weaker-than-standard aperiodic-unichain assumption (Assumption 4.1).

Policy design. Our proposed policies depart from the prevalent *priority-based* design of existing policies. A priority-based policy specifies a fixed priority order over all the *states* of a single arm. At each time step, the policy pulls arms from states of higher priorities to those of lower priorities, until the budget constraint is met. In contrast, each of our proposed policies selects a subset of arms based on the *empirical distribution* of their states and lets the selected arms take their *ideal actions* as much as possible. These ideal actions are computed using the solution of a single-armed, budget-relaxed problem. The subset selection is constructed in a way such that most arms in the subset can take their ideal actions and the subset expands over time. Note that the FTVA policy in Chapter 3 [149] also leverages the same single-armed problem to guide policy construction; however, it requires simulating additional virtual states, whereas our policies do not.

Proof techniques. We establish a meta-theorem that provides sufficient conditions for the asymptotic optimality of a focus-set policy. The proof of the meta-theorem highlights a class of bivariate Lyapunov functions we term *subset Lyapunov functions*, along with a global Lyapunov function constructed dynamically from one of the subset Lyapunov functions. Using these Lyapunov functions, we show that, under the stipulated sufficient conditions, the state-action distribution of arms in the selected subset converges to the optimal distribution, and the subset eventually expands to cover most arms. This meta-theorem allows us to prove the asymptotic optimality of the

two proposed instances of the focus-set policies by verifying the stipulated sufficient conditions.

4.1.2 Chapter organization

The remainder of the chapter is organized as follows. In Section 4.2, we give a more detailed review of the literature. In Section 4.3, we set up the problem of average-reward restless bandits and introduce the single-armed problem. In Section 4.4, we present our main results, where we propose the focus-set policies, present two concrete instances of focus-set policies, namely, the set-expansion policy and the ID policy, and establish their $O(1/\sqrt{N})$ optimality. In Section 4.5, we present a meta-theorem that provides sufficient conditions for focus-set policies to achieve $O(1/\sqrt{N})$ optimality gaps. In Section 4.6 and Section 4.7, we use the meta-theorem to prove the asymptotic optimality of the set-expansion policy and the ID policy. In Section 4.8, we conduct experiments comparing the performances of our policies with existing policies, and investigate the GAP and SA assumptions. We conclude the chapter in Section 4.9.

4.2 Additional related work

Backgrounds on unichain and aperiodicity. Most prior work for restless bandits have been reviewed in Sections 3.1 and 3.2. Here we make two additional comments on the unichain and aperiodicity condition. First, the unichain conditions assumed in these papers are different. In particular, the unichain condition assumed in [266, 274] requires the N -armed system to be unichain under every policy, which is stronger than the single-armed single-policy unichain condition considered in our chapter (Assumption 4.1). See Appendix 4.B.1 for a detailed discussion on different unichain conditions and their roles in these papers. Second, although none of the papers reviewed above explicitly assume any form of aperiodicity, this is either because they focus on the continuous-time setting [266, 274], where, loosely speaking, all policies are aperiodic, or because there is another assumption that plays a similar role as aperiodicity (i.e., Synchronization Assumption (SA) in Chapter 3 [149]). Our chapter considers the discrete-time setting without the SA assumption, where aperiodicity becomes crucial, as demonstrated by an example in Appendix 4.B.2.

Subsequent papers that further relax our conditions. In addition to the prior work reviewed above, two recent papers [121, 294] appeared a few months after the arXiv version of the work reviewed in this chapter. These papers propose new policies that are asymptotically optimal under weaker conditions than ours. In particular, [294] considers the discrete-time average-reward RB problem and [121] considers the so-called weakly-coupled MDP problem, which is a multi-action, multi-constraint generalization of the RB problem. The assumptions in these two papers are implied by the single-armed MDP being weakly communicating and aperiodic, whereas ours are not. However, [121, 294] only provide asymptotic results but do not characterize the order of the optimality gaps.

4.3 Problem setup

In this section, we restate the average-reward restless bandits problem and its single-armed relaxation, introducing the notation and assumptions for this chapter.

4.3.1 Restless bandit problem

We consider the discrete-time, infinite-horizon restless bandit problem with the average-reward criterion. The RB problem consists of N homogeneous arms and is henceforth referred to as the N -armed problem. Each arm is associated with an MDP called the single-armed MDP, which is defined by the tuple $(\mathbb{S}, \mathbb{A}, P, r)$. Here $\mathbb{S}$ is the state space, which is a finite set; $\mathbb{A} = \{0, 1\}$ is the action space, where the action 1 is interpreted as activating or pulling the arm; $P : \mathbb{S} \times \mathbb{A} \times \mathbb{S} \rightarrow [0, 1]$ is the transition kernel, where $P(s, a, s')$ is the probability of transitioning to state s' in the next time step conditioned on taking action a at state s in the current step; $r : \mathbb{S} \times \mathbb{A} \rightarrow \mathbb{R}$ is the reward function, where $r(s, a)$ is the expected reward for taking action a in state s . Let $r_{\max} = \max_{s \in \mathbb{S}, a \in \mathbb{A}} |r(s, a)|$. The RB problem has a *budget constraint*, which requires that exactly αN arms must be pulled at every time step for some given constant $\alpha \in (0, 1)$. Here αN is assumed to be an integer for simplicity. We note that although the equality budget constraint is standard in the literature and is the primary focus of this chapter, our algorithms and results also extend to a natural alternative formulation with an inequality constraint, which allows *at most* αN arms to be pulled at every time step.

We will describe the necessary modifications at appropriate points in the chapter. Finally, note that we focus on the setting in which all the model parameters, $\mathbb{S}, \mathbb{A}, P, r, \alpha$, are known.

We index the arms in an N -armed bandit by $[N]$, where $[n] \triangleq \{1, 2, \dots, n\}$ for integer $n \geq 1$ and $[0] \triangleq \emptyset$. We refer to the index i of Arm i as its *ID*, to avoid confusion with the Whittle index or other index notion.

A policy π for the N -armed problem chooses in each time step the action for each of the N arms. We allow the policy to be randomized and choose actions based on the whole history.

Under a policy π , we use the *state vector* $\mathbf{S}_t^\pi \triangleq (S_t^\pi(i))_{i \in [N]} \in \mathbb{S}^N$ to represent the states of all arms, where $S_t^\pi(i) \in \mathbb{S}$ denotes the state of the i -th arm at time t . Similarly, the *action vector* is defined as $\mathbf{A}_t^\pi \triangleq (A_t^\pi(i))_{i \in [N]} \in \mathbb{A}^N$, where $A_t^\pi(i) \in \mathbb{A}$ denotes the action applied to the i -th arm at time t .

Let the *limsup average reward* and *liminf average reward* be

$$R^+(\pi, \mathbf{S}_0) \triangleq \limsup_{T \rightarrow \infty} (1/(NT)) \sum_{t=0}^{T-1} \sum_{i \in [N]} \mathbb{E}[r(S_t^\pi(i), A_t^\pi(i))]$$

$$R^-(\pi, \mathbf{S}_0) \triangleq \liminf_{T \rightarrow \infty} (1/(NT)) \sum_{t=0}^{T-1} \sum_{i \in [N]} \mathbb{E}[r(S_t^\pi(i), A_t^\pi(i))].$$

When the limsup and liminf average rewards coincide, the infinite-horizon average reward (also known as the long-run average reward) exists and is given by

$$R(\pi, \mathbf{S}_0) \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i \in [N]} \mathbb{E}[r(S_t^\pi(i), A_t^\pi(i))].$$

Our goal is to solve the following optimization problem.

$$\begin{aligned} & \underset{\text{policy } \pi}{\text{maximize}} && R^-(\pi, \mathbf{S}_0) \end{aligned} \tag{RB}$$

$$\begin{aligned} & \text{subject to} && \sum_{i \in [N]} A_t^\pi(i) = \alpha N, \quad \forall t = 0, 1, 2, \dots \end{aligned} \tag{4.1a}$$

Let $R^*(N, \mathbf{S}_0)$ be the optimal value of the problem, referred to as the *optimal reward*. Note that $R^*(N, \mathbf{S}_0) = \sup_{\pi'} R^-(\pi', \mathbf{S}_0) = \sup_{\pi'} R^+(\pi', \mathbf{S}_0)$ because (RB) is an MDP with finite state and action spaces [230, Theorem 9.1.6]. For any policy π , we define its optimality gap as $R^*(N, \mathbf{S}_0) - R^-(\pi, \mathbf{S}_0)$; we say

the policy is *asymptotically optimal* if its optimality gap vanishes as $N \rightarrow \infty$, i.e., $R^*(N, \mathbf{S}_0) - R^-(\pi, \mathbf{S}_0) = o(1)$. This notion of asymptotic optimality is consistent with the literature; see, e.g., [266, Definition 4.11].

In later parts of the chapter, we will focus on policies under which the long-run average reward $R(\pi, \mathbf{S}_0)$ exists. These policies include any stationary Markovian policies, under which $\mathbf{S}_t$ is a finite-state Markov chain [230, Proposition 8.1.1]. More generally, with a similar argument, it is easy to show that $R(\pi, \mathbf{S}_0)$ is well-defined if π makes decisions based on augmented system states with a finite state space. Importantly, restricting to such policies is sufficient, because there always exists a stationary Markovian policy whose long-run average reward achieves the optimal reward, by standard results for the MDPs with finite state and action spaces [230, Theorem 9.1.8]. For simplicity, we will refer to $R(\pi, \mathbf{S}_0)$ as the objective function of (RB) and write the optimality gap as $R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0)$.

4.3.2 Scaled state-count vector

We introduce an alternative way, used extensively in the chapter, for representing the information contained in the state vector $\mathbf{S}_t^\pi$. For each subset $D \subseteq [N]$, we define the *scaled state-count vector on D* as $X_t^\pi(D) = (X_t^\pi(D, s))_{s \in \mathbb{S}}$, where

$$X_t^\pi(D, s) = \frac{1}{N} \sum_{i \in D} \mathbf{1}\{S_t^\pi(i) = s\}.$$

Note that each entry of the vector $X_t^\pi(D)$ is the number of arms in D in a certain state, scaled by $1/N$. When $D = [N]$ is the set of all arms, we simply call $X_t^\pi([N])$ the *scaled state-count vector*.

Sometimes we view $X_t^\pi(D)$ as a vector-valued function of $D \subseteq [N]$. We refer to this function X_t as the *system state* at time t . The system state X_t^π contains the same information as the state vector $\mathbf{S}_t^\pi$ does; in particular, from X_t^π one can deduce the state of each arm.

4.3.3 LP relaxation

In this section, we discuss a linear programming (LP) relaxation of the N -armed problem (RB) which is crucial for the design and analysis of RB policies. This LP is defined as follows.

$$\begin{aligned} & \underset{\{y(s,a)\}_{s \in \mathbb{S}, a \in \mathbb{A}}}{\text{maximize}} && \sum_{s \in \mathbb{S}, a \in \mathbb{A}} r(s, a) y(s, a) \end{aligned} \tag{LP}$$

$$\text{subject to} \quad \sum_{s \in \mathbb{S}} y(s, 1) = \alpha, \quad (4.2a)$$

$$\sum_{s' \in \mathbb{S}, a \in \mathbb{A}} y(s', a) P(s', a, s) = \sum_{a \in \mathbb{A}} y(s, a), \quad \forall s \in \mathbb{S}, \quad (4.2b)$$

$$\sum_{s' \in \mathbb{S}, a' \in \mathbb{A}} y(s', a') = 1, \quad y(s, a) \geq 0, \quad \forall s \in \mathbb{S}, a \in \mathbb{A}. \quad (4.2c)$$

To see why (LP) is a relaxation of (RB), for any stationary Markovian policy π , consider

$$y^\pi(s, a) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{1}{N} \sum_{i \in [N]} \mathbb{1}\{S_t^\pi(i) = s, A_t^\pi(i) = a\} \right] \quad \forall s \in \mathbb{S}, a \in \mathbb{A}.$$

It is not hard to see that $R(\pi, \mathbf{S}_0) = \sum_{s \in \mathbb{S}, a \in \mathbb{A}} r(s, a) y^\pi(s, a)$, and that $(y^\pi(s, a))_{s \in \mathbb{S}, a \in \mathbb{A}}$ satisfies the constraints (4.2a)–(4.2c). Therefore, letting R^{rel} be the optimal value of (LP), it can be shown that $R^{\text{rel}} \geq R^*(N, \mathbf{S}_0)$ (See Appendix 4.C for the detailed proof). This relation allows us to bound the optimality gap of any policy π using the inequality $R^*(N, \mathbf{S}_0) - R^-(\pi, \mathbf{S}_0) \leq R^{\text{rel}} - R^-(\pi, \mathbf{S}_0)$. Bounding the optimality gap via this LP relaxation is the standard approach in the literature [107, 108, 149, 266, 274].

4.3.4 Optimal single-armed policy

To understand how to achieve the average-reward upper bound R^{rel} given by the LP relaxation (LP), it is helpful to view (LP) as solving for a certain stationary state-action probability, $y(s, a)$, in the single-armed MDP, $(\mathbb{S}, \mathbb{A}, P, r)$. Specifically, the objective of (LP) equals the expected reward under the stationary probability. The constraint in (4.2a) can be interpreted as a budget constraint, which requires the arm to be activated with α probability in the steady state. The constraint (4.2b) is the stationary equation. The constraint (4.2c) ensures that $(y(s, a))_{s \in \mathbb{S}, a \in \mathbb{A}}$ is a valid probability distribution.

From each stationary state-action probability $(y(s, a))_{s \in \mathbb{S}, a \in \mathbb{A}}$, one can construct a policy for the single-armed MDP, which we call a *single-armed policy*, that achieves this state-action probability. In particular, let $\{y^*(s, a)\}_{s \in \mathbb{S}, a \in \mathbb{A}}$ be an optimal solution to (LP). We consider the following single-armed policy $\bar{\pi}^*$: for $s \in \mathbb{S}$, $a \in \mathbb{A}$,

$$\bar{\pi}^*(a|s) = \begin{cases} y^*(s, a) / (y^*(s, 0) + y^*(s, 1)), & \text{if } y^*(s, 0) + y^*(s, 1) > 0, \\ 1/2, & \text{if } y^*(s, 0) + y^*(s, 1) = 0. \end{cases} \quad (4.3)$$

We call $\bar{\pi}^*$ the *optimal single-armed policy*. Let $P_{\bar{\pi}^*}$ be the transition matrix induced by $\bar{\pi}^*$ in the single-armed MDP. We make the following assumption throughout the chapter:

Assumption 4.1 (Unichain and aperiodicity). There exists an optimal solution $\{y^*(s, a)\}_{s \in \mathbb{S}, a \in \mathbb{A}}$ to (LP), such that the optimal single-armed policy $\bar{\pi}^*$ defined in (4.3) induces an aperiodic unichain (i.e., a Markov chain with a single recurrent class and a possibly empty set of transient states) with state space $\mathbb{S}$ and transition matrix $P_{\bar{\pi}^*}$.

With Assumption 4.1, the Markov chain induced by $\bar{\pi}^*$ converges to a unique stationary distribution, which we denote as $\mu^* = (\mu^*(s))_{s \in \mathbb{S}}$. From the definition of $\bar{\pi}^*$ in (4.3), it is easy to verify that $\mu^*(s) = y^*(s, 0) + y^*(s, 1)$; thus the steady-state state-action probability under $\bar{\pi}^*$ is $(y^*(s, a))_{s \in \mathbb{S}, a \in \mathbb{A}}$. Consequently, the long-run average reward of $\bar{\pi}^*$ equals the optimal value of (LP), R^{rel} ; the long-run average budget usage of $\bar{\pi}^*$ equals α .

In Appendix 4.B, we discuss the generality of Assumption 4.1. In particular, we compare Assumption 4.1 with the assumptions in the literature; we also give an example to show that $R^{\text{rel}} - R^*(N, \mathbf{S}_0)$ can be non-diminishing as $N \rightarrow \infty$ when the single-armed MDP is periodic.

4.3.5 Additional notation

For a subset $D \subseteq [N]$, we let $m(D) = |D|/N$ denote the fraction of arms contained in D . We introduce a convenient shorthand $[0, 1]_N = \{0, 1/N, 2/N, \dots, 1\}$. Then $m(D) \in [0, 1]_N$ for any $D \subseteq [N]$. Let $\Delta(\mathbb{S})$ denote the set of probability distributions on the state space $\mathbb{S}$. We treat each distribution $v \in \Delta(\mathbb{S})$ as a row vector. Recall that π denotes a policy for the N -armed problem. In later sections, when the context is clear, we drop the superscript π from the vectors $\mathbf{S}_t^\pi$, $\mathbf{A}_t^\pi$, and X_t^π . We use $a^+ \triangleq \max\{a, 0\}$ to denote the positive part of $a \in \mathbb{R}$.

4.4 Main results: policies and optimality guarantees

In this section, we propose policies for the average-reward RB problems and bound their optimality gaps. In Section 4.4.1, we present an algorithmic idea based on the convergence of state distribution to μ^* under the optimal

single-armed policy $\bar{\pi}^*$, and propose a novel class of policies named *focus-set policies*. Then in Section 4.4.2 and Section 4.4.3, we present two instances of focus-set policies, named the *set-expansion policy* and the *ID policy*, and state their optimality gap bounds. In Section 4.4.4, we discuss the relationships between our policies and the policies in the literature to explain why they rely on different assumptions. Finally, in Section 4.4.5, we discuss how our results can be adapted for the alternative setting with the inequality budget constraint.

4.4.1 Algorithmic idea and focus-set policies

Consider the single-armed system and the optimal single-armed policy $\bar{\pi}^*$. Because $P_{\bar{\pi}^*}$ is an aperiodic unichain by Assumption 4.1, it follows that starting from any initial distribution in $\Delta(\mathbb{S})$, the state distribution of the Markov chain $P_{\bar{\pi}^*}$ converges to the steady-state distribution μ^* .

We observe the following simple fact based on the single-armed convergence under $\bar{\pi}^*$: an RB system would achieve the reward upper bound if all arms could follow the optimal single-armed policy $\bar{\pi}^*$. However, exactly achieving this goal is not possible due to the budget constraint. A natural way is to approach it gradually: let a subset of arms persistently follow $\bar{\pi}^*$ and wait for them to approach μ^* ; at this point, more arms can be included into the subset, as we explain later.

We capture this idea of letting a subset of arms follow $\bar{\pi}^*$ and then gradually expanding the subset in a class of policies named *focus-set policies*, the template of which is given in Algorithm 3. In particular, a focus-set policy first samples an *ideal action* $\hat{A}_t(i)$ using $\bar{\pi}^*$ for each arm $i \in [N]$ based on its state $S_t(i)$ at time t (Line 2). The policy then selects a subset of arms D_t , referred to as the *focus-set*, and gives them precedence to set the actual action equal to the ideal action, i.e., $A_t(i) = \hat{A}_t(i)$ (Line 4).

The performance benefit of a focus-set policy will not be realized until one specifies a proper rule to update the focus set D_t . In subsequent subsections, we propose two instances of focus-set policies with updating rules that lead to asymptotically optimal performance. In Section 4.5, we provide more general sufficient conditions for a focus-set policy to achieve asymptotic optimality.

Algorithm 3 Focus-set policy

Input: number of arms N , budget αN , the optimal single-armed policy $\bar{\pi}^*$, initial system state X_0 , initial state vector $\mathbf{S}_0$, initial focus set D_{-1}

- 1: **for** $t = 0, 1, 2, \dots$ **do**
 $\triangleright$ *Set update*
- 2: Choose a *focus set* $D_t \subseteq [N]$ based on X_t and D_{t-1}
 $\triangleright$ *Action sampling*
- 3: Independently sample $\hat{A}_t(i) \sim \bar{\pi}^*(\cdot | S_t(i))$ for $i \in [N]$
 $\triangleright$ *Action rectification*
- 4: Pick $A_t(i)$ for $i \in D_t$ based on X_t and D_t to achieve

$$\begin{aligned} & \underset{\{A_t(i) : i \in D_t\}}{\text{maximize}} && \left| \left\{ i \in D_t : A_t(i) = \hat{A}_t(i) \right\} \right| \\ & \text{subject to} && \alpha N - (N - |D_t|) \leq \sum_{i \in D_t} A_t(i) \leq \alpha N \end{aligned} \quad (4.4)$$

- 5: Pick $A_t(i)$ for $i \in D_t^c$ based on X_t and D_t such that

$$\sum_{i \in [N]} A_t(i) = \alpha N \quad (4.5)$$

- 6: Apply $A_t(i)$ and observe $S_{t+1}(i)$ for each arm $i \in [N]$

4.4.2 Set-expansion policy

In this section, we introduce an instance of the focus-set policies called the *set-expansion policy*. The set-expansion policy updates D_t based on a quantity referred to as the *slack*, which is defined for a subset $D \subseteq [N]$ based on the system state x as follows:

$$\delta(x, D) = \beta(1 - m(D)) - \frac{1}{2} \|x(D) - m(D)\mu^*\|_1, \quad (4.6)$$

where $\beta = \min(\alpha, 1 - \alpha)$ and recall that $m(D) = |D|/N$. The policy chooses D_t with the maximal cardinality such that $\delta(X_t, D_t) \geq 0$, under the constraint that it either expands upon D_{t-1} (i.e., $D_t \supseteq D_{t-1}$) if $\delta(X_t, D_{t-1}) > 0$ or shrinks (i.e., $D_t \subseteq D_{t-1}$) otherwise. See Algorithm 4 for the full definition of the set-expansion policy.

Algorithm 4 Set-expansion policy

Input: number of arms N , budget αN , the optimal single-armed policy $\bar{\pi}^*$, initial system state X_0 , initial state vector $\mathbf{S}_0$, initial focus set $D_{-1} = \emptyset$

- 1: **for** $t = 0, 1, 2, \dots$ **do**
 - ▷ *Set update*
 - 2: **if** $\delta(X_t, D_{t-1}) > 0$ **then**
 - 3: Let D_t be any set with the largest $m(D_t)$ such that $D_t \supseteq D_{t-1}$ and $\delta(X_t, D_t) \geq 0$
 - 4: **else**
 - 5: Let D_t be any set with the largest $m(D_t)$ such that $D_t \subseteq D_{t-1}$ and $\delta(X_t, D_t) \geq 0$

▷ *Lines below implement Lines 2–6 of Algorithm 3, with random tie-breaking for Lines 4–5*

- ▷ *Action sampling*
 - 6: Independently sample $\hat{A}_t(i) \sim \bar{\pi}^*(\cdot | S_t(i))$ for $i \in [N]$
 - ▷ *Action rectification*
 - 7: **if** $\sum_{i \in D_t} \hat{A}_t(i) \geq \alpha N$ **then**
 - 8: Select αN arms in D_t with $\hat{A}_t(i) = 1$ uniformly at random, and set $A_t(i) = 1$
 - 9: For the rest of $i \in [N]$, set $A_t(i) = 0$
 - 10: **else if** $\sum_{i \in D_t} \hat{A}_t(i) \leq \alpha N - (N - |D_t|)$ **then**
 - 11: Select $(1 - \alpha)N$ arms in D_t with $\hat{A}_t(i) = 0$ uniformly at random, and set $A_t(i) = 0$
 - 12: For the rest of $i \in [N]$, set $A_t(i) = 1$
 - 13: **else**
 - 14: Set $A_t(i) = \hat{A}_t(i)$ for $i \in D_t$
 - 15: Set $A_t(i) = \hat{A}_t(i)$ for as many $i \notin D_t$ as possible; break ties uniformly at random
 - 16: Apply $A_t(i)$ and observe $S_{t+1}(i)$ for each arm $i \in [N]$
-

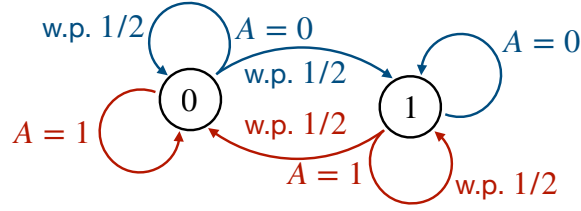


Figure 4.1: The single-armed MDP has two states $\{0, 1\}$; one unit of reward is generated if and only if the state changes, and $\alpha = 0.5$. The optimal policy $\bar{\pi}^*$ activates the arm if and only if it is in state 1, giving $\mu^* = (0.5, 0.5)$.

Here we briefly explain the design of the set-expansion policy and why it works. The slack $\delta(x, D)$ is carefully constructed such that $\delta(X_t, D_t) \geq 0$ ensures that most arms in D_t can follow $\bar{\pi}^*$. Moreover, $P_{\bar{\pi}^*}$ is non-expansive under the L_1 norm, so $\delta(X_t, D_t) \geq 0$ often implies that $\delta(X_{t+1}, D_t) \geq 0$, preventing D_{t+1} from shrinking significantly. Consequently, each arm in the focus set is likely to remain in the focus set for a long time, where they persistently follow $\bar{\pi}^*$ and converge to the optimal stationary distribution μ^* . As soon as the scaled state-count vector on D_t , $X_t(D_t)$, is sufficiently close to $m(D_t)\mu^*$, the focus set expands. Therefore, in the long run, the arms in the focus set converges to μ^* , allowing the focus set to cover most of the arms.

Next, we use an example to provide some more concrete intuition. For illustration purposes, let us temporarily suppose that the L_1 norm between μ^* and any distribution on $\mathbb{S}$ strictly decreases after right-multiplying $P_{\bar{\pi}^*}$.

Consider the RB problem defined by the two-state single-armed MDP in Figure 4.1. For any system state x and any subset of arms D , the scaled state-count vector on D , $x(D) = (X(D, 0), X(D, 1))$, can be represented by a point in the triangle $\{(a, b) : a \geq 0, b \geq 0, a + b \leq 1\}$, depicted in Figure 4.2a. The shaded region where all arms in D can follow $\bar{\pi}^*$ is $[0, 0.5] \times [0, 0.5]$. One can verify that the constraint $\delta(X_t, D_t) \geq 0$ exactly keeps $X_t(D_t)$ in the shaded region. In fact, $\delta(X_t, D_t)$ is proportional to the L_1 distance between $X_t(D_t)$ and the boundary of the shaded region, and is non-negative if $X_t(D_t)$ stays within this region. Consequently, under the set-expansion policy the arms in D_t mostly follow $\bar{\pi}^*$, so $X_{t+1}(D_t)$ moves closer to $m(D_t)\mu^*$ on the central dotted line; the subsequent expansion to D_{t+1} then pushes the point up-right while keeping it in the shaded region. Repeating these two steps, the iterates $X_1(D_1), X_2(D_1), X_2(D_2), X_3(D_2), X_3(D_3) \dots$ converge in a zigzag to a small neighborhood of μ^* , as illustrated by the solid arrows in Figure 4.2a.

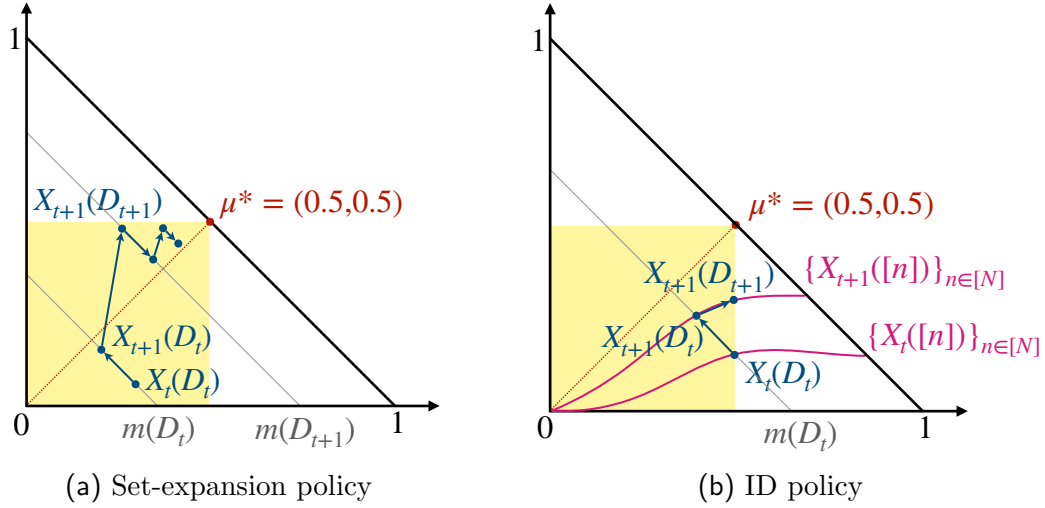


Figure 4.2: Intuitions of the set-expansion policy and the ID policy through the example in Figure 4.1. **(a)** A typical trajectory under the set-expansion policy: each system transition moves $X_t(D_t)$ toward the central line $\{m\mu^*\}$ (solid arrows), and each expansion of the focus set moves it up-right within the shaded region. **(b)** Under the ID policy, the system state is the solid curve $\{X_t([n])\}_{n \in [N]}$; its segment inside the shaded region follows $\bar{\pi}^*$, and D_t is defined so that $X_t(D_t)$ is where the curve meets the region's boundary.

Based on the above intuition, we can formally prove that the set-expansion policy is asymptotically optimal, as stated in Theorem 4.2 below. The proof of Theorem 4.2 is provided in Section 4.6.

Theorem 4.2 (Optimality gap of set-expansion policy). *Consider an N -armed restless bandit problem with the single-armed MDP $(\mathbb{S}, \mathbb{A}, P, r)$ and budget αN for $0 < \alpha < 1$. Assume that the optimal single-armed policy induces an aperiodic unichain (Assumption 4.1). Let π be the set-expansion policy (Algorithm 4). Then for all N and initial state vector $\mathbf{S}_0$, the optimality gap of π is bounded as*

$$R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0) \leq \frac{C_{\text{SE}}}{\sqrt{N}}, \quad (4.7)$$

where C_{SE} is a constant depending on $r_{\max}$, $|\mathbb{S}|$, $\beta \triangleq \min\{\alpha, 1 - \alpha\}$, and $P_{\bar{\pi}^*}$; the explicit expression of C_{SE} is given in the proof.

Theorem 4.2 shows that under Assumption 4.1, the set-expansion policy is asymptotically optimal with an $O(1/\sqrt{N})$ optimality gap. The bound (4.7)

Algorithm 5 ID policy

Input: number of arms N , budget αN , the optimal single-armed policy $\bar{\pi}^*$, initial system state X_0 , initial state vector $\mathbf{S}_0$

- 1: **for** $t = 0, 1, 2, \dots$ **do**
 $\triangleright$ *Action sampling*
- 2: Independently sample $\hat{A}_t(i) \sim \bar{\pi}^*(\cdot | S_t(i))$ for $i \in [N]$
 $\triangleright$ *Action rectification*
- 3: **if** $\sum_{i \in [N]} \hat{A}_t(i) \geq \alpha N$ **then**
- 4: $N_t^{\bar{\pi}^*} \leftarrow \max \{n \leq N : \sum_{i=1}^n \hat{A}_t(i) \leq \alpha N\}$
- 5: $A_t(i) \leftarrow \hat{A}_t(i)$ for $i \leq N_t^{\bar{\pi}^*}$, $A_t(i) \leftarrow 0$ for $i > N_t^{\bar{\pi}^*}$
- 6: **else**
- 7: $N_t^{\bar{\pi}^*} \leftarrow \max \{n \leq N : \sum_{i=1}^n (1 - \hat{A}_t(i)) \leq (1 - \alpha)N\}$
- 8: $A_t(i) \leftarrow \hat{A}_t(i)$ for $i \leq N_t^{\bar{\pi}^*}$, $A_t(i) \leftarrow 1$ for $i > N_t^{\bar{\pi}^*}$
- 9: Apply $A_t(i)$ and observe $S_{t+1}(i)$ for each arm $i \in [N]$

holds for any finite N , and only depends on intuitive quantities, including the problem primitives and a quantity reflecting the mixing time of the transition matrix $P_{\bar{\pi}^*}$.

Finally, note that Lines 7–15 of Algorithm 4 is just one way to implement the action rectification step of the focus-set policy by breaking ties uniformly at random. The analysis of the set-expansion policy still goes through as long as the constraints in Lines 4–5 of Algorithm 3 are satisfied. In Section 4.8, we will also consider an alternative way for action rectification for the set-expansion policy, where we let arms outside the focus-sets follow the LP index policy [108], which sometimes leads to a better empirical performance.

4.4.3 ID policy

Next, we introduce another instance of the focus-set policies, named the ID policy, whose pseudocode is given in Algorithm 5. The ID policy first samples an ideal action $\hat{A}_t(i)$ for each arm $i \in [N]$ using $\bar{\pi}^*$. It then proceeds sequentially through the arms by their IDs ($i = 1, 2, \dots, N$), assigning $A_t(i) = \hat{A}_t(i)$ for as many arms as the budget constraint allows. All subsequent arms with higher IDs are then forced to take the same action (either all 0 or all 1).

Although the focus set of the ID policy could in principle be any subset of the form $[n]$, for analysis purposes, we will construct a suitable one to use the same framework as the set-expansion policy. The formal definition of this focus set is deferred to Section 4.7.2. We will show that this constructed focus set has two properties: (1) most arms within it follow $\bar{\pi}^*$, and (2) the set expands almost monotonically over time to eventually contain most of the arms in the system. In Figure 4.2b, we illustrate the dynamics of X_t under the ID policy and the choice of the focus set.

The ID policy again has an $O(1/\sqrt{N})$ optimality gap, as stated in Theorem 4.3 below; the proof of Theorem 4.3 is provided in Section 4.7.

Theorem 4.3 (Optimality gap of ID policy). *Consider an N -armed restless bandit problem with the single-armed MDP $(\mathbb{S}, \mathbb{A}, P, r)$ and budget αN for $0 < \alpha < 1$. Assume that the optimal single-armed policy induces an aperiodic unichain (Assumption 4.1). Let π be the ID policy (Algorithm 5). Then for all N and initial state vector $\mathbf{S}_0$, the optimality gap of π is bounded as*

$$R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0) \leq \frac{C_{\text{ID}}}{\sqrt{N}}, \quad (4.8)$$

where C_{ID} is a constant depending on $r_{\max}$, $|\mathbb{S}|$, $\beta \triangleq \min\{\alpha, 1 - \alpha\}$, and $P_{\bar{\pi}^*}$; the explicit expression of C_{ID} is given in the proof.

Remark 4.4 (Comparison between the ID policy and the set-expansion policy). The ID policy is simpler to implement and does not require an explicit calculation of the focus set. Moreover, in simulations, the ID policy often performs slightly better than the set-expansion policy. We also notice in simulations that a larger set of arms are able to persistently follow $\bar{\pi}^*$ under the ID policy than under the set-expansion policy (see Appendix 4.H.5 for a closer investigation of this phenomenon). On the other hand, under the ID policy, the arms with higher IDs have less chance to follow $\bar{\pi}^*$, whereas the set-expansion policy is ID-oblivious.

Remark 4.5. Due to the homogeneity of the arms, it is actually sufficient to focus on the scaled state-count vector $X_t([N])$ as the state representation of the RB problem and design policies based on $X_t([N])$. In fact, most prior work on homogeneous RBs [107, 108, 266, 274] focuses on $X_t([N])$ rather than X_t as the state representation. However, in our case, neither the set-expansion policy nor the ID policy is Markovian with respect to $X_t([N])$, which naturally leads to the question: Does there exist a focus-set policy that makes $X_t([N])$ a Markov chain and achieves the $O(1/\sqrt{N})$ optimality gap? In Appendix 4.F,

we construct another focus-set policy termed *set-optimization policy* that indeed satisfies these two requirements. The basic idea of the set-optimization policy is to update the focus set D_t by minimizing a Lyapunov function of $X_t(D_t)$.

4.4.4 Connection with existing policies

As we intuitively explained in previous subsections, a focus-set policy achieves asymptotic optimality if most arms independently take actions according to the optimal single-armed policy $\bar{\pi}^*$ in the steady state. In this subsection, we consider the policies in the prior work, the LP-Priority policies [108, 266] and the FTVA policy [149], and investigate *the number of arms that follow $\bar{\pi}^*$* under these policies. This unified view through the number of arms that follow $\bar{\pi}^*$ can help us develop a better understanding of the connection among these policies and the roles of these assumptions in proving asymptotic optimality.

We first consider LP-Priority policies, which achieve asymptotic optimality under the GAP assumption [108, 266]. Recall that an LP-Priority policy assigns a priority order to each state, and prioritizes activating the arms in the high-priority states. The priority orders must be compatible with the optimal single-armed policy $\bar{\pi}^*$ in the following way. Assuming that $y^*(s, 1) + y^*(s, 0) > 0$ for all $s \in \mathbb{S}$, the state space can be partitioned into three subsets: $S^+ \triangleq \{s \in \mathbb{S} : \bar{\pi}^*(1|s) = 1\}$, $S^0 \triangleq \{s \in \mathbb{S} : 0 < \bar{\pi}^*(1|s) < 1\}$, and $S^- \triangleq \{s \in \mathbb{S} : \bar{\pi}^*(1|s) = 0\}$, where one can always find a $\bar{\pi}^*$ such that $|S^0| \leq 1$ and we let $S^0 = \{\tilde{s}\}$ if $|S^0| = 1$. The definition of an LP-Priority policy requires the priorities of the states in S^+ to be higher than those in S^0 , and further higher than those in S^- (Definition 4.4 of [266]).

Here is a heuristic way to see why GAP implies the asymptotic optimality of an LP-Priority policy: When GAP holds, the state-count vector of the system concentrates around μ^* in the steady state (see, e.g., Lemma 12 in [107]). The relation $\sum_{s \in S^+} \mu^*(s) + \sum_{s \in S^0} \mu^*(s) \bar{\pi}^*(1|s) = \sum_{s \in \mathbb{S}} y^*(s, 1) = \alpha$, along with the concentration of the state-count vector, suggests that, in an approximate sense, all arms in S^+ are activated, a fraction $\bar{\pi}^*(1|\tilde{s})$ of arms in S^0 are activated if $|S^0| = 1$, and all arms in S^- remain passive. Thus, in the steady state, for each state $s \in \mathbb{S}$, the fraction of active arms under an LP-Priority policy approximately coincides with the fraction of active arms if the actions were sampled by $\bar{\pi}^*$. Consequently, nearly all arms can be considered as following $\bar{\pi}^*$. On the other hand, when GAP fails, the scaled state-count vector may significantly deviate from μ^* , making it infeasible to

activate a $\bar{\pi}^*(1|s)$ fraction of arms for each state $s \in \mathbb{S}$. In this case, only a limited subset of arms can be considered as following $\bar{\pi}^*$.

Next, we discuss the FTVA policy from Chapter 3 [149], an asymptotically optimal policy that relies on the SA condition. FTVA is simulation-based: it first generates ideal actions by simulating a virtual, unconstrained system where each arm independently follows $\bar{\pi}^*$. It then lets as many arms follow these virtual actions as the budget constraint allows. The asymptotic optimality of FTVA hinges on the SA condition, which guarantees that the real states of most arms remain aligned with their virtual counterparts, thereby ensuring that their real actions are effectively drawn from $\bar{\pi}^*$ conditioned on the real states.

FTVA is similar to our policies in the sense that it also maintains a set of arms that follow $\bar{\pi}^*$ persistently for a long time (referred to as the “good arms” in Chapter 3 [149]). However, for the rest of the arms whose virtual and real states are misaligned, FTVA *passively* waits for them to re-align on their own, which is guaranteed to happen only when SA holds. In contrast, our policies *actively* expand the focus set whenever the empirical state distribution of the focus set is sufficiently close to μ^* , which is guaranteed to happen if the optimal single-armed policy $\bar{\pi}^*$ induces an aperiodic unichain.

4.4.5 Alternative formulation with inequality budget constraint

The set-expansion policy and the ID policy can be extended to the alternative formulation of the RB problem with the inequality budget constraint:

$$\underset{\text{policy } \pi}{\text{maximize}} \quad R^-(\pi, \mathbf{S}_0) \quad (4.9a)$$

$$\text{subject to} \quad \sum_{i \in [N]} A_t^\pi(i) \leq \alpha N, \quad \forall t = 0, 1, 2, \dots \quad (4.9b)$$

The modified pseudo-code of the two policies is given in Algorithms 7 and 8 of Appendix 4.G. The changes in the policies are summarized below.

- The equality budget constraint of the LP relaxation (4.2a) changes to an inequality.
- In the action rectification step of the set-expansion policy (Lines 10–15 of Algorithm 4), we should instead take $A_t(i) = \hat{A}_t(i)$ for all $i \in D_t$ whenever $\sum_{i \in D_t} \hat{A}_t(i) \leq \alpha N$.

- In the action rectification step of the ID policy (Lines 6–8 of Algorithm 5), we should instead take $N_t^{\bar{\pi}^*} = N$ and $A_t(i) = \hat{A}_t(i)$ for all $i \in [N]$ whenever $\sum_{i \in [N]} \hat{A}_t(i) \leq \alpha N$.

We show that the modified set-expansion policy and the modified ID policy also achieve $O(1/\sqrt{N})$ optimality gaps for the inequality-constraint setting, as stated in the following theorems.

Theorem 4.6. *Consider the N -armed restless bandit problem with the inequality constraint (4.9) and assume that the optimal single-armed policy $\bar{\pi}^*$ induces an aperiodic unichain. Let π be the set-expansion policy defined in Algorithm 7. For any N and initial state vector $\mathbf{S}_0$, the optimality gap of π is bounded as*

$$R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0) \leq \frac{C'_{\text{SE}}}{\sqrt{N}}, \quad (4.10)$$

for some constant C'_{SE} depending on $r_{\max}$, $|\mathbb{S}|$, $\beta \triangleq \min\{\alpha, 1 - \alpha\}$, and $P_{\bar{\pi}^*}$.

Theorem 4.7. *Consider the N -armed restless bandit problem with the inequality constraint (4.9) and assume that the optimal single-armed policy $\bar{\pi}^*$ induces an aperiodic unichain. Let π be the ID policy defined in Algorithm 8. For any N and initial state vector $\mathbf{S}_0$, the optimality gap of π is bounded as*

$$R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0) \leq \frac{C'_{\text{ID}}}{\sqrt{N}}, \quad (4.11)$$

for some constant C'_{ID} depending on $r_{\max}$, $|\mathbb{S}|$, $\beta \triangleq \min\{\alpha, 1 - \alpha\}$, and $P_{\bar{\pi}^*}$.

The proofs of Theorems 4.6 and 4.7 are almost identical to the proofs of Theorems 4.2 and 4.3 with the only differences lying in the proofs of Lemmas 4.14 and 4.19. We refer the reader to Appendix 4.G for the details.

4.5 A meta-theorem for focus-set policies and the proof

In this section, we establish a meta-theorem, Theorem 4.9, which provides sufficient conditions for a focus-set policy to have an $O(1/\sqrt{N})$ optimality gap. The meta-theorem and its conditions are stated in Section 4.5.1, followed by its proof in Section 4.5.2.

The meta-theorem contains the main technical novelty of our analysis. In the subsequent sections, we will simply verify that the set-expansion

policy and the ID policy satisfy these conditions under the aperiodic unichain assumption, thereby proving the optimality gap bounds in Theorem 4.2 and Theorem 4.3.

4.5.1 Meta-theorem on $O(1/\sqrt{N})$ optimality gaps of focus-set policies

We now state a set of conditions which, once satisfied by a focus-set policy, guarantees an $O(1/\sqrt{N})$ optimality gap.

To begin with, we define a class of functions called the *subset Lyapunov functions*, which are indexed by a collection of subsets $D \subseteq [N]$. The subset Lyapunov function indexed by D upper bounds the distance between $x(D)$ and $m(D)\mu^*$, and decreases geometrically if the arms in D follow the optimal single-armed policy $\bar{\pi}^*$ indefinitely. The formal definition is given below.

Definition 4.8 (Subset Lyapunov functions). Let $\mathcal{D}$ be a collection of subsets of $[N]$. Consider a class of functions $\{h(\cdot, D) : D \in \mathcal{D}\}$, where each $h(\cdot, D)$ maps a system state x to a real value that depends only on the states of the arms in D . This class of functions is referred to as the *subset Lyapunov functions* for the policy $\bar{\pi}^*$ if they satisfy the following conditions:

1. (Drift condition for a fixed D). There exist constants $\rho_2 \in (0, 1)$ and $K_{\text{drift}} > 0$ such that for any $D \in \mathcal{D}$ and any system state x ,

$$\mathbb{E}[h(X_1, D) \mid X_0 = x, A_0(i) \sim \bar{\pi}^*(\cdot \mid S_0(i)) \forall i \in D] \leq \rho_2 h(x, D) + \frac{K_{\text{drift}}}{\sqrt{N}}. \quad (4.12)$$

2. (Distance domination). There exists a constant $K_{\text{dist}} > 0$ such that for any $D \in \mathcal{D}$ and any system state x ,

$$h(x, D) \geq K_{\text{dist}} \|x(D) - m(D)\mu^*\|_1. \quad (4.13)$$

3. (Lipschitz continuity in D). There exists a constant $L_h > 0$ such that for any $D, D' \in \mathcal{D}$ with $D \subseteq D'$ and any system state x ,

$$|h(x, D') - h(x, D)| \leq L_h (m(D') - m(D)). \quad (4.14)$$

As an example, in Section 4.6, we will define a weighted L_2 norm, $\|v\|_W \triangleq \sqrt{vWv^\top}$ for some weight matrix W . We will show that the class of functions $\{h_W(\cdot, D)\}_{D \subseteq [N]}$ with $h_W(x, D) = \|x(D) - m(D)\mu^*\|_W$ satisfies the definition of subset Lyapunov functions.

While the subset Lyapunov function $h(\cdot, D)$ is constructed to witness the convergence of $X_t(D)$ to $m(D)\mu^*$ for a *fixed* set D , in a focus-set policy, the set D_t is not fixed but rather is chosen dynamically. Below we introduce three conditions on D_t , which would allow us to use the subset Lyapunov functions to establish the asymptotic optimality of a focus set policy.

Condition 1 requires that most arms in the focus set D_t conform to the actions sampled from $\bar{\pi}^*$.

Condition 1 (Majority conformity). *Let $K_{\text{conf}} > 0$ be a constant. For any time step $t \geq 0$, there exists $D'_t \subseteq D_t$ such that for any $i \in D'_t$, the policy chooses $A_t(i) = \hat{A}_t(i)$, and*

$$\mathbb{E}[m(D_t \setminus D'_t) \mid X_t, D_t] \leq \frac{K_{\text{conf}}}{\sqrt{N}} \quad a.s. \quad (4.15)$$

Condition 2 requires that D_t changes in a set-inclusive manner and does not shrink much in expectation.

Condition 2 (Almost non-shrinking). *For any time step $t \geq 0$, either $D_{t+1} \supseteq D_t$ or $D_{t+1} \subseteq D_t$. Moreover, there exists a constant $K_{\text{mono}} > 0$ such that for any $t \geq 0$,*

$$\mathbb{E}[(m(D_t) - m(D_{t+1}))^+ \mid X_t, D_t] \leq \frac{K_{\text{mono}}}{\sqrt{N}} \quad a.s. \quad (4.16)$$

Condition 3 requires that $m(D_t)$, the fraction of arms covered by D_t , is sufficiently large with respect to a subset Lyapunov function on D_t .

Condition 3 (Sufficient coverage). *There exist a class of subset Lyapunov functions $\{h(\cdot, D) : D \in \mathcal{D}\}$ and constants $L_{\text{cov}} > 0, K_{\text{cov}} > 0$ such that for any time step $t \geq 0$,*

$$1 - m(D_t) \leq L_{\text{cov}} h(X_t, D_t) + \frac{K_{\text{cov}}}{\sqrt{N}} \quad a.s. \quad (4.17)$$

We remark that Conditions 1 and 2 are generally easier to satisfy when the focus set D_t is small, while Condition 3 requires D_t to be large.

We are now ready to state the meta-theorem, which establishes an $O(1/\sqrt{N})$ bound on the optimality gap of a focus-set policy that satisfies the above conditions.

Theorem 4.9 (Meta-theorem on optimality gap of focus-set policies). *Consider an N -armed restless bandit problem with the single-armed MDP $(\mathbb{S}, \mathbb{A}, P, r)$ and budget αN for $0 < \alpha < 1$. Assume that the optimal single-armed policy induces an aperiodic unichain (Assumption 4.1). Let π be a focus-set policy given in Algorithm 3 satisfying Conditions 1, 2, and 3 for a class of subset Lyapunov functions $\{h(\cdot, D)\}_{D \in \mathcal{D}}$, then*

$$R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0) \leq r_{\max} \left(\left(\frac{1}{K_{\text{dist}}} + \frac{2}{L_h} \right) \frac{K_1}{1 - \rho_1} + 2K_{\text{conf}} \right) \frac{1}{\sqrt{N}} \quad \forall N, \mathbf{S}_0, \quad (4.18)$$

where $\rho_1 = 1 - (1 - \rho_2)/(1 + L_h L_{\text{cov}})$ and $K_1 = K_{\text{drift}} + 2L_h K_{\text{conf}} + 2L_h K_{\text{mono}} + (1 - \rho_2)L_h K_{\text{cov}}/(1 + L_h L_{\text{cov}})$.

4.5.2 Proof of Theorem 4.9 assuming convergence to stationary distribution.

In this section, we prove Theorem 4.9 under the assumption that the focus set policy induces a Markov chain that converges to a unique stationary distribution. This simplified setting allows us to present the key proof ideas more clearly. The proof for the general case follows essentially the same line of argument and is given in Appendix 4.D. The general proof also bounds the finite-time expected reward as a by-product.

Under the above assumption, we use $\mathbf{S}_\infty, \hat{\mathbf{A}}_\infty, \mathbf{A}_\infty, X_\infty, D_\infty$ to denote the random variables following the stationary distributions of $\mathbf{S}_t, \hat{\mathbf{A}}_t, \mathbf{A}_t, X_t, D_t$, respectively. With this notation, the long-run average reward of the policy π is equal to $R(\pi, \mathbf{S}_0) = \frac{1}{N} \sum_{i \in [N]} \mathbb{E}[r(S_\infty(i), A_\infty(i))]$.

Proof of Theorem 4.9 assuming convergence to stationary distribution. Our proof is structured into two steps: understanding the optimality gap, and bounding the Lyapunov function.

Understanding the optimality gap. Recall that the optimality gap can be upper bounded as $R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0) \leq R^{\text{rel}} - R(\pi, \mathbf{S}_0)$, where R^{rel} is the expected reward associated with the optimal steady-state state-action distribution $y^* = (y^*(s, a))_{s \in \mathbb{S}, a \in \mathbb{A}}$. Then

$$\begin{aligned} R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0) \\ \leq R^{\text{rel}} - R(\pi, \mathbf{S}_0) \end{aligned}$$

$$\begin{aligned}
&= \sum_{s \in \mathbb{S}, a \in \mathbb{A}} r(s, a) y^*(s, a) - \frac{1}{N} \sum_{i \in [N]} \mathbb{E} \left[r(S_\infty(i), A_\infty(i)) \right] \\
&\leq \sum_{s \in \mathbb{S}, a \in \mathbb{A}} r(s, a) y^*(s, a) - \frac{1}{N} \sum_{i \in [N]} \mathbb{E} \left[r(S_\infty(i), \hat{A}_\infty(i)) \right] \\
&\quad + \frac{2r_{\max}}{N} \sum_{i \in [N]} \mathbb{P} \left(\hat{A}_\infty(i) \neq A_\infty(i) \right) \\
&\leq \sum_{s \in \mathbb{S}, a \in \mathbb{A}} r(s, a) \left(y^*(s, a) - \bar{\pi}^*(a|s) \mathbb{E} [X_\infty([N], s)] \right) \\
&\quad + 2r_{\max} \mathbb{E} [1 - m(D'_\infty)] \\
&= \sum_{s \in \mathbb{S}, a \in \mathbb{A}} r(s, a) \bar{\pi}^*(a|s) \left(\mu^*(s) - \mathbb{E} [X_\infty([N], s)] \right) \\
&\quad + 2r_{\max} \mathbb{E} [1 - m(D'_\infty)] \\
&\leq r_{\max} \mathbb{E} [\|\mu^* - X_\infty([N])\|_1] + 2r_{\max} \mathbb{E} [1 - m(D_\infty)] \\
&\quad + \frac{2r_{\max} K_{\text{conf}}}{\sqrt{N}}, \tag{4.19}
\end{aligned}$$

where D'_∞ is the subset of D_∞ assumed in Condition 1, which satisfies $m(D'_\infty) \geq m(D_\infty) - K_{\text{conf}}/\sqrt{N}$. Therefore, to bound the optimality gap, it suffices to bound $\mathbb{E}[\|\mu^* - X_\infty([N])\|_1]$, which is the distributional distance, and $\mathbb{E}[1 - m(D_\infty)]$, which is the size of the complement of the focus set.

In this proof, we construct a Lyapunov function that can be viewed as an upper bound on a weighted sum of the two terms in (4.19). In particular, consider the following Lyapunov function

$$V(x, D) = h(x, D) + L_h(1 - m(D)). \tag{4.20}$$

Let us first see how the terms in (4.19) are upper bounded by $\mathbb{E}[V(X_\infty, D_\infty)]$. For the first term, it is easy to see that $K_{\text{dist}} \|\mu^* - X_\infty([N])\|_1 \leq h(X_\infty, [N])$ by the distance domination property of h (cf. Equation (4.13)). Then by the Lipschitz continuity of h , we have $h(X_\infty, [N]) \leq h(X_\infty, D_\infty) + L_h(1 - m(D_\infty)) = V(X_\infty, D_\infty)$. Thus, $\mathbb{E}[\|\mu^* - X_\infty([N])\|_1] \leq \mathbb{E}[V(X_\infty, D_\infty)]/K_{\text{dist}}$. For the second term, clearly

$\mathbb{E}[1 - m(D_\infty)] \leq \mathbb{E}[V(X_\infty, D_\infty)]/L_h$. Therefore, the upper bound in (4.19) can be further bounded as

$$R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0) \leq r_{\max} \left(\frac{1}{K_{\text{dist}}} + \frac{2}{L_h} \right) \mathbb{E}[V(X_\infty, D_\infty)] + \frac{2r_{\max}K_{\text{conf}}}{\sqrt{N}}, \quad (4.21)$$

which makes it sufficient to bound $\mathbb{E}[V(X_\infty, D_\infty)]$.

Bounding the Lyapunov function. We establish an upper bound on $\mathbb{E}[V(X_\infty, D_\infty)]$ by proving the following drift condition: for any $t \geq 0$,

$$\mathbb{E}[V(X_{t+1}, D_{t+1}) \mid X_t, D_t] \leq \rho_1 V(X_t, D_t) + \frac{K_1}{\sqrt{N}}, \quad (4.22)$$

for some constants $\rho_1 \in (0, 1)$ and $K_1 > 0$. To prove (4.22), observe that for any time step $t \geq 0$,

$$\begin{aligned} V(X_{t+1}, D_{t+1}) &= h(X_{t+1}, D_{t+1}) + L_h(1 - m(D_{t+1})) \\ &\leq \left(h(X_{t+1}, D_t) + L_h |m(D_{t+1}) - m(D_t)| \right) \\ &\quad + \left(L_h(1 - m(D_t)) + L_h(m(D_t) - m(D_{t+1})) \right) \\ &= h(X_{t+1}, D_t) + L_h(1 - m(D_t)) + 2L_h(m(D_t) - m(D_{t+1}))^+, \end{aligned} \quad (4.23)$$

where we have used the facts that $D_{t+1} \supseteq D_t$ or $D_{t+1} \subseteq D_t$ (Condition 2) and the Lipschitz continuity of $h(x, D)$ in D . Subtracting $V(X_t, D_t)$ and taking the expectation, we obtain the *key decomposition* below:

$$\begin{aligned} &\mathbb{E}[V(X_{t+1}, D_{t+1}) \mid X_t, D_t] - V(X_t, D_t) \\ &\leq \mathbb{E}[h(X_{t+1}, D_t) \mid X_t, D_t] - h(X_t, D_t) \end{aligned} \quad (4.24)$$

$$+ 2L_h \mathbb{E}[(m(D_t) - m(D_{t+1}))^+ \mid X_t, D_t]. \quad (4.25)$$

where the term in (4.24) represents the contribution of state transitions to the drift of $V(X_t, D_t)$, and the term in (4.25) represents the contribution of set updates.

We first upper bound the term $\mathbb{E}[h(X_{t+1}, D_t) \mid X_t, D_t] - h(X_t, D_t)$ in (4.24). Note that this bound would immediately follow from the drift condition of subset Lyapunov functions if all the arms in D_t were to follow the ideal actions. By the majority conformity property of the focus set D_t (Condition 1), there

exists $D'_t \subseteq D_t$ such that for any $i \in D'_t$, the policy chooses $A_t(i) = \hat{A}_t(i)$, and $\mathbb{E}[m(D_t \setminus D'_t) \mid X_t, D_t] = O(1/\sqrt{N})$. Let X'_{t+1} be a random element denoting the system state at time $t+1$ if $A_t(i) = \hat{A}_t(i)$ for all $i \in [N]$. We couple X_{t+1} with X'_{t+1} such that they have the same states on the set D'_t , and thus $h(X_{t+1}, D'_t) = h(X'_{t+1}, D'_t)$. Then

$$\begin{aligned} & \mathbb{E}[h(X_{t+1}, D_t) \mid X_t, D_t] \\ &= \mathbb{E}[h(X'_{t+1}, D_t) + (h(X_{t+1}, D_t) - h(X'_{t+1}, D_t)) \mid X_t, D_t] \\ &= \mathbb{E}[h(X'_{t+1}, D_t) + (h(X_{t+1}, D_t) - h(X_{t+1}, D'_t)) \\ &\quad + (h(X'_{t+1}, D'_t) - h(X'_{t+1}, D_t)) \mid X_t, D_t] \\ &\leq \rho_2 h(X_t, D_t) + \frac{K_{\text{drift}}}{\sqrt{N}} + 2L_h \mathbb{E}[m(D_t \setminus D'_t) \mid X_t, D_t] \\ &\leq \rho_2 h(X_t, D_t) + \frac{K_{\text{drift}} + 2L_h K_{\text{conf}}}{\sqrt{N}}, \end{aligned}$$

where we have used the drift condition and the Lipschitz continuity of h . It follows that

$$\mathbb{E}[h(X_{t+1}, D_t) \mid X_t, D_t] - h(X_t, D_t) \leq -(1 - \rho_2)h(X_t, D_t) + \frac{K_{\text{drift}} + 2L_h K_{\text{conf}}}{\sqrt{N}}. \quad (4.26)$$

Next, to bound the term in (4.25), we simply apply Condition 2:

$$2L_h \mathbb{E}[(m(D_t) - m(D_{t+1}))^+ \mid X_t, D_t] \leq \frac{2L_h K_{\text{mono}}}{\sqrt{N}}. \quad (4.27)$$

Combining the above bounds for (4.24) and (4.25), we get

$$\begin{aligned} & \mathbb{E}[V(X_{t+1}, D_{t+1}) \mid X_t, D_t] - V(X_t, D_t) \\ &\leq -(1 - \rho_2)h(X_t, D_t) + \frac{K_{\text{drift}} + 2L_h K_{\text{conf}} + 2L_h K_{\text{mono}}}{\sqrt{N}}. \end{aligned} \quad (4.28)$$

To get (4.22), it remains to upper bound the $-(1 - \rho_2)h(X_t, D_t)$ term. By the sufficient coverage condition (Condition 3), $1 - m(D_t) \leq L_{\text{cov}}h(X_t, D_t) + K_{\text{cov}}/\sqrt{N}$, so

$$V(X_t, D_t) = h(X_t, D_t) + L_h(1 - m(D_t)) \leq (1 + L_h L_{\text{cov}})h(X_t, D_t) + \frac{L_h K_{\text{cov}}}{\sqrt{N}}.$$

Upper bounding the $-(1 - \rho_2)h(X_t, D_t)$ term in (4.28) using the above inequality, we get

$$\mathbb{E}[V(X_{t+1}, D_{t+1}) \mid X_t, D_t] \leq \rho_1 V(X_t, D_t) + \frac{K_1}{\sqrt{N}},$$

where $\rho_1 = 1 - (1 - \rho_2)/(1 + L_h L_{\text{cov}})$ and $K_1 = K_{\text{drift}} + 2L_h K_{\text{conf}} + 2L_h K_{\text{mono}} + (1 - \rho_2)L_h K_{\text{cov}}/(1 + L_h L_{\text{cov}})$. This is the bound in (4.22) that we set out to prove.

Taking the expectation on both sides of (4.22) letting $t \rightarrow \infty$, we have

$$\mathbb{E}[V(X_\infty, D_\infty)] \leq \rho_1 \mathbb{E}[V(X_\infty, D_\infty)] + \frac{K_1}{\sqrt{N}},$$

which implies that

$$\mathbb{E}[V(X_\infty, D_\infty)] \leq \frac{K_1}{(1 - \rho_1)\sqrt{N}}. \quad (4.29)$$

This completes the proof of Theorem 4.9. $\square$

Remark 4.10. We conclude this section by a remark on our use of the *bivariate Lyapunov functions* $h(x, D)$ and $V(x, D) = h(x, D) + L_h(1 - m(D))$. By definition, the subset Lyapunov function $h(x, D)$ depends on the system state x only through $x(D)$. This means that for fixed D , the drifts of $h(x, D)$ and $V(x, D)$ only depend on the state transitions of the arms in D . When D is chosen appropriately, most arms in D can follow $\bar{\pi}^*$ under the budget constraint, thus inheriting the convergence and concentration properties of the aperiodic unichain induced by $\bar{\pi}^*$. Therefore, the auxiliary variable D provides the flexibility of focusing on a subset of arms so that the drift is easy to bound and expanding the subset gradually to the entire system.

For the ID policy, D_t is determined by the system state X_t , and hence $h(X_t, D_t)$ can be written as a function of X_t alone. Even in this case, using a bivariate h is beneficial, as it allows us to decouple the two variables—in particular, quantities like $h(X_{t+1}, D_t)$ play a prominent role in our proof of Theorem 4.9.

Our use of bivariate Lyapunov functions departs from most prior work on the RB problem [107, 108, 266, 274, 283], whose analysis is in terms of the full system state $X_t([N])$, under which the dynamics of arms in a subset is less visible. We expect that our approach is useful for a broader class of

problems where the system state consists of multiple components, a subset of which have a more tractable dynamic at a given time. In this case, one may construct a Lyapunov function that can zoom into this more tractable subset and seek to gradually expand the subset based on the system state.

4.6 Proof of Theorem 4.2 (Optimality gap of set-expansion policy)

In this section, we prove Theorem 4.2 using the framework established in Section 4.5. This section is organized as follows. In Section 4.6.1, we define the subset Lyapunov functions for the set-expansion policy. In Section 4.6.2, we present three lemmas verifying that the set-expansion policy satisfies Conditions 1, 2 and 3, and prove Theorem 4.2 by citing the meta-theorem Theorem 4.9. These three lemmas are subsequently proved in Sections 4.6.3, 4.6.4 and 4.6.5.

4.6.1 Subset Lyapunov functions

To construct the subset Lyapunov functions, we consider the L_2 norm weighted by a carefully constructed matrix W defined below.

Definition 4.11. Let W be an $|\mathbb{S}|$ -by- $|\mathbb{S}|$ matrix given by

$$W = \sum_{k=0}^{\infty} (P_{\bar{\pi}^*} - \Xi)^k (P_{\bar{\pi}^*}^\top - \Xi^\top)^k, \quad (4.30)$$

where Ξ is an $|\mathbb{S}|$ -by- $|\mathbb{S}|$ matrix with each row being μ^* . Let λ_W denote maximal eigenvalue of W .

Intuitively, the infinite series in (4.30) is convergent because $P_{\bar{\pi}^*}$ is aperiodic unichain, Ξ is the limit of $P_{\bar{\pi}^*}^k$ as $k \rightarrow \infty$, and $(P_{\bar{\pi}^*} - \Xi)^k = P_{\bar{\pi}^*}^k - \Xi$. In Appendix 4.E.1, we formally show that the matrix W is well-defined and positive definite, with eigenvalues in the range $[1, \lambda_W]$. Our next lemma shows that $P_{\bar{\pi}^*}$ is a pseudo-contraction under the W -weighted L_2 norm.

Lemma 4.12 (Pseudo-contraction under W -weighted L_2 norm). *Suppose $P_{\bar{\pi}^*}$ is an aperiodic unichain on $\mathbb{S}$. Then for any distribution $v \in \Delta(\mathbb{S})$,*

$$\|v P_{\bar{\pi}^*} - \mu^*\|_W \leq \left(1 - \frac{1}{2\lambda_W}\right) \|v - \mu^*\|_W, \quad (4.31)$$

where λ_W is the maximal eigenvalue of the matrix W given in Definition 4.11, and $\|\cdot\|_W$ is the W -weighted L_2 norm, that is, $\|u\|_W = \sqrt{uWu^\top}$ for any row vector $u \in \mathbb{R}^{|\mathbb{S}|}$.

Now we are ready to define the subset Lyapunov functions. For any system state x and $D \subseteq [N]$, let

$$h_W(x, D) = \|x(D) - m(D)\mu^*\|_W, \quad (4.32)$$

which measures the distance between $x(D)$, the scaled state-count vector on D , and $m(D)\mu^*$, the correspondingly scaled optimal stationary distribution. Note that $h_W(x, D)$ depends only on the states of the arms in D , as required by the definition of subset Lyapunov functions. The next lemma, Lemma 4.13, shows that the class of functions $\{h_W(x, D)\}_{D \subseteq [N]}$ satisfies the definition of subset Lyapunov functions (Definition 4.8). The proof of Lemma 4.13 is provided in Appendix 4.E.2.

Lemma 4.13. *The class of functions $\{h_W(\cdot, D)\}_{D \subseteq [N]}$ defined in (4.32) satisfies that for any system state x and any pair of subsets $D, D' \subseteq [N]$ with $D \subseteq D'$,*

$$\begin{aligned} \mathbb{E}[h_W(X_1, D) \mid X_0 = x, A_0(i) \sim \bar{\pi}^*(\cdot \mid S_0(i)) \forall i \in D] \\ \leq \left(1 - \frac{1}{2\lambda_W}\right) h_W(x, D) + \frac{2\lambda_W^{1/2}}{\sqrt{N}} \end{aligned} \quad (4.33)$$

$$h_W(x, D) \geq \frac{1}{|\mathbb{S}|^{1/2}} \|x(D) - m(D)\mu^*\|_1 \quad (4.34)$$

$$|h_W(x, D) - h_W(x, D')| \leq L_W(m(D') - m(D)), \quad (4.35)$$

where the Lipschitz constant $L_W = 2\lambda_W^{1/2}$. These inequalities imply the drift condition, distance dominance property, and Lipschitz continuity in Definition 4.8, respectively. Consequently, $\{h_W(x, D)\}_{D \subseteq [N]}$ is a class of subset Lyapunov functions for the optimal single-armed policy $\bar{\pi}^*$.

4.6.2 Lemmas verifying Conditions 1, 2 and 3; proof of Theorem 4.2

Next, we establish Lemmas 4.14, 4.15 and 4.16, which verify that the set-expansion policy in Algorithm 4 satisfies Conditions 1, 2 and 3, respectively. Then we apply Theorem 4.9 to prove Theorem 4.2.

Lemma 4.14 (Set-expansion policy satisfies Condition 1). *Consider the set-expansion policy (Algorithm 4). For any $t \geq 0$, there exists a subset $D'_t \subseteq D_t$ such that for any $i \in D'_t$, $A_t(i) = \hat{A}_t(i)$, and*

$$\mathbb{E}[m(D_t \setminus D'_t) \mid X_t, D_t] \leq \frac{1}{\sqrt{N}} + \frac{1}{N} \quad a.s. \quad (4.36)$$

Lemma 4.15 (Set-expansion policy satisfies Condition 2). *Consider the set-expansion policy in Algorithm 4. For any $t \geq 0$,*

$$\mathbb{E}[(m(D_t) - m(D_{t+1}))^+ \mid X_t, D_t] \leq \frac{|\mathbb{S}|^{1/2} + 1}{\beta\sqrt{N}} + \frac{1 + (\beta + 1)|\mathbb{S}|}{\beta N} \quad a.s. \quad (4.37)$$

Lemma 4.16 (Set-expansion policy satisfies Condition 3). *Consider the set-expansion policy in Algorithm 4. For any $t \geq 0$,*

$$1 - m(D_t) \leq \frac{|\mathbb{S}|^{1/2}}{\beta} h_W(X_t, D_t) + \frac{2}{\beta N} \quad a.s. \quad (4.38)$$

Proof of Theorem 4.2. By Lemmas 4.14, 4.15 and 4.16, the set-expansion policy satisfies Conditions 1, 2 and 3 with the subset Lyapunov functions $\{h_W(x, D)\}_{D \subseteq [N]}$. Applying Theorem 4.9 and substituting the constants, we get

$$R^{\text{rel}} - R(\pi, \mathbf{S}_0) \leq \frac{256r_{\max}\lambda_W^2|\mathbb{S}|^2}{\beta^2\sqrt{N}},$$

which implies the optimality gap bound in Theorem 4.2. Note that we relax all $1/N$ factors to $1/\sqrt{N}$ when deriving the bound. $\square$

4.6.3 Proof of Lemma 4.14

Proof of Lemma 4.14. Recall that under the set-expansion policy, the actions in the focus set $(A_t(i))_{i \in D_t}$ are chosen according to the ideal actions $(\hat{A}_t(i))_{i \in D_t}$ based on the following rules:

- When $\sum_{i \in D_t} \hat{A}_t(i) > \alpha N$, the set-expansion policy chooses αN arms in D_t with $\hat{A}_t(i) = 1$ and sets $A_t(i) = \hat{A}_t(i)$.
- When $\sum_{i \in D_t} (1 - \hat{A}_t(i)) > (1 - \alpha)N$, the set-expansion policy chooses $(1 - \alpha)N$ arms in D_t with $\hat{A}_t(i) = 0$ and sets $A_t(i) = \hat{A}_t(i)$.
- Otherwise, the set-expansion policy sets $A_t(i) = \hat{A}_t(i)$ for all $i \in D_t$.

Let $D'_t = \{i \in D_t : A_t(i) = \hat{A}_t(i)\}$. Then we have

$$|D_t \setminus D'_t| = \left(\sum_{i \in D_t} \hat{A}_t(i) - \alpha N \right)^+ + \left(\sum_{i \in D_t} (1 - \hat{A}_t(i)) - (1 - \alpha)N \right)^+. \quad (4.39)$$

Therefore, if we can show that for any $t \geq 0$,

$$\mathbb{E} \left[\left(\sum_{i \in D_t} \hat{A}_t(i) - \alpha N \right)^+ + \left(\sum_{i \in D_t} (1 - \hat{A}_t(i)) - (1 - \alpha)N \right)^+ \mid X_t, D_t \right] \leq \sqrt{N} \quad a.s., \quad (4.40)$$

we will have $\mathbb{E}[m(D_t \setminus D'_t) \mid X_t, D_t] \leq 1/\sqrt{N}$, which will complete the proof. The remainder of the proof is dedicated to proving (4.40).

Consider the *scaled expected budget requirement* for a given system state x and subset $D \subseteq [N]$, defined as

$$C_{\bar{\pi}^*}(x, D) \triangleq \frac{1}{N} \mathbb{E} \left[\sum_{i \in D} \hat{A}_t(i) \mid X_t = x \right] = \sum_{s \in \mathbb{S}} x(D, s) \bar{\pi}^*(1|s). \quad (4.41)$$

Then $\mathbb{E}[\sum_{i \in D_t} \hat{A}_t(i) \mid X_t, D_t] = NC_{\bar{\pi}^*}(X_t, D_t)$. By the Cauchy-Schwarz inequality and the fact that given X_t and D_t , $\hat{A}_t(i)$'s are independent across $i \in D_t$, we have

$$\begin{aligned} & \mathbb{E} \left[\left| \sum_{i \in D_t} \hat{A}_t(i) - NC_{\bar{\pi}^*}(X_t, D_t) \right| \mid X_t, D_t \right] \\ & \leq \mathbb{E} \left[\left(\sum_{i \in D_t} \hat{A}_t(i) - NC_{\bar{\pi}^*}(X_t, D_t) \right)^2 \mid X_t, D_t \right]^{\frac{1}{2}} \\ & = \left(\sum_{i \in D_t} \text{Var}[\hat{A}_t(i) \mid X_t, D_t] \right)^{\frac{1}{2}} \\ & \leq \sqrt{N}. \end{aligned} \quad (4.42)$$

We next prove (4.40) utilizing the bound (4.42). Recall that D_t is chosen with $\delta(X_t, D_t) \geq 0$, i.e., $\|X_t(D_t) - m(D_t)\mu^*\|_1/2 \leq \beta(1 - m(D_t))$. Then $|C_{\bar{\pi}^*}(X_t, D_t) - \alpha m(D_t)|$ can be bounded as

$$|C_{\bar{\pi}^*}(X_t, D_t) - \alpha m(D_t)| = \left| \sum_{s \in \mathbb{S}} (X_t(D_t, s) - m(D_t)\mu^*(s)) \bar{\pi}^*(1|s) \right| \quad (4.43)$$

$$\leq \left| \sum_{s \in \mathbb{S}} \left(X_t(D_t, s) - m(D_t) \mu^*(s) \right) \left(\bar{\pi}^*(1|s) - \frac{1}{2} \right) \right| \quad (4.44)$$

$$\leq \frac{1}{2} \|X_t(D_t) - m(D_t) \mu^*\|_1 \quad (4.45)$$

$$\leq \beta(1 - m(D_t)), \quad (4.46)$$

where (4.43) is because $\sum_{s \in \mathbb{S}} \mu^*(s) \bar{\pi}^*(1|s) = \alpha$, (4.44) is because

$$\sum_{s \in \mathbb{S}} (X_t(D_t, s) - m(D_t) \mu^*(s)) = 0,$$

and (4.45) is because $|\bar{\pi}^*(1|s) - 1/2| \leq 1/2$ for all $s \in \mathbb{S}$. Thus,

$$\begin{aligned} NC_{\bar{\pi}^*}(X_t, D_t) &\leq \alpha|D_t| + \beta(N - |D_t|) \\ &\leq \alpha|D_t| + \alpha(N - |D_t|) = \alpha N, \\ NC_{\bar{\pi}^*}(X_t, D_t) &\geq \alpha|D_t| - \beta(N - |D_t|) \\ &\geq \alpha|D_t| - (1 - \alpha)(N - |D_t|) = \alpha N - (N - |D_t|). \end{aligned}$$

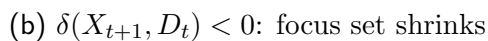
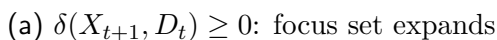
Therefore, we have

$$\begin{aligned} &\mathbb{E} \left[\left(\sum_{i \in D_t} \hat{A}_t(i) - \alpha N \right)^+ + \left(\sum_{i \in D_t} (1 - \hat{A}_t(i)) - (1 - \alpha)N \right)^+ \mid X_t, D_t \right] \\ &= \mathbb{E} \left[\left(\sum_{i \in D_t} \hat{A}_t(i) - \alpha N \right)^+ + \left(\alpha N - (N - |D_t|) - \sum_{i \in D_t} \hat{A}_t(i) \right)^+ \mid X_t, D_t \right] \\ &\leq \mathbb{E} \left[\left| \sum_{i \in D_t} \hat{A}_t(i) - NC_{\bar{\pi}^*}(X_t, D_t) \right| \mid X_t, D_t \right] \\ &\leq \sqrt{N}, \end{aligned} \quad (4.47)$$

where the inequality (4.47) follows from the fact that $(x-a)^+ + (b-x)^+ \leq |x-c|$ for any real numbers x, a, b, c such that $b \leq c \leq a$. This proves (4.40), concluding the proof of Lemma 4.14. $\square$

4.6.4 Proof of Lemma 4.15

In this subsection, we prove Lemma 4.15, which shows that the focus set D_t does not shrink significantly every time step. The intuition of the proof is illustrated in Figure 4.3.



Each dot is a pair $(m(D), h_1(x, D))$, where $h_1(x, D)$ is a shorthand for $0.5\|x(D) - m(D)\mu^*\|_1$. The slack is the vertical gap from the dot to the line $m \mapsto \beta(1 - m)$, so $\delta(x, D) \geq 0$ exactly when the dot lies below the line. The state transition from X_t to X_{t+1} moves the dot vertically to $(m(D_t), h_1(X_{t+1}, D_t))$, and the two panels show the two possible outcomes. The triangular dot in Figure 4.3b is an auxiliary subset $\overline{D} \subseteq D_t$ constructed in the proof; it lies below the line and is nearly as large as D_t , which is what makes the shrinkage small.

Proof of Lemma 4.15. Our proof consists of two steps. In Step 1, we focus on proving the following inequality for each time step t :

$$(m(D_t) - m(D_{t+1}))^+ \leq \frac{1}{\beta}(-\delta(X_{t+1}, D_t))^+ + \frac{K}{N}, \quad (4.48)$$

where $K = (1 + 1/\beta)|\mathbb{S}|$. In Step 2, we utilize (4.48) to bound $\mathbb{E}[(m(D_t) - m(D_{t+1}))^+ | X_t, D_t]$. For the ease of reference, we first recall the definition of the slack $\delta(x, D)$ for some system state x and subset $D \subseteq [N]$, which is heavily used in this proof:

$$\delta(x, D) = \beta(1 - m(D)) - \frac{1}{2} \|x(D) - m(D)\mu^*\|_1. \quad (4.49)$$

Step 1: Proving (4.48). When $\delta(X_{t+1}, D_t) > 0$, the focus set expands and we have $D_{t+1} \supseteq D_t$, so (4.48) trivially holds. We therefore focus on the non-trivial case of $\delta(X_{t+1}, D_t) < 0$, where the focus set shrinks. By definition, D_{t+1} has the largest cardinality among all subsets $\bar{D} \subseteq D_t$ such that $\delta(X_{t+1}, \bar{D}) \geq 0$. We can thus look for a subset $\bar{D} \subseteq D_t$ with the following two properties:

$$\delta(X_{t+1}, \bar{D}) \geq 0 \quad (4.50)$$

$$(m(D_t) - m(\bar{D}))^+ \leq \frac{1}{\beta}(-\delta(X_{t+1}, D_t))^+ + \frac{K}{N}. \quad (4.51)$$

The existence of such a subset $\bar{D}$ implies (4.48) because $(m(D_t) - m(D_{t+1}))^+ \leq (m(D_t) - m(\bar{D}))^+ \leq (-\delta(X_{t+1}, D_t))^+/\beta + K/N$. Geometrically, (4.50) asks for the point $(m(\bar{D}), h_1(X_{t+1}, \bar{D}))$ to be back below the line $m \mapsto \beta(1 - m)$ in Figure 4.3b, while (4.51) asks for $\bar{D}$ to be almost as large as D_t .

We construct $\bar{D}$ based on two cases:

- If $|\delta(X_{t+1}, D_t)|/\beta + K/N \geq m(D_t)$, we let $\bar{D} = \emptyset$;
- Otherwise, we let

$$\gamma = 1 - \frac{1}{m(D_t)} \left(\frac{|\delta(X_{t+1}, D_t)|}{\beta} + \frac{K - |\mathbb{S}|}{N} \right). \quad (4.52)$$

One can verify that $0 < \gamma < 1$. We define $\bar{D}$ as a subset of D_t such that

$$X_{t+1}(\bar{D}, s) = \frac{\lfloor \gamma X_{t+1}(D_t, s)N \rfloor}{N} \quad \forall s \in \mathbb{S}. \quad (4.53)$$

Intuitively, $\bar{D}$ is obtained by sampling γ fraction of arms in each state from D_t , which is the triangular dot in Figure 4.3b; the sampling shrinks m by $m(D_t) - m(\bar{D}) \approx |\delta(X_{t+1}, D_t)| / \beta$, which is just enough to bring the point back below the line $m \mapsto \beta(1 - m)$. Since $|\delta(X_{t+1}, D_t)|$ is small, $\gamma \approx 1 - |\delta(X_{t+1}, D_t)| / (\beta m(D_t))$ is close to 1, so $|D_{t+1}| \geq |\bar{D}|$ is not significantly smaller than $|D_t|$.

We now verify that this construction of $\bar{D}$ satisfies the properties (4.50) and (4.51). When $|\delta(X_{t+1}, D_t)| / \beta + K/N \geq m(D_t)$, we have $\bar{D} = \emptyset$, so these two conditions are trivially true. We can thus focus on the case where $|\delta(X_{t+1}, D_t)| / \beta + K/N < m(D_t)$.

We first show (4.50), i.e., $\delta(X_{t+1}, \bar{D}) \geq 0$. By the definition of $\bar{D}$, we have $m(\bar{D}) \leq \gamma m(D_t)$. Substituting the definitions of γ , K , and $\delta(X_{t+1}, D_t)$, we upper bound $m(\bar{D})$ as

$$m(\bar{D}) \leq m(D_t) + \frac{1}{\beta} \delta(X_{t+1}, D_t) - \frac{|\mathbb{S}|}{\beta N} = 1 - \frac{1}{2\beta} \|X_{t+1}(D_t) - m(D_t)\mu^*\|_1 - \frac{|\mathbb{S}|}{\beta N}.$$

Then we can lower bound $\delta(X_{t+1}, \bar{D})$ using the above upper bound of $m(\bar{D})$:

$$\begin{aligned} \delta(X_{t+1}, \bar{D}) &= \beta(1 - m(\bar{D})) - \frac{1}{2} \|X_{t+1}(\bar{D}) - m(\bar{D})\mu^*\|_1 \\ &\geq \frac{1}{2} \|X_{t+1}(D_t) - m(D_t)\mu^*\|_1 - \frac{1}{2} \|X_{t+1}(\bar{D}) - m(\bar{D})\mu^*\|_1 + \frac{|\mathbb{S}|}{N}. \end{aligned} \tag{4.54}$$

We further lower bound $\|X_{t+1}(D_t) - m(D_t)\mu^*\|_1 - \|X_{t+1}(\bar{D}) - m(\bar{D})\mu^*\|_1$ in (4.54) as

$$\begin{aligned} &\|X_{t+1}(D_t) - m(D_t)\mu^*\|_1 - \|X_{t+1}(\bar{D}) - m(\bar{D})\mu^*\|_1 \\ &\geq \gamma \|X_{t+1}(D_t) - m(D_t)\mu^*\|_1 - \|X_{t+1}(\bar{D}) - m(\bar{D})\mu^*\|_1 \\ &\geq -\|\gamma X_{t+1}(D_t) - \gamma m(D_t)\mu^* - X_{t+1}(\bar{D}) + m(\bar{D})\mu^*\|_1 \\ &\geq -\|\gamma X_{t+1}(D_t) - X_{t+1}(\bar{D})\|_1 - |\gamma m(D_t) - m(\bar{D})| \cdot \|\mu^*\|_1 \\ &\geq -\frac{2|\mathbb{S}|}{N}, \end{aligned}$$

where the first inequality is due to $\gamma < 1$; the second and third inequalities are due to the triangular inequality; the last inequality is due to the definition of $\bar{D}$ in (4.53). Therefore, $\delta(X_{t+1}, \bar{D}) \geq 0$.

Next, we show (4.51). By the definition of $\overline{D}$, we have $m(\overline{D}) \geq \gamma m(D) - |\mathbb{S}|/N$. Substituting the definition of γ in (4.52), we get

$$m(\overline{D}) \geq m(D_t) + \frac{1}{\beta} \delta(X_{t+1}, D_t) - \frac{K - |\mathbb{S}|}{N} - \frac{|\mathbb{S}|}{N} = m(D_t) + \frac{1}{\beta} \delta(X_{t+1}, D_t) - \frac{K}{N},$$

which implies (4.51). Therefore, we have proved the inequality (4.48) claimed at the beginning of this proof.

Step 2: Utilizing (4.48) to bound $\mathbb{E}[(m(D_t) - m(D_{t+1}))^+ \mid X_t, D_t]$. Taking the expectation on both sides of (4.48), we get

$$\mathbb{E}[(m(D_t) - m(D_{t+1}))^+ \mid X_t, D_t] \leq \frac{1}{\beta} \mathbb{E}[(-\delta(X_{t+1}, D_t))^+ \mid X_t, D_t] + \frac{K}{N}. \quad (4.55)$$

Now, it remains to upper bound $\mathbb{E}[(-\delta(X_{t+1}, D_t))^+ \mid X_t, D_t]$. Let X'_{t+1} be a random element denoting the system state at time $t + 1$ if we were able to set $A_t(i) = \hat{A}_t(i)$ for all $i \in [N]$. Let $D'_t \triangleq \{i \in D_t : \hat{A}_t(i) = A_t(i)\}$ (i.e., the subset given in Lemma 4.14), we couple X'_{t+1} and X_{t+1} such that $X'_{t+1}(D'_t) = X_{t+1}(D'_t)$. We make the following decomposition with the help of X'_{t+1} :

$$\begin{aligned} & \mathbb{E}[(-\delta(X_{t+1}, D_t))^+ \mid X_t, D_t] \\ & \leq \mathbb{E}[(-\delta(X_{t+1}, D_t) + \delta(X'_{t+1}, D_t))^+ \mid X_t, D_t] \\ & \quad + \mathbb{E}[(-\delta(X'_{t+1}, D_t))^+ \mid X_t, D_t]. \end{aligned} \quad (4.56)$$

Below, we bound the two terms on the right-hand side of (4.56) separately.

To bound $\mathbb{E}[(-\delta(X_{t+1}, D_t) + \delta(X'_{t+1}, D_t))^+ \mid X_t, D_t]$, note that the coupling implies that

$$\begin{aligned} & (-\delta(X_{t+1}, D_t) + \delta(X'_{t+1}, D_t))^+ \\ & = \frac{1}{2} (\|X_{t+1}(D_t) - m(D_t)\mu^*\|_1 - \|X'_{t+1}(D_t) - m(D_t)\mu^*\|_1)^+ \\ & \leq \frac{1}{2} \|X_{t+1}(D_t) - X'_{t+1}(D_t)\|_1 \\ & \leq \frac{1}{2} \|X_{t+1}(D_t \setminus D'_t)\|_1 + \frac{1}{2} \|X'_{t+1}(D_t \setminus D'_t)\|_1 \\ & \leq m(D_t \setminus D'_t), \end{aligned} \quad (4.57)$$

where the last inequality uses the fact that

$$\|X_{t+1}(D_t \setminus D'_t)\|_1 = \|X'_{t+1}(D_t \setminus D'_t)\|_1 = m(D_t \setminus D'_t).$$

Taking the expectation and applying Lemma 4.14, we get

$$\begin{aligned} & \mathbb{E}[(-\delta(X_{t+1}, D_t) + \delta(X'_{t+1}, D_t))^+ \mid X_t, D_t] \\ & \leq \mathbb{E}[m(D_t \setminus D'_t) \mid X_t, D_t] \\ & \leq \frac{1}{\sqrt{N}} + \frac{1}{N}. \end{aligned} \quad (4.58)$$

Next, we bound $\mathbb{E}[(-\delta(X'_{t+1}, D_t))^+ \mid X_t, D_t]$. By the definition of $\delta(X'_{t+1}, D_t)$, we have

$$\begin{aligned} & \mathbb{E}[(-\delta(X'_{t+1}, D_t))^+ \mid X_t, D_t] \\ & = \mathbb{E}\left[\left(\frac{1}{2}\|X'_{t+1}(D_t) - m(D_t)\mu^*\|_1 - \beta(1 - m(D_t))\right)^+ \mid X_t, D_t\right] \quad (4.59) \\ & \leq \frac{1}{2}\mathbb{E}[(\|X'_{t+1}(D_t) - m(D_t)\mu^*\|_1 - \|X_t(D_t) - m(D_t)\mu^*\|_1)^+ \mid X_t, D_t], \end{aligned} \quad (4.60)$$

where (4.60) follows from the facts that $\delta(X_t, D_t) \geq 0$ and $\delta(X_t, D_t) \triangleq \beta(1 - m(D_t)) - \|X_t(D_t) - m(D_t)\mu^*\|_1$. Intuitively, the expression on the right-hand side of (4.60) represents the expected one-step increase in the L_1 distance from the stationary distribution if all arms follow $\bar{\pi}^*$. Because the transition matrix $P_{\bar{\pi}^*}$ is non-expansive, this increase should be small. Indeed, as we formally prove in Lemma 4.26 (Appendix 4.E.3), this expression is bounded as:

$$\mathbb{E}[(\|X'_{t+1}(D_t) - m(D_t)\mu^*\|_1 - \|X_t(D_t) - m(D_t)\mu^*\|_1)^+ \mid X_t, D_t] \leq \frac{2|\mathbb{S}|^{1/2}}{\sqrt{N}}. \quad (4.61)$$

Therefore,

$$\mathbb{E}[(-\delta(X'_{t+1}, D_t))^+ \mid X_t, D_t] \leq \frac{|\mathbb{S}|^{1/2}}{\sqrt{N}}. \quad (4.62)$$

Plugging (4.58) and (4.62) into (4.48), we get

$$\mathbb{E}[(m(D_t) - m(D_{t+1}))^+ \mid X_t, D_t] \leq \frac{|\mathbb{S}|^{1/2} + 1}{\beta\sqrt{N}} + \frac{1 + (\beta + 1)|\mathbb{S}|}{\beta N}. \quad \square$$

4.6.5 Proof of Lemma 4.16

Proof of Lemma 4.16. Recall that D_t is taken to be a maximal set such that $\delta(X_t, D_t) \geq 0$, where $\delta(x, D) = \beta(1 - m(D)) - \|x(D) - m(D)\mu^*\|_1 / 2$. We first prove that

$$\delta(X_t, D_t) \leq \frac{2}{N}.$$

Assume, for the sake of contradiction, that $\delta(X_t, D_t) > 2/N$. Then by the definition of $\delta(X_t, D_t)$, we can see that $m(D_t) < 1$ and thus $D_t^c \neq \emptyset$. Picking an arbitrary $i \in D_t^c$, we have

$$\begin{aligned} & \delta(X_t, D_t \cup \{i\}) - \delta(X_t, D_t) \\ &= -\frac{\beta}{N} - \frac{1}{2} \|X_t(D_t \cup \{i\}) - m(D_t \cup \{i\})\mu^*\|_1 \\ &\quad + \frac{1}{2} \|X_t(D_t) - m(D_t)\mu^*\|_1 \\ &\geq -\frac{\beta}{N} - \frac{1}{2} \|X_t(\{i\}) - m(\{i\})\mu^*\|_1 \\ &\geq -\frac{2}{N}. \end{aligned}$$

Therefore, $\delta(X_t, D_t \cup \{i\}) > 0$, which contradicts the maximality of D_t .

Since $\delta(X_t, D_t) \leq 2/N$, we have

$$1 - m(D_t) \leq \frac{1}{\beta} \|X_t(D_t) - m(D_t)\mu^*\|_1 + \frac{2}{\beta N} \leq \frac{|\mathbb{S}|^{1/2}}{\beta} h_W(X_t, D_t) + \frac{2}{\beta N},$$

where the second inequality is due to the distance dominance property of $h_W(x, D)$ in (4.34). $\square$

4.7 Proof of Theorem 4.3 (Optimality gap of ID policy)

In this section, we prove Theorem 4.3 using the framework established in Section 4.5. This section is organized as follows. We first define the subset Lyapunov functions for the ID policy in Section 4.7.1. We then define the focus set for the ID policy in Section 4.7.2. In Section 4.7.3, we present three lemmas verifying that the ID policy satisfies Conditions 1, 2 and 3,

respectively, and prove Theorem 4.3 by combining these three lemmas and applying the meta-theorem Theorem 4.9. In Sections 4.7.4, 4.7.5, and 4.7.6, we prove the three lemmas.

4.7.1 Subset Lyapunov functions

To construct the subset Lyapunov functions for the ID policy, consider the following class of functions, $\{h_{\text{ID}}(x, [Nm])\}_{m \in [0,1]_N}$: For any system state x and $m \in [0,1]_N$, we let $h_{\text{ID}}(x, [Nm])$ be a non-decreasing “envelop” of $h_W(x, [Nm])$, given by

$$h_{\text{ID}}(x, [Nm]) = \max_{\substack{m' \in [0,1]_N \\ m' \leq m}} h_W(x, [Nm']). \quad (4.63)$$

In the rest of the chapter, we write $h_W(x, m)$ and $h_{\text{ID}}(x, m)$ as shorthands for $h_W(x, [Nm])$ and $h_{\text{ID}}(x, [Nm])$. Note that $h_{\text{ID}}(x, m)$ depends only on the states of the arms in $[Nm]$, as required by the definition of subset Lyapunov functions. The lemma below verifies that $\{h_{\text{ID}}(x, m)\}_{m \in [0,1]_N}$ satisfies Definition 4.8 and is proved in Appendix 4.E.2.

Lemma 4.17. *The class of functions $\{h_{\text{ID}}(\cdot, m)\}_{m \in [0,1]_N}$ defined in (4.63) satisfies that for any system state x and any $m, m' \in [0,1]_N$,*

$$\begin{aligned} & \mathbb{E} \left[\left(h_{\text{ID}}(X_1, m) - \left(1 - \frac{1}{2\lambda_W}\right) h_{\text{ID}}(x, m) \right)^+ \right. \\ & \quad \left. \middle| X_0 = x, A_0(i) \sim \bar{\pi}^*(\cdot | S_0(i)) \forall i \in [Nm] \right] \\ & \leq \frac{4\lambda_W^{1/2}}{\sqrt{N}}, \end{aligned} \quad (4.64)$$

$$h_{\text{ID}}(x, m) \geq \frac{1}{|\mathbb{S}|^{1/2}} \|x([Nm]) - m\mu^*\|_1, \quad (4.65)$$

$$|h_{\text{ID}}(x, m) - h_{\text{ID}}(x, m')| \leq 2\lambda_W^{1/2} |m' - m|. \quad (4.66)$$

These inequalities imply the drift condition, distance dominance property, and Lipschitz continuity in Definition 4.8, respectively. Consequently, $\{h_{\text{ID}}(x, m)\}_{m \in [0,1]_N}$ are subset Lyapunov functions for the optimal single-armed policy $\bar{\pi}^*$.

We note that the inequality (4.64) is stronger than the drift condition required by the definition of subset Lyapunov functions. This stronger version is needed for later analysis.

4.7.2 Focus set

The ID policy, as previously noted, does not explicitly specify focus sets within its algorithm. Nonetheless, for analysis purposes, we can introduce a set D_t at each time step t , effectively serving as the focus set for the ID policy. Specifically, let $D_t = [Nm_d(X_t)]$, where $m_d(\cdot)$ is a function that maps a system state to a number in $[0, 1]_N = \{0, 1/n, \dots, 1\}$, formally defined as follows:

$$m_d(x) = \max\{m \in [0, 1]_N : K_{c/h} h_{\text{ID}}(x, m) \leq \beta(1 - m)\}, \quad (4.67)$$

where $\beta \triangleq \min\{\alpha, 1 - \alpha\}$ and $K_{c/h}$ is a constant. More concretely, the constant $K_{c/h} = \|c_{\bar{\pi}^*}\|_{W^{-1}} = \sqrt{c_{\bar{\pi}^*} W^{-1} c_{\bar{\pi}^*}^\top}$, where $c_{\bar{\pi}^*}$ denotes the row vector $(\bar{\pi}^*(1|s))_{s \in \mathcal{S}}$ and W is the matrix given in Definition 4.11.

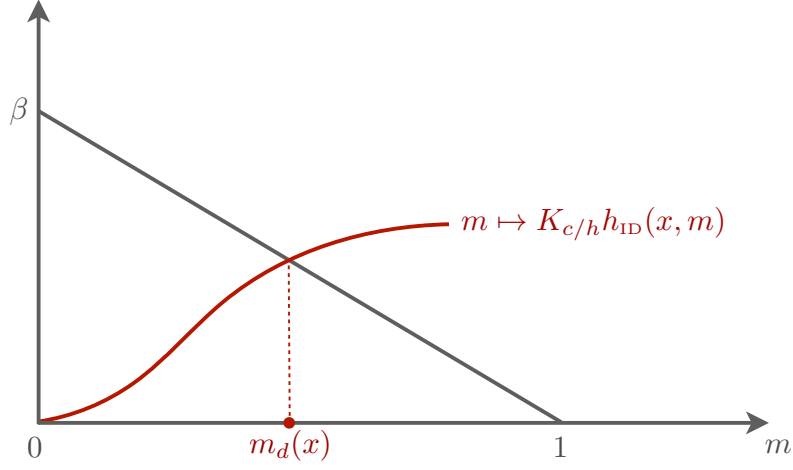
The definition of $m_d(x)$ has a nice geometric representation, as shown in Figure 4.4a. For a system state x , note that $h_{\text{ID}}(x, [0]) = 0$ and recall that $h_{\text{ID}}(x, [Nm])$ is non-decreasing in m . Then $m_d(x)$ is the value of m at which the curve $m \mapsto K_{c/h} h_{\text{ID}}(x, [Nm])$ intersects with the line $m \mapsto \beta(1 - m)$, ignoring the integer effect.

Remark 4.18. Here we comment on the different choices of subset Lyapunov functions in the analysis of the set-expansion policy and the ID policy. In the analysis of the set-expansion policy, $\{h_W(x, D)\}_{D \subseteq [N]}$ is constructed to satisfy Definition 4.8, allowing the application of the meta-theorem Theorem 4.9. In the analysis of the ID policy, $\{h_{\text{ID}}(x, m)\}_{m \in [0, 1]_N}$ is constructed to satisfy Definition 4.8 and to be non-decreasing in m . This monotonicity of $h_{\text{ID}}(x, m)$ ensures that the focus set $D_t = [Nm_d(X_t)]$, defined via $h_{\text{ID}}(x, m)$, can be proved to satisfy the almost non-shrinking condition (Lemma 4.20).

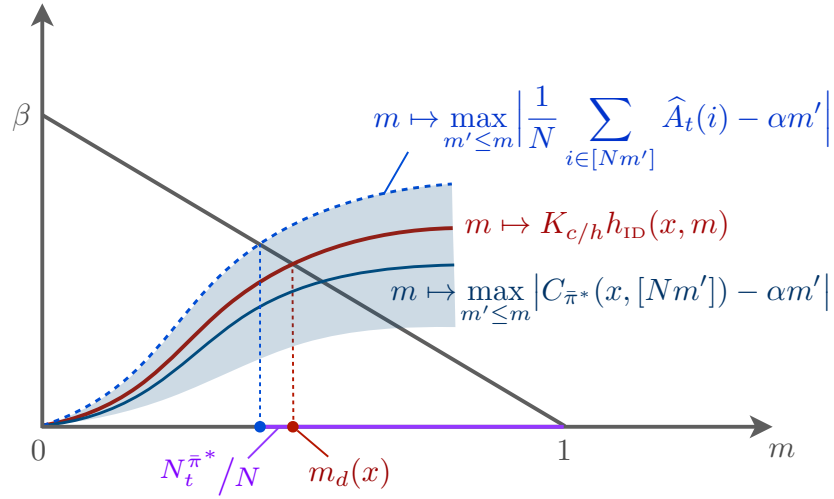
4.7.3 Lemmas for verifying Conditions 1, 2 and 3; proof of Theorem 4.3

Having defined the subset Lyapunov functions $\{h_{\text{ID}}(x, [Nm])\}_{m \in [0, 1]_N}$ and the focus set $D_t = [Nm_d(X_t)]$, we proceed to establish Lemmas 4.19, 4.20 and 4.21, which verify that the ID policy satisfies Conditions 1, 2 and 3, respectively. Then we apply Theorem 4.9 to prove Theorem 4.3.

Lemma 4.19 (ID policy satisfies Condition 1). *Consider the ID policy in Algorithm 5. For any $t \geq 0$, let $D'_t = [\min(N_t^{\bar{\pi}^*}, Nm_d(X_t))]$, where recall that*



(a) Subset Lyapunov functions and the focus set



(b) Illustration of the proof of Lemma 4.19

Figure 4.4: Geometric illustrations of the focus set and the proof of Lemma 4.19 for the ID policy.

$N_t^{\bar{\pi}^*} \in [N]$ is defined in Algorithm 5 as the largest number such that for any $i \in [N_t^{\bar{\pi}^*}]$, $A_t(i) = \hat{A}_t(i)$. Then

$$\begin{aligned} \mathbb{E}[m(D_t \setminus D'_t) \mid X_t] &= \frac{1}{N} \mathbb{E}[(Nm_d(X_t) - N_t^{\bar{\pi}^*})^+ \mid X_t] \\ &\leq \frac{2}{\beta\sqrt{N}} + \frac{1}{N} \quad a.s. \end{aligned} \quad (4.68)$$

Lemma 4.20 (ID policy satisfies Condition 2). *Consider the ID policy in Algorithm 5. For any $t \geq 0$,*

$$\begin{aligned} \mathbb{E}[(m(D_t) - m(D_{t+1}))^+ \mid X_t] &= \mathbb{E}[(m_d(X_t) - m_d(X_{t+1}))^+ \mid X_t] \\ &\leq \frac{4K_{c/h}\lambda_W^{1/2}(1+\beta)}{\beta^2\sqrt{N}} + \frac{2K_{c/h}\lambda_W^{1/2} + \beta}{\beta N} \quad a.s. \end{aligned} \quad (4.69)$$

Lemma 4.21 (ID policy satisfies Condition 3). *Consider the ID policy in Algorithm 5. For any $t \geq 0$,*

$$1 - m(D_t) \leq \frac{K_{c/h}}{\beta} h_{\text{ID}}(X_t, D_t) + \frac{2K_{c/h}\lambda_W^{1/2} + \beta}{\beta N} \quad a.s. \quad (4.70)$$

Proof of Theorem 4.3. By Lemma 4.19, 4.20 and 4.21, the ID policy satisfies Conditions 1, 2 and 3 with the subset Lyapunov functions $\{h_{\text{ID}}(x, [Nm])\}_{m \in [0,1]_N}$. Applying Theorem 4.9 and substituting the constants, we get

$$R^{\text{rel}} - R(\pi, \mathbf{S}_0) \leq \frac{684r_{\max}\lambda_W^{5/2}|\mathbb{S}|^{3/2}}{\beta^3\sqrt{N}},$$

which implies the optimality gap bound in the theorem statement. Note that we bound $K_{c/h}$ by $|\mathbb{S}|^{1/2}$ and relax all $1/N$ factors to $1/\sqrt{N}$ when deriving this bound. $\square$

4.7.4 Proof of Lemma 4.19

Before delving into the proof, we first offer a high-level understanding of Lemma 4.19. Recall that $[N_t^{\bar{\pi}^*}]$ is defined to be the largest set of arms that follow their ideal actions under the ID policy. Then Lemma 4.19 states that the focus set we define, $D_t = [Nm_d(X_t)]$, is close to $[N_t^{\bar{\pi}^*}]$, differing by only $O(\sqrt{N})$ elements. Note that whether a set of arms $[Nm]$ can follow their

ideal actions or not is determined by the amount of budget required by them, i.e., the number of action 1's in their ideal actions. Our proof of Lemma 4.19 utilizes the relationship between the budget requirement by arms in $[Nm]$ and the distributional distance $\|x([Nm]) - m\mu^*\|_W$.

More precisely, conditioning on $X_t = x$, define the scaled conditional expected number of ideal active actions in D by

$$C_{\bar{\pi}^*}(x, D) \triangleq \frac{1}{N} \mathbb{E} \left[\sum_{i \in D} \hat{A}_t(i) \mid X_t = x \right].$$

Figure 4.4b illustrates the relationship between the following three functions of m and the noise term $Z_{t,x}$:

$$b_x(m) \triangleq \max_{\substack{m' \in [0,1]_N \\ m' \leq m}} \left| C_{\bar{\pi}^*}(x, [Nm']) - \alpha m' \right|, \quad (4.71)$$

$$\hat{b}_{t,x}(m) \triangleq \max_{\substack{m' \in [0,1]_N \\ m' \leq m}} \left| \frac{1}{N} \sum_{i \in [Nm']} \hat{A}_t(i) - \alpha m' \right|, \quad (4.72)$$

$$\ell_x(m) \triangleq K_{c/h} h_{\text{ID}}(x, m), \quad (4.73)$$

$$Z_{t,x} \triangleq \max_{m' \in [0,1]_N} \left| \frac{1}{N} \sum_{i \in [Nm']} \hat{A}_t(i) - C_{\bar{\pi}^*}(x, [Nm']) \right|. \quad (4.74)$$

The lower solid curve, dashed blue curve, and red solid curve in Figure 4.4b are b_x , $\hat{b}_{t,x}$, and ℓ_x , respectively; the descending line is $m \mapsto \beta(1 - m)$. The proof shows that $b_x(m) \leq \ell_x(m)$ and $\hat{b}_{t,x}(m) \leq b_x(m) + Z_{t,x}$ for every m , with $\mathbb{E}[Z_{t,x} \mid X_t = x] \leq 2/\sqrt{N}$. Comparing the intersections of these functions with $m \mapsto \beta(1 - m)$ yields the normalized focus-set discrepancy bound in Lemma 4.19:

$$\mathbb{E}[(m_d(X_t) - N_t^{\bar{\pi}^*}/N)^+ \mid X_t] \leq \frac{2}{\beta\sqrt{N}} + \frac{1}{N}.$$

Proof of Lemma 4.19. In this proof, the variables n and n' are by default non-negative integers. We fix a time step $t \geq 0$, and condition on $X_t = x$ for a fixed system state realization x . We first prove a lower bound for $N_t^{\bar{\pi}^*}$ in terms of the *budget requirements* of the arms in $[n]$, $\sum_{i \in [n]} \hat{A}_t(i)$, for each $n \leq N$. Observe that the arms in $[n]$ can follow their ideal actions if and only if

$$\sum_{i \in [n]} \hat{A}_t(i) \leq \alpha N \quad \text{and} \quad \sum_{i \in [n]} (1 - \hat{A}_t(i)) \leq (1 - \alpha)N. \quad (4.75)$$

Here the first condition in (4.75) requires that the number of action 1's is within budget. For the second condition in (4.75), the easiest way to understand it is that it requires the number of action 0's to not exceed $(1 - \alpha)N$, where $(1 - \alpha)N$ can be interpreted as the “budget for passive actions”. As a result, a sufficient condition for the arms in $[n]$ to follow their ideal actions is

$$\left| \sum_{i \in [n]} \hat{A}_t(i) - \alpha n \right| \leq \beta(N - n), \quad (4.76)$$

where recall that $\beta = \min\{\alpha, 1 - \alpha\}$. In this proof, we use a further sufficient condition for the inequality (4.76) above, which is

$$\max_{n' \leq n} \left| \sum_{i \in [n']} \hat{A}_t(i) - \alpha n' \right| \leq \beta(N - n). \quad (4.77)$$

Therefore, by the definition of $N_t^{\bar{\pi}^*}$, we have

$$\begin{aligned} N_t^{\bar{\pi}^*} &\geq \max \left\{ n \leq N : \max_{n' \leq n} \left| \sum_{i \in [n']} \hat{A}_t(i) - \alpha n' \right| \leq \beta(N - n) \right\} \\ &= N \max \left\{ m \in [0, 1]_N : \max_{\substack{m' \in [0, 1]_N \\ m' \leq m}} \left| \frac{1}{N} \sum_{i \in [Nm']} \hat{A}_t(i) - \alpha m' \right| \leq \beta(1 - m) \right\}. \end{aligned} \quad (4.78)$$

Our next step is to upper bound the following quantity:

$$\hat{b}_{t,x}(m) = \max_{m' \in [0, 1]_N, m' \leq m} \left| \frac{1}{N} \sum_{i \in [Nm']} \hat{A}_t(i) - \alpha m' \right|. \quad (4.79)$$

This quantity is defined in (4.72) when we give the intuition of the proof. Bounding $\hat{b}_{t,x}(m)$ is useful because (4.78) lower bounds $N_t^{\bar{\pi}^*}/N$ by the largest $m \in [0, 1]_N$ such that $\hat{b}_{t,x}(m) \leq \beta(1 - m)$. To bound $\hat{b}_{t,x}(m)$, we relate it to the subset Lyapunov function $h_{\text{ID}}(x, m)$ and recall the scaled expected budget requirement $C_{\bar{\pi}^*}(x, D)$:

$$C_{\bar{\pi}^*}(x, D) \triangleq \frac{1}{N} \mathbb{E} \left[\sum_{i \in D} \hat{A}_t(i) \mid X_t = x \right] = \sum_{s \in \mathbb{S}} x(D, s) \bar{\pi}^*(1|s) = x(D) c_{\bar{\pi}^*}^\top, \quad (4.80)$$

where $c_{\bar{\pi}^*}$ is the row vector $(\bar{\pi}^*(1|s))_{s \in \mathbb{S}}$. We first decompose the target quantity using the triangle inequality: for any $m \in [0, 1]_N$, we have

$$\begin{aligned}
\widehat{b}_{t,x}(m) &= \max_{\substack{m' \in [0,1]_N \\ m' \leq m}} \left| \frac{1}{N} \sum_{i \in [Nm']} \widehat{A}_t(i) - \alpha m' \right| \\
&\leq \max_{\substack{m' \in [0,1]_N \\ m' \leq m}} \left(\left| C_{\bar{\pi}^*}(x, [Nm']) - \alpha m' \right| \right. \\
&\quad \left. + \left| \frac{1}{N} \sum_{i \in [Nm']} \widehat{A}_t(i) - C_{\bar{\pi}^*}(x, [Nm']) \right| \right) \\
&\leq \max_{\substack{m' \in [0,1]_N \\ m' \leq m}} \left| C_{\bar{\pi}^*}(x, [Nm']) - \alpha m' \right| \\
&\quad + \max_{m' \in [0,1]_N} \left| \frac{1}{N} \sum_{i \in [Nm']} \widehat{A}_t(i) - C_{\bar{\pi}^*}(x, [Nm']) \right|, \tag{4.81}
\end{aligned}$$

In the notation introduced above, (4.81) states that $\widehat{b}_{t,x}(m) \leq b_x(m) + Z_{t,x}$. We will bound the noise term $Z_{t,x}$ later. We now focus on $b_x(m)$, which captures the deviation of the instantaneous expected budget requirement from its steady-state expectation. Using the fact that $\mu^* c_{\bar{\pi}^*}^\top = \sum_{s \in \mathbb{S}} \mu^*(s) \bar{\pi}^*(1|s) = \alpha$, we have:

$$\begin{aligned}
|C_{\bar{\pi}^*}(x, [Nm']) - \alpha m'| &= |(x([Nm']) - m' \mu^*) c_{\bar{\pi}^*}^\top| \\
&= |(x([Nm']) - m' \mu^*) W^{1/2} W^{-1/2} c_{\bar{\pi}^*}^\top| \\
&\leq \|x([Nm']) - m' \mu^*\|_W \|c_{\bar{\pi}^*}\|_{W^{-1}} \\
&= K_{c/h} h_W(x, m') \tag{4.82}
\end{aligned}$$

$$\max_{\substack{m' \in [0,1]_N \\ m' \leq m}} |C_{\bar{\pi}^*}(x, [Nm']) - \alpha m'| \leq K_{c/h} h_{\text{ID}}(x, m). \tag{4.83}$$

That is, $b_x(m) \leq \ell_x(m)$. Using the monotonicity of $h_{\text{ID}}(x, m)$ in m and the definition of $m_d(x)$, we have that for any $m \leq m_d(x)$,

$$K_{c/h} h_{\text{ID}}(x, m) \leq K_{c/h} h_{\text{ID}}(x, m_d(x)) \leq \beta(1 - m_d(x)), \tag{4.84}$$

i.e., $\ell_x(m) \leq \beta(1 - m_d(x))$ for any $m \leq m_d(x)$. Substituting these bounds back into the decomposition in (4.81) implies that, for any $m \in [0, 1]_N$ such that $m \leq m_d(x)$, we have the following bound:

$$\begin{aligned} \widehat{b}_{t,x}(m) &= \max_{\substack{m' \in [0,1]_N \\ m' \leq m}} \left| \frac{1}{N} \sum_{i \in [Nm']} \widehat{A}_t(i) - \alpha m' \right| \\ &\leq \beta(1 - m_d(x)) + \max_{m' \in [0,1]_N} \left| \frac{1}{N} \sum_{i \in [Nm']} \widehat{A}_t(i) - C_{\bar{\pi}^*}(x, [Nm']) \right|, \end{aligned} \quad (4.85)$$

which is $\widehat{b}_{t,x}(m) \leq \beta(1 - m_d(x)) + Z_{t,x}$.

Next, we derive a lower bound for $N_t^{\bar{\pi}^*}$ in terms of $Z_{t,x}$. By (4.78) and (4.85), $N_t^{\bar{\pi}^*}/N$ is no smaller than any $\bar{m} \in [0, 1]_N$ satisfying two conditions: (1) $\bar{m} \leq m_d(x)$, and (2) $\beta(1 - m_d(x)) + Z_{t,x} \leq \beta(1 - \bar{m})$. First, suppose that $m_d(x) - Z_{t,x}/\beta \geq 0$, then there exists $\bar{m}$ satisfying these two conditions. Taking the largest $\bar{m}$ satisfying these two conditions yields

$$\begin{aligned} N_t^{\bar{\pi}^*} &\geq \min \left\{ Nm_d(x), \left\lfloor N - \frac{N}{\beta} \left(\beta(1 - m_d(x)) + Z_{t,x} \right) \right\rfloor \right\} \\ &\geq \min \left\{ Nm_d(x), N - \frac{N}{\beta} \left(\beta(1 - m_d(x)) + Z_{t,x} \right) - 1 \right\} \\ &= \min \left\{ Nm_d(x), Nm_d(x) - 1 - \frac{N}{\beta} Z_{t,x} \right\} \\ &= Nm_d(x) - 1 - \frac{N}{\beta} Z_{t,x}. \end{aligned} \quad (4.86)$$

Rearranging the terms in (4.86) gives

$$(Nm_d(x) - N_t^{\bar{\pi}^*})^+ \leq 1 + \frac{N}{\beta} Z_{t,x}. \quad (4.87)$$

If $m_d(x) - Z_{t,x}/\beta < 0$, then $N_t^{\bar{\pi}^*} \geq 0$ implies

$$(Nm_d(x) - N_t^{\bar{\pi}^*})^+ \leq Nm_d(x) < \frac{N}{\beta} Z_{t,x},$$

so (4.87) also holds in this case. Therefore, taking the conditional expectation in (4.87), we obtain

$$\mathbb{E} \left[(Nm_d(x) - N_t^{\bar{\pi}^*})^+ \mid X_t = x \right] \leq 1 + \frac{N}{\beta} \mathbb{E} [Z_{t,x} \mid X_t = x].$$

It remains to prove $\mathbb{E}[Z_{t,x} \mid X_t = x] \leq 2/\sqrt{N}$. By the definition of $Z_{t,x}$ in (4.74), this is equivalent to proving

$$\mathbb{E}\left[\max_{n \leq N} \left| \sum_{i \in [n]} \hat{A}_t(i) - NC_{\bar{\pi}^*}(x, [n]) \right| \mid X_t = x\right] \leq 2\sqrt{N}. \quad (4.88)$$

Our proof uses Doob's maximum inequality. First, we simplify the expression by defining the centered random variables $\xi(i) = \hat{A}_t(i) - \mathbb{E}[\hat{A}_t(i) \mid X_t = x]$. This allows us to rewrite the left-hand side of (4.88) as:

$$\mathbb{E}\left[\max_{n \leq N} \left| \sum_{i \in [n]} \hat{A}_t(i) - NC_{\bar{\pi}^*}(x, [n]) \right| \mid X_t = x\right] = \mathbb{E}\left[\max_{n \leq N} \left| \sum_{i \in [n]} \xi(i) \right| \mid X_t = x\right]. \quad (4.89)$$

The sequence of partial sums $(\sum_{i \in [n]} \xi(i))_{n \in [N]}$ forms a martingale when conditioned on $X_t = x$. This holds because the ideal actions $\hat{A}_t(i)$ are sampled independently, making the $\xi(i)$ terms independent random variables, each with a conditional expectation of zero. Applying Cauchy-Schwarz and Doob's L_2 maximum inequality [see, e.g., 86, Theorem 4.4.4], we bound this term as:

$$\begin{aligned} \mathbb{E}\left[\max_{n \leq N} \left| \sum_{i \in [n]} \xi(i) \right| \mid X_t = x\right] &\leq \mathbb{E}\left[\max_{n \leq N} \left| \sum_{i \in [n]} \xi(i) \right|^2 \mid X_t = x\right]^{1/2} \\ &\leq \left(4\mathbb{E}\left[\left| \sum_{i \in [N]} \xi(i) \right|^2 \mid X_t = x\right]\right)^{1/2} \\ &= \left(4 \sum_{i \in [N]} \mathbb{E}[\xi(i)^2 \mid X_t = x]\right)^{1/2} \\ &\leq 2\sqrt{N}. \end{aligned}$$

Here, the first inequality follows from Cauchy-Schwarz, the second applies Doob's maximal L_2 inequality, and the equality follows from the independence of the $\xi(i)$ terms. The final inequality holds because $|\xi(i)| = |\hat{A}_t(i) - \mathbb{E}[\hat{A}_t(i) \mid X_t = x]| \leq 1$. This completes the proof. $\square$

4.7.5 Proof of Lemma 4.20

Next, we prove that the focus set that we choose for the ID policy, $D_t = [Nm_d(X_t)]$, satisfies the almost non-shrinking condition (Condition 2). The

key intuition lies in the geometric interpretation shown in Figure 4.4a, where $m_d(X_t)$ is the intersection point of two functions: the non-decreasing curve $m \mapsto h_{\text{ID}}(X_t, m)$ and the strictly decreasing line $m \mapsto \beta(1 - m)$. Our proof shows that the vertical shift of the curve $m \mapsto h_{\text{ID}}(X_t, m)$ from one time step to the next is small (in particular, $h_{\text{ID}}(X_{t+1}, m_d(X_t)) - h_{\text{ID}}(X_t, m_d(X_t))$ is small). Because the non-decreasing curve's vertical movement is small as it intersects the strictly decreasing line, the intersection point $m_d(X_t)$ cannot shift significantly to the left. This implies that the focus set $D_t = [Nm_d(X_t)]$ is almost non-shrinking.

Proof of Lemma 4.20. Fixing a time step $t \geq 0$, we first prove the following inequality, which will be used to establish an upper bound on $\mathbb{E}[(m_d(X_t) - m_d(X_{t+1}))^+ \mid X_t]$:

$$m_d(X_{t+1}) \geq m_d(X_t) - \frac{K_{c/h}}{\beta} (h_{\text{ID}}(X_{t+1}, m_d(X_t)) - h_{\text{ID}}(X_t, m_d(X_t)))^+ - \frac{1}{N}. \quad (4.90)$$

By the maximality of $m_d(X_{t+1})$, it suffices to show $K_{c/h} h_{\text{ID}}(X_{t+1}, \bar{m}) \leq \beta(1 - \bar{m})$ for each $\bar{m} \in [0, 1]_N$ with $\bar{m} \leq m_d(X_t) - \frac{K_{c/h}}{\beta} (h_{\text{ID}}(X_{t+1}, m_d(X_t)) - h_{\text{ID}}(X_t, m_d(X_t)))^+$. For any such $\bar{m}$,

$$\begin{aligned} \beta(1 - \bar{m}) &\geq \beta(1 - m_d(X_t)) \\ &\quad + K_{c/h} (h_{\text{ID}}(X_{t+1}, m_d(X_t)) - h_{\text{ID}}(X_t, m_d(X_t)))^+ \\ &\geq K_{c/h} h_{\text{ID}}(X_t, m_d(X_t)) \\ &\quad + K_{c/h} (h_{\text{ID}}(X_{t+1}, m_d(X_t)) - h_{\text{ID}}(X_t, m_d(X_t)))^+ \\ &\geq K_{c/h} h_{\text{ID}}(X_{t+1}, m_d(X_t)) \\ &\geq K_{c/h} h_{\text{ID}}(X_{t+1}, \bar{m}), \end{aligned}$$

where the second inequality is because $K_{c/h} h_{\text{ID}}(X_t, m_d(X_t)) \leq \beta(1 - m_d(X_t))$, and the last inequality is because $h_{\text{ID}}(X_t, m)$ is *non-decreasing* in m and $\bar{m} \leq m_d(X_t)$. This proves (4.90).

The inequality (4.90) implies that

$$\begin{aligned} &\mathbb{E}[(m_d(X_t) - m_d(X_{t+1}))^+ \mid X_t] \\ &\leq \frac{K_{c/h}}{\beta} \mathbb{E}[(h_{\text{ID}}(X_{t+1}, m_d(X_t)) - h_{\text{ID}}(X_t, m_d(X_t)))^+ \mid X_t] + \frac{1}{N}. \quad (4.91) \end{aligned}$$

We now upper bound $\mathbb{E}[(h_{\text{ID}}(X_{t+1}, m_d(X_t)) - h_{\text{ID}}(x, m_d(x)))^+ \mid X_t]$ by coupling X_{t+1} with a random element X'_{t+1} constructed below. Let X'_{t+1} be the random element denoting the system state at time step $t + 1$ if we were able to set $A_t(i) = \hat{A}_t(i)$ for all $i \in [N]$. By the drift condition of the subset Lyapunov function $h_{\text{ID}}(\cdot, D)$ established as (4.64) in Lemma 4.17,

$$\begin{aligned} & \mathbb{E}[(h_{\text{ID}}(X'_{t+1}, m_d(X_t)) - h_{\text{ID}}(X_t, m_d(X_t)))^+ \mid X_t] \\ & \leq \mathbb{E}\left[\left(h_{\text{ID}}(X'_{t+1}, m_d(X_t)) - \left(1 - \frac{1}{2\lambda_W}\right)h_{\text{ID}}(X_t, m_d(X_t))\right)^+ \mid X_t\right] \\ & \leq \frac{4\lambda_W^{1/2}}{\sqrt{N}}. \end{aligned} \quad (4.92)$$

We couple X'_{t+1} and X_{t+1} such that $X'_{t+1}(\{i\}) = X_{t+1}(\{i\})$ for all $i \leq \min(Nm_d(X_t), N_t^{\bar{\pi}^*})$. Then

$$\begin{aligned} & \mathbb{E}\left[(h_{\text{ID}}(X_{t+1}, m_d(X_t)) - h_{\text{ID}}(X_t, m_d(X_t)))^+ \right. \\ & \quad \left. - (h_{\text{ID}}(X'_{t+1}, m_d(X_t)) - h_{\text{ID}}(X_t, m_d(X_t)))^+ \mid X_t\right] \\ & \leq \mathbb{E}\left[(h_{\text{ID}}(X_{t+1}, m_d(X_t)) - h_{\text{ID}}(X'_{t+1}, m_d(X_t)))^+ \mid X_t\right] \\ & = \mathbb{E}\left[\left(\max_{m' \in [0,1]_N, m' \leq m_d(X_t)} h_W(X_{t+1}, m') \right. \right. \\ & \quad \left. \left. - \max_{m' \in [0,1]_N, m' \leq m_d(X_t)} h_W(X'_{t+1}, m')\right)^+ \mid X_t\right] \\ & \leq \mathbb{E}\left[\max_{m' \in [0,1]_N, m' \leq m_d(X_t)} (h_W(X_{t+1}, m') - h_W(X'_{t+1}, m'))^+ \mid X_t\right] \\ & \leq \mathbb{E}\left[\max_{m' \in [0,1]_N, m' \leq m_d(X_t)} \|X_{t+1}([Nm']) - X'_{t+1}([Nm'])\|_W \mid X_t\right] \\ & \leq \mathbb{E}\left[\max_{m' \in [0,1]_N, m' \leq m_d(X_t)} \left(\|X_{t+1}([Nm'] \setminus [N_t^{\bar{\pi}^*}])\|_W \right. \right. \\ & \quad \left. \left. + \|X'_{t+1}([Nm'] \setminus [N_t^{\bar{\pi}^*}])\|_W\right) \mid X_t\right] \\ & \leq \frac{2\lambda_W^{1/2}}{N} \mathbb{E}\left[\max_{m' \in [0,1]_N, m' \leq m_d(X_t)} (Nm' - N_t^{\bar{\pi}^*})^+ \mid X_t\right] \end{aligned} \quad (4.93)$$

$$\begin{aligned} & \leq \frac{2\lambda_W^{1/2}}{N} \mathbb{E}\left[(Nm_d(X_t) - N_t^{\bar{\pi}^*})^+ \mid X_t\right] \\ & \leq \frac{4\lambda_W^{1/2}}{\beta\sqrt{N}} + \frac{2\lambda_W^{1/2}}{N}, \end{aligned} \quad (4.94)$$

where (4.93) follows from the facts $\|v\|_W \leq \lambda_W^{1/2} \|v\|_1$ for any vector v and that $\|X_{t+1}(D)\|_1 = \|X'_{t+1}(D)\|_1 = m(D)$ for any $D \subseteq [N]$, and (4.94) applies the bound on $\mathbb{E}[(Nm_d(X_t) - N_t^{\pi^*})^+ | X_t]$ in Lemma 4.19.

Combining (4.91), (4.92) and (4.94), we get

$$\mathbb{E}[(m_d(X_t) - m_d(X_{t+1}))^+ | X_t] \leq \frac{4K_{c/h}\lambda_W^{1/2}(1+\beta)}{\beta^2\sqrt{N}} + \frac{2K_{c/h}\lambda_W^{1/2} + \beta}{\beta N}. \quad \square$$

4.7.6 Proof of Lemma 4.21

Proof of Lemma 4.21. This lemma almost directly follows from the definition of the focus set $D_t \triangleq [Nm_d(X_t)]$, where

$$m_d(X_t) \triangleq \max\{m \in [0, 1]_N : K_{c/h}h_{\text{ID}}(X_t, m) \leq \beta(1 - m)\}. \quad (4.95)$$

The main technical step is to account for the discretization effect, i.e., $m_d(X_t)$ must be a multiple of $1/N$.

We focus on the non-trivial case where $m_d(X_t) < 1$. By the maximality in the definition (4.95), we have

$$K_{c/h}h_{\text{ID}}\left(X_t, m_d(X_t) + \frac{1}{N}\right) > \beta\left(1 - m_d(X_t) - \frac{1}{N}\right). \quad (4.96)$$

Next, by the Lipschitz continuity of $h_{\text{ID}}(x, m)$ stated in (4.66), we have

$$K_{c/h}h_{\text{ID}}\left(X_t, m_d(X_t) + \frac{1}{N}\right) \leq K_{c/h}h_{\text{ID}}\left(X_t, m_d(X_t)\right) + \frac{2K_{c/h}\lambda_W^{1/2}}{N}. \quad (4.97)$$

Combining (4.96) with (4.97), we get

$$\beta(1 - m_d(X_t)) < K_{c/h}h_{\text{ID}}(x, m_d(X_t)) + \frac{2K_{c/h}\lambda_W^{1/2} + \beta}{N}. \quad \square$$

4.8 Experiments

In our theoretical analysis, we have shown that our policies achieve asymptotic optimality assuming only the aperiodic unichain assumption, removing GAP or SA assumed in prior work. In this section, we compare the numerical performance of our policies with the policies in the prior work when the number of arms N is finite. These numerical results complement our theory,

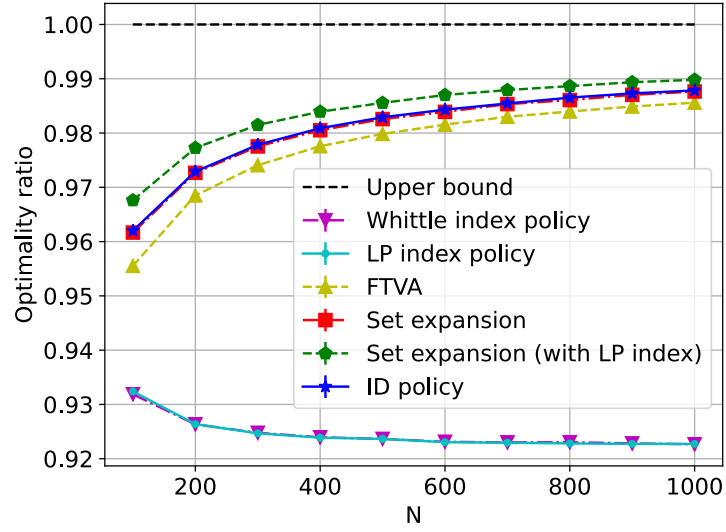
showing that our policies also empirically outperform previous policies on some RB problems that violate GAP or SA (Section 4.8.1 and 4.8.3) but still satisfy our Assumption 4.1. Moreover, such RB problems are not rare. For GAP, there exists some natural classes of randomly-generated RB instances, a decent fraction of which do not satisfy GAP under all LP-Priority policies (Section 4.8.2); for SA, we give two counterexamples for SA and discuss ways to construct more counterexamples in Appendix 4.A. The code for all the experiments are available on Github [152].

4.8.1 Comparing policies on two non-GAP examples

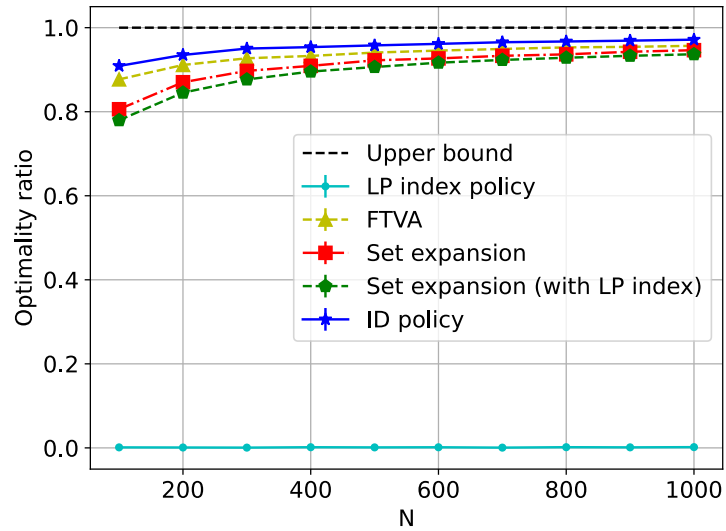
In this section, we consider two examples where two prevalent versions of LP-Priority policies, the Whittle index policy [283] and the LP index policy [108] (whose variants are also studied as primal-dual heuristics, Lagrange-based policies, or the Optimal Lagrangian Index Policy [32, 56, 142, 155]) are not asymptotically optimal, either because of the failure of GAP or because the policy itself is not well-defined. We will simulate the Whittle index policy and the LP-index policy on these two examples, where the Whittle index policy is implemented using the algorithm in [106]. Along with these two LP-Priority policies, we also evaluate the performance of the FTVA policy (Chapter 3 [149]), the set-expansion policy (Section 4.4.2), and the ID policy (Section 4.4.3). Note that for the set-expansion policy, we will consider two versions of implementations that perform action rectification differently: the vanilla version performs action rectification in the uniformly random way as described in Algorithm 4; the “LP index version” of the set-expansion policy applies the LP index policy to the arms not in the focus set. See Appendix 4.H.1 for implementation details of the set-expansion policy.

For all the simulations, we compare the *optimality ratios* of the policies, which are their average rewards normalized by the optimal value of the LP relaxation in (LP). Note that the optimality ratio of an asymptotically optimal policy converges to 1 as $N \rightarrow \infty$. Each optimality ratio is estimated from multiple replications. For each replication, the initial state of each arm is independently sampled from the uniform distribution over the state space. Details of the simulation setting can be found in Appendix 4.H.2.

The first non-GAP example is an RB problem defined by a single-armed MDP with three states and was obtained in a random search by [105]; it was also evaluated in Figure 1 of [149]; see Appendix 4.H.2 for its detailed definition. Note that in this three-state example, there is only one LP-Priority



(a) Performance comparison on the three-state example



(b) Performance comparison on the eight-state example

Figure 4.5: Performance comparison on two examples where GAP fails.

policy, so the Whittle index policy and the LP index policy are identical. Our simulation is shown in Figure 4.5a: The Whittle index policy and the LP index policy are asymptotically suboptimal; FTVA outperforms the LP index policy and appears to be asymptotically optimal; the set-expansion policy and the ID policy are strictly better than FTVA; the LP-index version of set-expansion policy has the best performance among all these policies.

The second non-GAP example is defined by a single-armed MDP with eight states, and is adapted from Figure 2 of [149]; see Appendix 4.H.2 for its detailed definition. A notable feature of this example is the existence of a local attractor, where the scaled state-count vector of the arms is attracted to a distribution other than the optimal stationary distribution, which is a mode of non-GAP-ness not observed in earlier literature. Our simulation result on this example is shown in Figure 4.5b: The Whittle index policy is not included since this example is non-indexable; the LP index policy has nearly zero reward; FTVA, the set-expansion policy, and the ID policy are asymptotically optimal, and the ID policy has the best performance among all these policies.

4.8.2 Commonness of non-GAP examples

Although counterexamples to GAP are well-known to exist starting from [274], such examples are rare in previously known classes of RB problems. In particular, it has been found in [107] that when $|\mathbb{S}| = 3$, no more than 0.2% of the uniformly random examples are non-indexable or violate GAP for Whittle index policy; when $|\mathbb{S}|$ gets large, this fraction among the uniformly random examples further decreases and becomes less than $10^{-4}\%$ when $|\mathbb{S}| = 7$.

In this section, we study some classes of RB problems that are *sparser* than the RB problems following the uniform distribution, i.e., RB problems with fewer non-zero entries in the transition kernels. Specifically, we generate random RB instances with $|\mathbb{S}| = 10$, where the transition distributions $P(s, a, \cdot)$ for $s \in \mathbb{S}, a \in \mathbb{A}$ are independently sampled from symmetric Dirichlet distributions, a natural class of distributions for generating points on the probability simplex. The reward functions $r(\cdot, a)$ for $a \in \mathbb{A}$ are sampled from the same Dirichlet distribution. We have also experimented with uniformly distributed reward functions and observed similar results, as discussed in Appendix 4.H.4. Note that all entries less than 10^{-7} are rounded to zero.

To count the number of non-GAP examples, we focus on identifying the RB problems that are *locally unstable*, which implies the violation of GAP

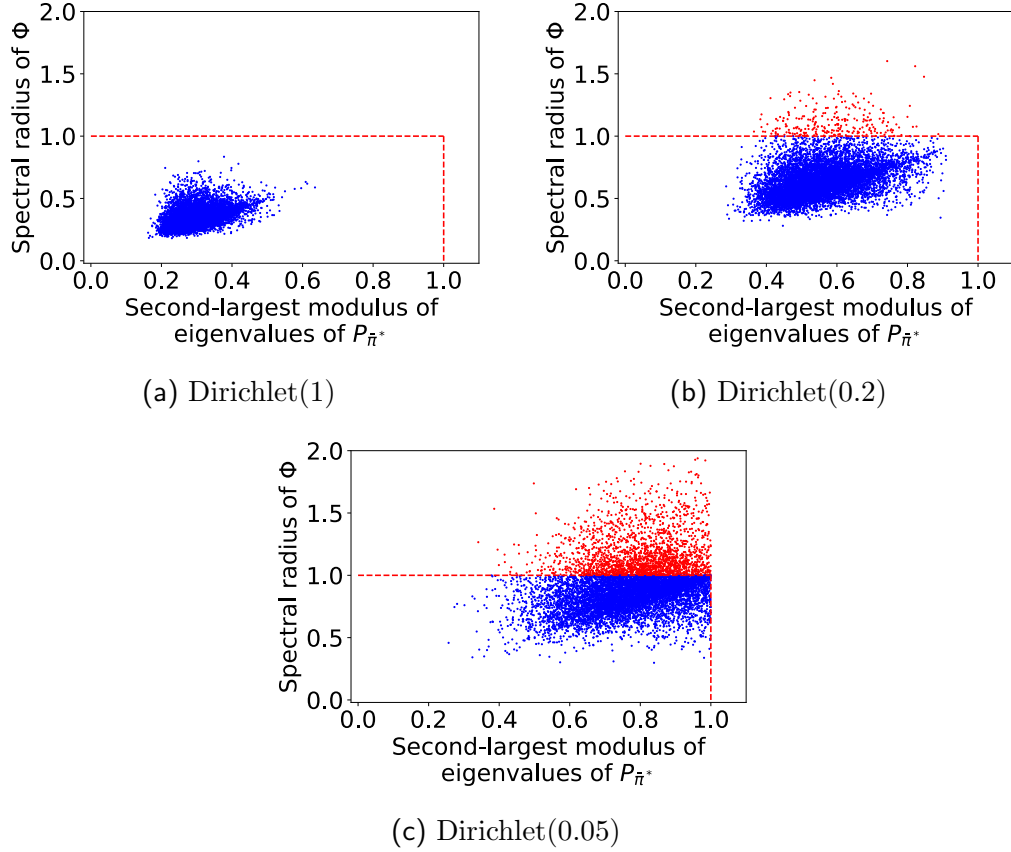


Figure 4.6: Scatter plots illustrating the eigenvalues of about 10^4 random RB problems, whose transition probabilities and reward function follow the symmetric Dirichlet distribution with different parameters in each subplot. Each point in a scatter plot represents an RB problem, whose x -coordinate (or y -coordinate) represents the second largest modulus of eigenvalues of $P_{\tilde{\pi}^*}$ (or spectral radius of Φ). Each RB problem has $|\mathbb{S}| = 10$. The points outside the dashed box (highlighted in red) represent RB problems that violate GAP under all LP-Priority policies.

under all LP-Priority policies. The local instability of an RB problem is easy to certify: it happens when the spectral radius of a certain matrix Φ representing the local mean-field dynamics under the LP-Priority policies is larger than 1, when the definition of Φ is unambiguous (See Appendix 4.H.3 for details). Based on this fact, we make three scatter plots in Figure 4.6, each visualizing 10^4 independently generated RB problems following the symmetric Dirichlet distribution with different parameters. Each point in a scatter plot represents an RB problem, whose y -coordinate is the spectral radius of Φ , and whose x -coordinate represents the second-largest absolute value of P_{π^*} 's eigenvalues.

In Figures 4.6a to 4.6c, the parameter of the symmetric Dirichlet distribution decreases as 1, 0.2, and 0.05, indicating the increased sparsity of the single-armed MDPs. In particular, the average fractions of non-zero entries of the transition distributions $P(s, a, \cdot)$ are approximately 100%, 95.3%, and 58.7%, respectively. In Figure 4.6a, because Dirichlet(1) is the uniform distribution on probability simplex, the fact that no RB examples are found to be locally unstable is consistent with the findings in [108]. In contrast, in Figure 4.6c, under Dirichlet(0.05) distribution, a significant proportion of the problem instances are locally unstable (marked in red); moreover, even if we focus on the examples whose x -coordinate is less than 0.95, that is, the examples whose P_{π^*} is aperiodic unichain with a decently large spectral gap, 1844 out of 9131 examples (about 20.2%) are locally unstable. Note that when generating the 10^4 random examples for each of the three scatter plots, the definition of Φ is ambiguous for only a small number of examples, which we do not display in the figures; specifically, the actual number of examples displayed in Figures 4.6a to 4.6c are 9776, 9918, and 9914, respectively.

The experiment in Figure 4.6 shows that for the RB problems whose single-armed MDPs are sparse, a significant fraction of them could be aperiodic unichain, but violate GAP for all LP-Priority policies.

LP-Priority policies on random non-GAP examples Violation of GAP invalidates the asymptotic optimality guarantee of an LP-Priority policy, but how suboptimal is an LP-Priority policy when GAP does not hold? In Figure 4.7, we plot some cumulative distribution function (CDF) curves representing the optimality ratios of the LP index policy, the Whittle index policy, and their maximal performance when $N = 500$, among 2034 locally unstable examples generated from the Dirichlet(0.05) distribution.

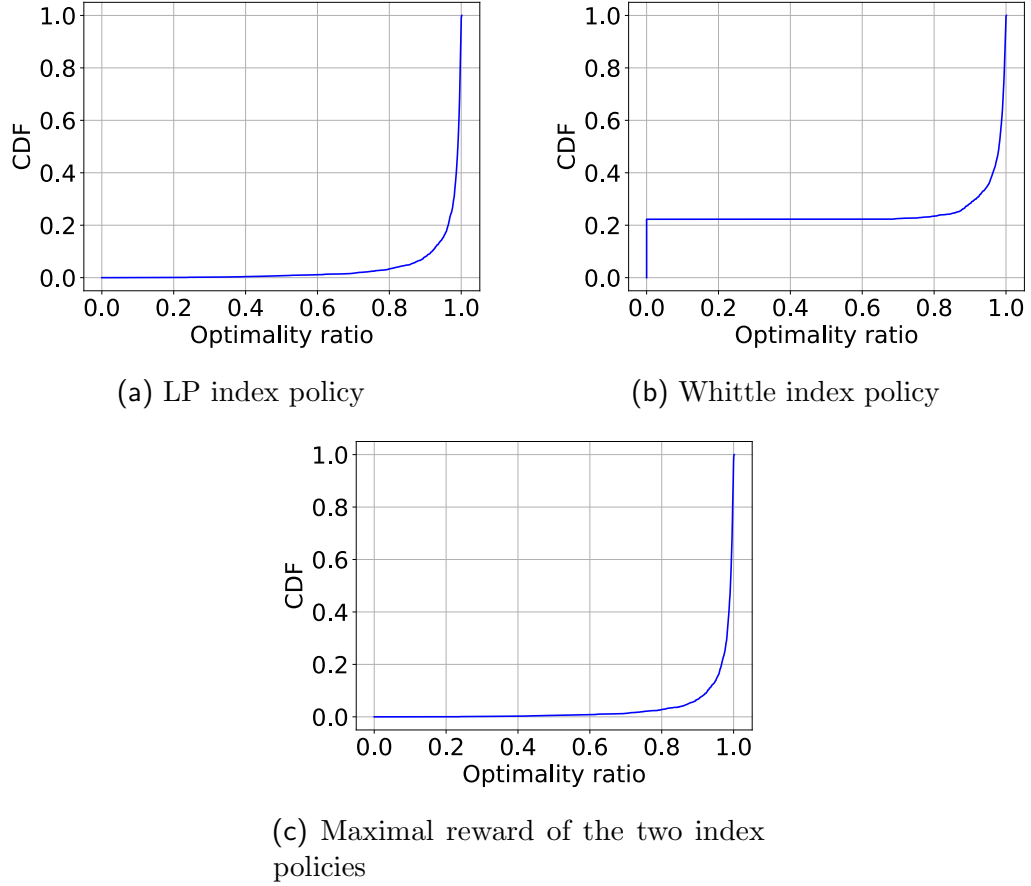


Figure 4.7: CDF of the optimality ratios of LP-Priority policies when $N = 500$, among 2049 non-GAP examples generated from $\text{Dirichlet}(0.05)$. We regard the average reward of Whittle index policy as 0 if it is not well-defined.

Each policy is simulated for 2×10^4 time steps on every example. As we can see from Figure 4.7c, the average rewards under the LP index policy or the Whittle index policy are close to the LP upper bound in most non-GAP examples, which explains the good performance of LP-Priority policies observed in practice; on the other hand, there are about 6.7% of the examples where the average rewards of both policies are less than 90% of the LP upper bound. This experiment shows that it is not uncommon for LP-Priority policies to be substantially suboptimal when the single-armed MDPs of the

RB problems are sparse. In these cases, relaxing the GAP condition could bring a practical benefit.

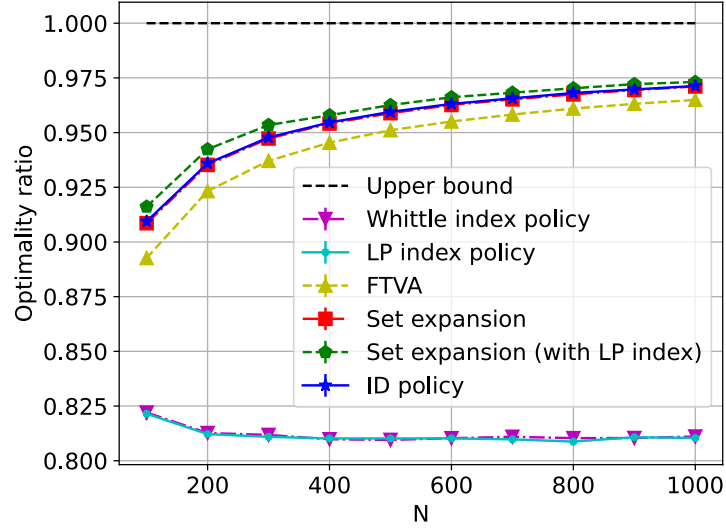
In Figure 4.8a and Figure 4.8b, we pick two non-GAP examples where both the Whittle index and LP index policies have optimality ratios of less than 90%, and compare the performance of different policies there. In both examples, FTVA, the two versions of set-expansion policy, and the ID policy outperform the LP-Priority policies, with clear and discernible differences.

4.8.3 Comparing policies on two non-SA examples

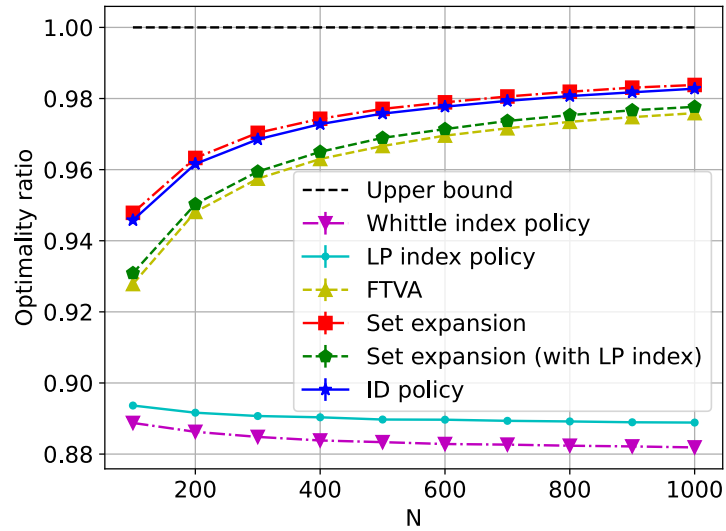
In this section, we consider the same set of policies as in Section 4.8.1 and compare their performances on two examples that violate SA (Assumption 1 of [149]) but satisfy our aperiodic unichain assumption. We give one of the examples at the end of this subsection, and another one in Appendix 4.A, where we also discuss ways to construct more counterexamples to SA. On a high level, in each example that we construct, an arm needs to strictly follow a particular policy to reach and remain in the states with high rewards, which is hard to achieve by following some virtual actions that are not generated based on the true state of the arm.

The simulation results are shown in Figure 4.9. Note that in each replication, the initial state of each arm is independently sampled from the uniform distribution over the state space; FTVA is simulated for 1.6×10^5 time steps with five replications, whereas the other policies are simulated for 2×10^4 time steps with five replications. Despite the longer simulations, the performances of FTVA still exhibit significant variability with large confidence intervals. In both examples, FTVA perform worse than the other policies, especially in the RB problem considered in Figure 4.9b, whose single-armed MDP has a larger state space. In contrast, the ID policy and the two version of the set-expansion policy demonstrate solid performances, though not quite reaching the performances of the LP index policy. The Whittle index policy is not well-defined on these two examples due to the multichain nature of the single-armed MDPs (see [106] for details).

Definition of one of the non-SA examples. Now we define the example considered in Figure 4.9a. The single-armed MDP of this example is defined using Figure 4.10. This figure consists of a set of circles denoting states and a set of arrows in solid lines and dashed lines denoting possible transitions under different actions. The states are indexed by $0, 1, 2, \dots, 7$. Each solid

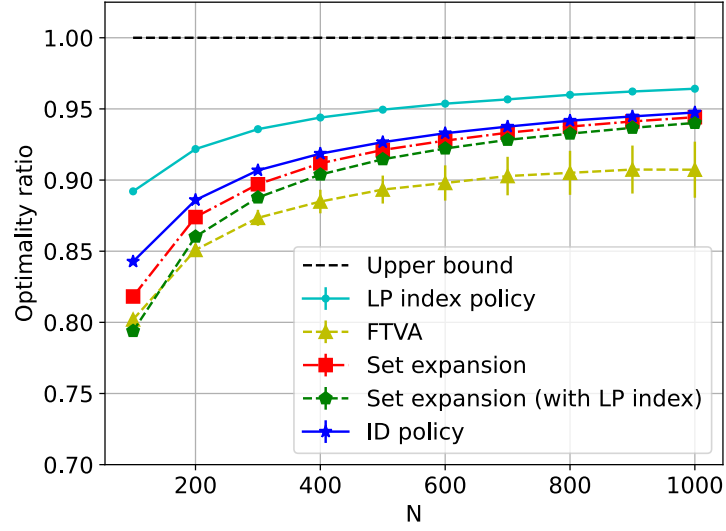


(a) First Dirichlet(0.05) example

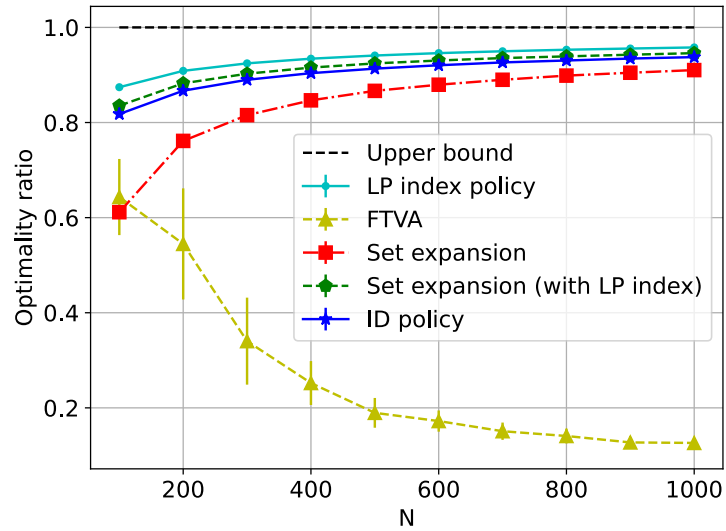


(b) Second Dirichlet(0.05) example

Figure 4.8: Performance comparison on two Dirichlet(0.05) examples where GAP fails to hold.



(a) RB problem defined in Figure 4.10



(b) RB problem defined in Figure 4.11

Figure 4.9: Performance comparison of the policies on counterexamples to SA defined in Appendix 4.A.

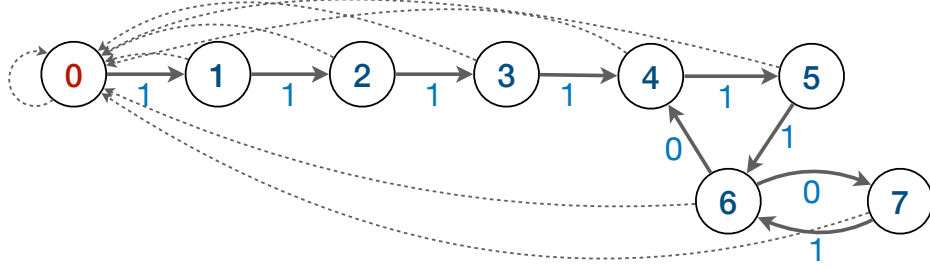


Figure 4.10: A counterexample to the Synchronization Assumption in [149].

Each circle denotes a state, indexed by $0, 1, 2, \dots, 7$. Each arrow denotes a possible transition. The numbers labeled on the solid-line arrows denote actions. If an arm takes an action that is labeled on one of the outward solid-line arrows at its current state, it picks such an arrow labeled by the action uniformly at random and transitions to a nearby state along the arrow; otherwise, the arm jumps to state 0. The reward is 1 if an arm is in states $\{4, 5, 6, 7\}$ and takes the action on an outward solid-line arrow at its current state. Otherwise, the reward is zero. The budget parameter α is set to be 0.6.

arrow is labeled by an action, which is 0 or 1. In each time step, when the arm takes an action labeled on one of the outward solid-line arrows adjacent to its current state, the arm transitions to a random nearby state through one of such arrows. When the arm takes an action that does not exist on any of the adjacent outward solid-line arrows, it jumps to state 0. For example, if the arm takes action 1 at state 7, it goes to state 6 with probability 1; if the arm takes action 0 at state 6, it goes to state 7 or 4 with equal probabilities; if the arm takes action 0 at state 2, it jumps to state 0 with probability 1. For the reward functions, one unit of reward is generated if the arm is in states $\{4, 5, 6, 7\}$ and takes the action on an adjacent outward solid-line arrow; no reward is generated otherwise. We let $\alpha = 0.6$, that is, the arm is activated for 0.6 fraction of the time in the long run.

One can verify that the optimal single-armed policy $\bar{\pi}^*$ takes each action with probability 0.5 in the states $\{0, 1, 2, 3\}$, and always takes the actions labeled on the solid line arrows in the states $\{4, 5, 6, 7\}$. Thus, $\bar{\pi}^*$ induces an aperiodic unichain with the recurrent class $\{4, 5, 6, 7\}$. The long-run average reward of $\bar{\pi}^*$ is 1.

On the other hand, we argue that SA is violated in this example. To see this, recall the leader-and-follower system in the SA (see Section 4.1 in

[149]), which consists of two arms, the leader arm and the follower arm, whose states are denoted as $\hat{S}_t$ and S_t ; the leader arm takes the action $\hat{A}_t \sim \bar{\pi}^*(\cdot|\hat{S}_t)$, and the follower arm takes the same action, $A_t = \hat{A}_t$. SA requires that the stopping time $\tau \triangleq \inf\{t: S_t = \hat{S}_t\}$ has a finite expectation for any possible pair of initial states. However, in the above example, if we initialize the pair of states as $\hat{S}_0 = 7$ and $S_0 = 0$, $\hat{S}_t$ will remain in states $\{4, 5, 6, 7\}$ under $\bar{\pi}^*$, and the action sequences applied by both arms will not contain more than two subsequent 1's. Consequently, S_t always falls back to the state 0 before reaching state 3, so the two arms never reach the same state, implying $\tau = \infty$.

4.9 Conclusion and discussions

In this chapter, we consider the infinite-horizon, average-reward restless bandit problem. We introduce a new class of policies that are asymptotically optimal with $O(1/\sqrt{N})$ optimality gaps, if the optimal single-armed policy induces an aperiodic unichain. This work is the first to show that asymptotic optimality can be achieved without any additional assumptions like GAP and SA.

Our policy design and analysis highlight the use of multiple, bivariate Lyapunov functions. This novel approach holds promise beyond restless bandits, showing potential for a broader class of large stochastic systems consisting of many coupled components. In such complex systems, it can be challenging to directly design a policy that steers the whole system towards optimality or to construct a Lyapunov function that certifies such convergence.

To complement our theory, we simulate our policies on examples where either GAP or SA fails, along with the policies from prior work. Our policies consistently demonstrate good performance, whereas the policies from the prior work may perform suboptimally in some examples. Additionally, we identify some natural classes of RB instances where GAP is violated with considerable probabilities and discuss a method for constructing more counterexamples of SA.

There are several interesting directions for future work. Some directions have been largely addressed by the follow-up papers reviewed in Section 4.2. These directions include multiple actions, heterogeneous arms, and exponential optimality gaps without the global attractor property. However, many interesting open problems remain. For example, an interesting open direction is to improve the dependency of the optimality gap on the problem parameters other than N , such as the spectral gap of the single-armed system. Another

open direction is exploring restless bandits (or weakly-coupled MDPs) with an infinite state space, which could naturally arise in various applications. The third open direction is learning to optimize restless bandits with unknown parameters. Finally, our bivariate Lyapunov function technique (Section 4.5) seems to be a flexible tool that allows us to build a complex Lyapunov function from a collection of simple Lyapunov functions. It would be intriguing to investigate its potential applications beyond the restless bandit setting.

4.A Additional counterexamples for the Synchronization Assumption

Recall that in Section 4.8.3, we have compared the performances of our policies and the policies from prior work on two examples where the Synchronization Assumption (SA), required by the FTVA policy in Chapter 3 [149], does not hold. One of the examples has been defined in Figure 4.10.

In this appendix, we discuss how to generalize the graphical way of defining the example in Figure 4.10 to construct non-SA examples of arbitrary sizes. In particular, we use this method to define the example simulated in Figure 4.9b.

To specify the single-armed MDP for a non-SA example, we can first pick a set of transient states (like the states $\{0, 1, 2, 3\}$ in Figure 4.10), and a set of recurrent states (like the states $\{4, 5, 6, 7\}$ in Figure 4.10). Then we set a *required action* for each state, such that the arm goes to a next state if it follows the required action, and jumps back to a fixed transient state (we call it state 0) otherwise. The reward is positive if the arm follows the required action on the recurrent states, and is zero otherwise. The budget parameter α is chosen to be the long-run average fraction of activations if the arm always follows the required actions. To make the SA fail, we can specify the transition structure and the required actions such that from state 0, the arm can reach a recurrent state only after strictly following a *particular sequence of actions*, which should be different from all possible action sequences taken by the arms in the recurrent states. In this way, the leader arm $\hat{S}_t$ keeps circulating among the recurrent states, so the action sequence of the leader arm cannot bring the follower arm from state 0 to a recurrent state.

Using the above method, we construct another example, whose single-armed MDP is illustrated in Figure 4.11, with budget $1/2$. One can verify that if the leader arm and the follower arm are initialized as $\hat{S}_0 = 5$ and

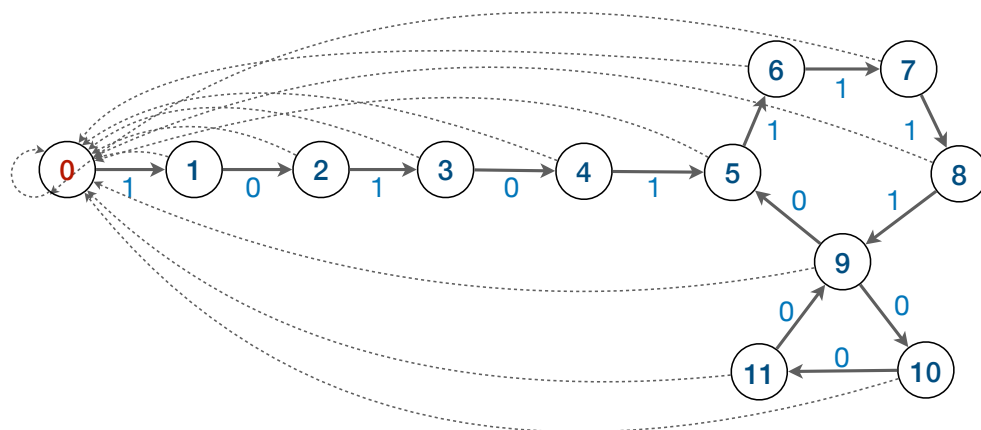


Figure 4.11: Another counterexample to the Synchronization Assumption in [149]. This figure can be interpreted in the way as Figure 4.10. The reward is 1 if the arms takes the required action in states $\{5, 6, 7, 8, 9, 10, 11\}$. The budget parameter α is set to be $1/2$.

$S_0 = 0$, the two arms never reach the same state. On the other hand, note that the optimal single-armed policy $\bar{\pi}^*$ defined by (4.3) induces an aperiodic unichain on this example, implying the compliance of Assumption 4.1.

4.B Discussion of Assumption 4.1

4.B.1 Unichain conditions in prior work

In this appendix, we discuss our version of the unichain condition stated in Assumption 4.1, which assumes that the optimal single-armed policy $\bar{\pi}^*$ induces a unichain. We compare it with other unichain-like assumptions in the literature.

The all-policy unichain condition commonly used in the average-reward MDP literature [230, Section 8.3] assumes that every stationary policy induces a unichain. Our single-policy unichain condition in Assumption 4.1 is weaker, because we only require a particular policy $\bar{\pi}^*$ to induce a unichain.

Another commonly used condition in the average-reward MDP literature is the weakly-communicating condition, which assumes that the state space can be partitioned into two sets: a closed set of states where every pair of states in the set can be reached from each other under some policy, and a

possibly empty set of states that are transient under every policy. The weakly-communicating condition, a weaker alternative to the all-policy unichain condition, ensures that an MDP has an initial-state-independent optimal average reward.

Our single-policy unichain condition in Assumption 4.1 and the weakly-communicating condition are not directly comparable. In particular,

- Our unichain condition does not imply the weakly-communicating condition because the transient states under $\bar{\pi}^*$ may not be transient under every policy.
- The weakly-communicating condition does not imply our unichain condition either. A counterexample is given in Example 3.1 of [200], as paraphrased below. Consider the following two-state MDP with the state space $\{0, 1\}$. The state of the MDP transitions to 0 (resp., 1) in the next time step with probability 1 after taking action 0 (resp., 1), regardless of the current state; the reward function is $r(1, 1) = r(0, 0) = 1$ and $r(1, 0) = r(0, 1) = 0$. This MDP is clearly communicating. However, if we consider the RB problem defined by this MDP with the budget parameter $\alpha = 1/2$, then the optimal solution to the LP relaxation (LP) is $y^*(1, 1) = y^*(0, 0) = 1/2$ and $y^*(1, 0) = y^*(0, 1) = 0$. The optimal single-armed policy is thus given by $\bar{\pi}^*(1|1) = \bar{\pi}^*(0|0) = 1$ and $\bar{\pi}^*(1|0) = \bar{\pi}^*(0|1) = 0$, which induces a Markov chain with no transitions between the two states, violating our unichain condition.

Now we review the unichain-like conditions considered in the RB literature. To the best of our knowledge, all prior work on average-reward RBs assumes the all-policy unichain assumption: In [107, 108, 266, 274], *the N -armed restless bandit system* is assumed to be unichain under every policy; in Chapter 3 [149], the single-armed MDP is assumed to be unichain under every policy. All these assumptions are stronger than our unichain condition in Assumption 4.1. In particular, assuming the N -armed restless bandit system to be unichain under every policy implies that the single-armed MDP is unichain under every policy, because if a policy for the single-armed MDP induces more than one recurrent classes, one can construct a policy for the N -armed system that also induces more than one recurrent classes. Nevertheless, all these unichain-like conditions in prior work are mostly for simplifying the presentation and can often be relaxed: For example, [108] mentions that their analysis still goes through if they assume the N -armed system to be weakly communicating; Chapter 3 [149] discusses in its appendices that the

unichain condition can be dropped as long as the Synchronization Assumption holds, albeit at the cost of a slightly more complicated formulation of the single-armed problem.

As mentioned in Section 4.2, two papers [121, 294] appearing after the arXiv version of this work further relax the unichain condition assumed here. Specifically, [294] considers the discrete-time average-reward RB problem and shows that a policy named “align-and-steer” achieves $o(1)$ optimality gap. They assume a certain control problem has certain “reachability”; this control problem captures the optimal control of the single-armed MDP’s state distribution under the expected budget constraint. [121] considers the weakly-coupled MDPs with homogeneous arms. Their policy achieves asymptotic optimality when there *exists* a single-armed policy that induces an aperiodic unichain, such that the support of the corresponding stationary distribution contains the support of the optimal stationary distribution. The assumptions in both papers are weaker than the aperiodic-unichain assumption in our work. Their assumptions also hold if the single-armed MDP is aperiodic in a mild sense and weakly-communicating.

4.B.2 Necessity of aperiodicity

In this subsection, we provide an example showing that without aperiodicity, the gap between the optimal value of the N -armed RB problem, $R^*(N, \mathbf{S}_0)$, and the optimal value of its single-armed relaxation, R^{rel} , can be non-diminishing as $N \rightarrow \infty$.

Consider a single-armed problem with two states, A and B . At each time step, the arm transitions to the other state with probability 1, regardless of the action applied. The reward function is given by $r(A, 0) = r(B, 1) = 1$ and $r(A, 1) = r(B, 0) = 0$. Let α be $\frac{1}{2}$ in the relaxed budget constraint, i.e., the arm is pulled half of the time in the long run. One can verify that an optimal policy $\bar{\pi}^*$ of this single-armed problem is given by $\bar{\pi}^*(0|A) = \bar{\pi}^*(1|B) = 1$ and $\bar{\pi}^*(1|A) = \bar{\pi}^*(0|B) = 0$, and achieves the optimal value $R^{\text{rel}} = 1$. Note that any policy in this single-armed problem induces a *periodic* unichain.

Now we consider the RB system consisting of N copies of the single-armed MDP defined above, with budget constraint $\alpha N = N/2$. Suppose all arms of the RB system are initialized in state A . Then at any time t , either all arms are in state A or all arms are in state B . In this case, all policies have the same outcome: when all arms are in state A , $N/2$ arms take action 0 and generate $N/2$ unit of reward; when all arms are in state B , $N/2$ arms

take action 1 and generate $N/2$ unit of reward. Therefore, under any policy, the long-run average reward per time step and per arm is $1/2$, which has a non-diminishing gap with the upper bound $R^{\text{rel}} = 1$.

4.C Proof of LP relaxation upper bound

In this appendix, we prove a lemma to show that the linear program (LP) is a relaxation of the restless bandit problem (RB). Although the lemma has been proved and is used in all prior work on average-reward restless bandit [see, e.g. 266, Lemma 4.3], we prove it here for completeness.

For ease of reference, we first restate (RB) and (LP).

$$\underset{\text{policy } \pi}{\text{maximize}} \quad R^-(\pi, \mathbf{S}_0) \triangleq \liminf_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i \in [N]} \mathbb{E}[r(S_t^\pi(i), A_t^\pi(i))] \quad (\text{RB})$$

$$\text{subject to} \quad \sum_{i \in [N]} A_t^\pi(i) = \alpha N, \quad \forall t \geq 0, \quad (4.1a)$$

$$\underset{\{y(s,a)\}_{s \in \mathbb{S}, a \in \mathbb{A}}}{\text{maximize}} \quad \sum_{s \in \mathbb{S}, a \in \mathbb{A}} r(s, a) y(s, a) \quad (\text{LP})$$

$$\text{subject to} \quad \sum_{s \in \mathbb{S}} y(s, 1) = \alpha, \quad (4.2a)$$

$$\sum_{s' \in \mathbb{S}, a \in \mathbb{A}} y(s', a) P(s', a, s) = \sum_{a \in \mathbb{A}} y(s, a), \quad \forall s \in \mathbb{S}, \quad (4.2b)$$

$$\sum_{s \in \mathbb{S}, a \in \mathbb{A}} y(s, a) = 1, \quad y(s, a) \geq 0, \quad \forall s \in \mathbb{S}, a \in \mathbb{A}. \quad (4.2c)$$

Next, we show that the optimal value of (LP) is an upper bound for that of (RB).

Lemma 4.22 (LP relaxation). *Let R^{rel} be the optimal value of the linear program (LP), and let $R^*(N, \mathbf{S}_0)$ be the optimal reward of the N -armed restless bandit problem (RB). Then we have*

$$R^{\text{rel}} \geq R^*(N, \mathbf{S}_0). \quad (4.98)$$

Proof of Lemma 4.22. By standard results for MDP with finite state and action spaces, there always exists a stationary Markovian policy whose

long-run average reward achieves the optimal reward $R^*(N, \mathbf{S}_0)$ [230, Theorem 9.1.8]. Therefore, to show that $R^*(N, \mathbf{S}_0) \leq R^{\text{rel}}$, it suffices to show that for each stationary Markovian policy π and initial state vector $\mathbf{S}_0$, we have $R(\pi, \mathbf{S}_0) \leq R^{\text{rel}}$.

Fixing an arbitrary stationary Markovian policy π and initial state vector $\mathbf{S}_0$, we define

$$y^\pi(s, a) \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{1}{N} \sum_{i \in [N]} \mathbf{1}\{S_t^\pi(i) = s, A_t^\pi(i) = a\} \right] \quad \forall s \in \mathbb{S}, a \in \mathbb{A}.$$

We first show that $R(\pi, \mathbf{S}_0) = \sum_{s \in \mathbb{S}, a \in \mathbb{A}} r(s, a) y^\pi(s, a)$.

$$\begin{aligned} \sum_{s \in \mathbb{S}, a \in \mathbb{A}} r(s, a) y^\pi(s, a) &= \sum_{s \in \mathbb{S}, a \in \mathbb{A}} r(s, a) \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{1}{N} \sum_{i \in [N]} \right. \\ &\quad \left. \cdot \mathbf{1}\{S_t^\pi(i) = s, A_t^\pi(i) = a\} \right] \\ &= \lim_{T \rightarrow \infty} \frac{1}{NT} \sum_{t=0}^{T-1} \sum_{i \in [N]} \mathbb{E} \left[\sum_{s \in \mathbb{S}, a \in \mathbb{A}} r(s, a) \right. \\ &\quad \left. \cdot \mathbf{1}\{S_t^\pi(i) = s, A_t^\pi(i) = a\} \right] \\ &= \lim_{T \rightarrow \infty} \frac{1}{NT} \sum_{t=0}^{T-1} \sum_{i \in [N]} \mathbb{E} [r(S_t^\pi(i), A_t^\pi(i))] \\ &= R(\pi, \mathbf{S}_0). \end{aligned}$$

Then we show that $(y^\pi(s, a))_{s \in \mathbb{S}, a \in \mathbb{A}}$ satisfies the constraints of (LP). We first consider the constraint (4.2a):

$$\begin{aligned} \sum_{s \in \mathbb{S}} y^\pi(s, 1) &= \sum_{s \in \mathbb{S}} \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{1}{N} \sum_{i \in [N]} \mathbf{1}\{S_t^\pi(i) = s, A_t^\pi(i) = 1\} \right] \\ &= \lim_{T \rightarrow \infty} \frac{1}{NT} \sum_{t=0}^{T-1} \mathbb{E} \left[\sum_{i \in [N]} \sum_{s \in \mathbb{S}} \mathbf{1}\{S_t^\pi(i) = s, A_t^\pi(i) = 1\} \right] \\ &= \lim_{T \rightarrow \infty} \frac{1}{NT} \sum_{t=0}^{T-1} \alpha N \end{aligned}$$

$$= \alpha.$$

Next, we look at the constraint (4.2b):

$$\begin{aligned}
& \sum_{s' \in \mathbb{S}, a \in \mathbb{A}} y^\pi(s', a) P(s', a, s) \\
&= \lim_{T \rightarrow \infty} \frac{1}{NT} \sum_{t=0}^{T-1} \sum_{i \in [N]} \sum_{s \in \mathbb{S}, a \in \mathbb{A}} P(s', a, s) \mathbb{P}(S_t^\pi(i) = s', A_t^\pi(i) = a) \\
&= \lim_{T \rightarrow \infty} \frac{1}{NT} \sum_{t=0}^{T-1} \sum_{i \in [N]} \mathbb{P}(S_{t+1}^\pi(i) = s) \\
&= \lim_{T \rightarrow \infty} \frac{1}{NT} \sum_{t=1}^T \sum_{i \in [N]} \mathbb{P}(S_t^\pi(i) = s) \\
&= \sum_{a \in \mathbb{A}} y^\pi(s, a).
\end{aligned}$$

Finally, we consider the constraint (4.2c):

$$\begin{aligned}
& \sum_{s' \in \mathbb{S}, a' \in \mathbb{A}} y^\pi(s', a') \\
&= \lim_{T \rightarrow \infty} \frac{1}{NT} \sum_{t=0}^{T-1} \sum_{i \in [N]} \mathbb{E} \left[\sum_{s' \in \mathbb{S}, a' \in \mathbb{A}} \mathbb{1}\{S_t^\pi(i) = s', A_t^\pi(i) = a'\} \right] = 1,
\end{aligned}$$

and it is obvious that $y^\pi(s, a) \geq 0$ for any $s \in \mathbb{S}$ and $a \in \mathbb{A}$.

Combining the above argument, $(y^\pi(s, a))_{s \in \mathbb{S}, a \in \mathbb{A}}$ is a feasible solution to (LP), so $R(\pi, \mathbf{S}_0) = \sum_{s \in \mathbb{S}, a \in \mathbb{A}} r(s, a) y^\pi(s, a) \leq R^{\text{rel}}$. Letting π be the optimal stationary Markovian policy finishes the proof. $\square$

4.D Proof of Theorem 4.9 in general case

Recall that in Section 4.5.1, we have proved Theorem 4.9 assuming that the focus-set policy induces a Markov chain that converges to a unique stationary distribution. Here we provide the general proof without this simplifying assumption.

Proof of Theorem 4.9 in the general case. Most steps in the general proof go through almost verbatim if we replace any steady-state expectations of the form $\mathbb{E}[f(S_\infty, A_\infty, X_\infty, D_\infty)]$ with the long-run averages of the form:

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[f(S_t, A_t, X_t, D_t)].$$

Note that the long-run averages of the above form always exist because (S_t, A_t, X_t, D_t) is a finite-state Markov chain and Proposition 8.1.1 in [230] can be applied with a trivial generalization of its proof, although the values of the long-run averages could depend on the initial states. With the steady-state expectations replaced by the long-run averages, we get the following analogs of (4.19) and (4.21):

$$\begin{aligned} R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0) \\ \leq r_{\max} \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \left(\mathbb{E}[\|\mu^* - X_t([N])\|_1] + 2r_{\max} \mathbb{E}[1 - m(D_t)] \right) \\ + \frac{2r_{\max} K_{\text{conf}}}{\sqrt{N}} \end{aligned} \quad (4.99)$$

$$\begin{aligned} \leq r_{\max} \left(\frac{1}{K_{\text{dist}}} + \frac{2}{L_h} \right) \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[V(X_t, D_t)] \\ + \frac{2r_{\max} K_{\text{conf}}}{\sqrt{N}}. \end{aligned} \quad (4.100)$$

The only place that needs a different treatment in the general case is in the last few steps, after deriving the drift condition for each finite t :

$$\mathbb{E}[V(X_{t+1}, D_{t+1}) \mid X_t, D_t] \leq \rho_1 V(X_t, D_t) + \frac{K_1}{\sqrt{N}}. \quad (4.22 \text{ restated})$$

We take expectation on both sides of (4.22) to get the recursive inequality on $\mathbb{E}[V(X_t, D_t)]$:

$$\mathbb{E}[V(X_{t+1}, D_{t+1})] \leq \rho_1 \mathbb{E}[V(X_t, D_t)] + \frac{K_1}{\sqrt{N}}.$$

We expand the recursion to get

$$\mathbb{E}[V(X_t, D_t)] \leq \rho_1^t \mathbb{E}[V(X_0, D_0)] + \frac{K_1}{(1 - \rho_1)\sqrt{N}}$$

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[V(X_t, D_t)] \leq \frac{1}{(1 - \rho_1)T} \mathbb{E}[V(X_0, D_0)] + \frac{K_1}{(1 - \rho_1)\sqrt{N}}.$$

Therefore, the long-run average of $\mathbb{E}[V(X_t, D_t)]$ can be bounded as

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[V(X_t, D_t)] \leq \frac{K_1}{(1 - \rho_1)\sqrt{N}}. \quad (4.101)$$

Combining (4.101) with (4.100), we finish the proof. $\square$

Remark 4.23. The proof of Theorem 4.9 also yields finite-time performance bounds. Under the same conditions, we can bound the difference between the long-run optimal reward $R^*(N, \mathbf{S}_0)$ and two finite-time performance metrics: the expected reward at time T , and the average expected reward over the first T time steps. These bounds are given by:

$$\begin{aligned} R^*(N, \mathbf{S}_0) - \mathbb{E}[R(S_T, A_T)] \\ \leq r_{\max} \left(\frac{1}{K_{\text{dist}}} + \frac{2}{L_h} \right) \mathbb{E}[V(X_0, D_0)] \rho_1^T + \frac{C_{\text{FS}}}{\sqrt{N}} \end{aligned} \quad (4.102)$$

$$\begin{aligned} R^*(N, \mathbf{S}_0) - \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[R(S_t, A_t)] \\ \leq r_{\max} \left(\frac{1}{K_{\text{dist}}} + \frac{2}{L_h} \right) \frac{\mathbb{E}[V(X_0, D_0)]}{(1 - \rho_1)T} + \frac{C_{\text{FS}}}{\sqrt{N}}, \end{aligned} \quad (4.103)$$

where $C_{\text{FS}} = r_{\max} \left(\left(\frac{1}{K_{\text{dist}}} + \frac{2}{L_h} \right) \frac{K_1}{1 - \rho_1} + 2K_{\text{conf}} \right)$. As T goes to infinity, these two upper bounds converge to the upper bound in Theorem 4.9 exponentially fast in T or at the rate $1/T$, respectively.

4.E Supplementary lemmas and proofs

In this appendix, we provide lemmas and proofs that serve as preliminaries for analyzing our policies. In Appendix 4.E.1, we show that the weight matrix W in Definition 4.11 is well-defined and prove Lemma 4.12, which claims that the state distribution of the Markov chain $P_{\bar{\pi}^*}$ converges to the steady-state distribution μ^* geometrically fast under the W -weighted L_2 norm. Then in Appendix 4.E.2, we show that two classes of functions, $\{h_W(x, D)\}_{D \subseteq [N]}$ and

$\{h_{\text{ID}}(\cdot, m)\}_{m \in [0,1]_N}$, are subset Lyapunov functions. Finally, in Appendix 4.E.3, we prove two lemmas about the L_1 norm that are used when analyzing the set-expansion policy.

4.E.1 Lemmas and proofs about matrix W and W -weighted L_2 norm

For ease of reference, we first restate Definition 4.11 below.

Definition 4.11. Let W be an $|\mathbb{S}|$ -by- $|\mathbb{S}|$ matrix given by

$$W = \sum_{k=0}^{\infty} (P_{\bar{\pi}^*} - \Xi)^k (P_{\bar{\pi}^*}^\top - \Xi^\top)^k, \quad (4.30)$$

where Ξ is an $|\mathbb{S}|$ -by- $|\mathbb{S}|$ matrix with each row being μ^* . Let λ_W denote maximal eigenvalue of W .

The next lemma shows that the matrix W is well-defined and positive definite, with eigenvalues in the range $[1, \lambda_W]$.

Lemma 4.24. *The matrix W given in Definition 4.11 is well-defined. Moreover, W is positive definite with eigenvalues lower bounded by 1.*

Proof of Lemma 4.24. Consider the sum of the spectral norm of all terms in the definition of W :

$$\sum_{k=0}^{\infty} \|(P_{\bar{\pi}^*} - \Xi)^k (P_{\bar{\pi}^*}^\top - \Xi^\top)^k\|_2.$$

Note that $(P_{\bar{\pi}^*} - \Xi)^k = P_{\bar{\pi}^*}^k - \Xi$. Because $\bar{\pi}^*$ induces an aperiodic unichain, $P_{\bar{\pi}^*}^k \rightarrow \Xi$ as $k \rightarrow \infty$. Consequently, there exist $k_0 \in \mathbb{N}^+$ and $\bar{\rho} < 1$ such that $\|(P_{\bar{\pi}^*} - \Xi)^{k_0}\|_2 = \bar{\rho}$. Then we have

$$\begin{aligned} \sum_{k=0}^{\infty} \|(P_{\bar{\pi}^*} - \Xi)^k (P_{\bar{\pi}^*}^\top - \Xi^\top)^k\|_2 &= \sum_{j=0}^{\infty} \sum_{k=j k_0}^{(j+1)k_0-1} \|(P_{\bar{\pi}^*} - \Xi)^k (P_{\bar{\pi}^*}^\top - \Xi^\top)^k\|_2 \\ &= \sum_{j=0}^{\infty} \sum_{k=0}^{k_0-1} \|(P_{\bar{\pi}^*} - \Xi)^{j k_0} (P_{\bar{\pi}^*} - \Xi)^k (P_{\bar{\pi}^*}^\top - \Xi^\top)^k (P_{\bar{\pi}^*}^\top - \Xi^\top)^{j k_0}\|_2 \\ &\leq \sum_{j=0}^{\infty} \sum_{k=0}^{k_0-1} \|(P_{\bar{\pi}^*} - \Xi)^{k_0}\|_2^j \|(P_{\bar{\pi}^*} - \Xi)^k (P_{\bar{\pi}^*}^\top - \Xi^\top)^k\|_2 \|(P_{\bar{\pi}^*}^\top - \Xi^\top)^{k_0}\|_2^j \end{aligned}$$

$$\begin{aligned}
&= \sum_{j=0}^{\infty} \bar{\rho}^{2j} \sum_{k=0}^{k_0-1} \|(P_{\bar{\pi}^*} - \Xi)^k (P_{\bar{\pi}^*}^{\top} - \Xi^{\top})^k\|_2 \\
&= \frac{C_0}{1 - \bar{\rho}^2} < \infty,
\end{aligned}$$

where $C_0 = \sum_{k=0}^{k_0-1} \|(P_{\bar{\pi}^*} - \Xi)^k (P_{\bar{\pi}^*}^{\top} - \Xi^{\top})^k\|_2$. Therefore, the infinite sum is absolutely convergent.

To show that W is positive definite, observe that each term in its definition, $(P_{\bar{\pi}^*} - \Xi)^k (P_{\bar{\pi}^*}^{\top} - \Xi^{\top})^k$, is positive semi-definite; and its first term is the identity matrix. Therefore, for any row vector $v \in \mathbb{R}^{|\mathbb{S}|}$ such that $v \neq 0$, $vWv^{\top} \geq vv^{\top}$. Therefore, W is positive definite and its eigenvalues are lower bounded by 1. $\square$

Next, we restate and prove Lemma 4.12.

Lemma 4.12 (Pseudo-contraction under W -weighted L_2 norm). *Suppose $P_{\bar{\pi}^*}$ is an aperiodic unichain on $\mathbb{S}$. Then for any distribution $v \in \Delta(\mathbb{S})$,*

$$\|vP_{\bar{\pi}^*} - \mu^*\|_W \leq \left(1 - \frac{1}{2\lambda_W}\right) \|v - \mu^*\|_W, \quad (4.31)$$

where λ_W is the maximal eigenvalue of the matrix W given in Definition 4.11, and $\|\cdot\|_W$ is the W -weighted L_2 norm, that is, $\|u\|_W = \sqrt{uWu^{\top}}$ for any row vector $u \in \mathbb{R}^{|\mathbb{S}|}$.

Proof of Lemma 4.12. Recall that by Lemma 4.24, the eigenvalues of W lie in the range $[1, \lambda_W]$, where λ_W is the maximal eigenvalue of W . It is not hard to see from the definition that W satisfies the equation

$$(P_{\bar{\pi}^*} - \Xi)W(P_{\bar{\pi}^*}^{\top} - \Xi^{\top}) - W + I = 0.$$

It follows that

$$\begin{aligned}
&\|(v - \mu^*)P_{\bar{\pi}^*}\|_W - \|v - \mu^*\|_W \\
&\leq \frac{(v - \mu^*)P_{\bar{\pi}^*}WP_{\bar{\pi}^*}^{\top}(v - \mu^*)^{\top} - (v - \mu^*)W(v - \mu^*)^{\top}}{2\|v - \mu^*\|_W} \\
&= \frac{(v - \mu^*)(P_{\bar{\pi}^*} - \Xi)W(P_{\bar{\pi}^*} - \Xi)^{\top}(v - \mu^*)^{\top} - (v - \mu^*)W(v - \mu^*)^{\top}}{2\|v - \mu^*\|_W} \\
&= \frac{(v - \mu^*)(W - I)(v - \mu^*)^{\top} - (v - \mu^*)W(v - \mu^*)^{\top}}{2\|v - \mu^*\|_W}
\end{aligned}$$

$$= -\frac{\|v - \mu^*\|_2^2}{2\|v - \mu^*\|_W}, \quad (4.104)$$

where the inequality is because $x - y \leq (x^2 - y^2)/(2y)$ for any positive numbers x, y . To change the norm in the numerator of (4.104) to the W -weighted L_2 norm, we use the following bound:

$$\|v - \mu^*\|_W^2 = (v - \mu^*)W(v - \mu^*)^\top \leq \lambda_W \|v - \mu^*\|_2^2.$$

Therefore, we have

$$\|(v - \mu^*)P_{\pi^*}\|_W - \|v - \mu^*\|_W \leq -\frac{1}{2\lambda_W} \|v - \mu^*\|_W.$$

This implies (4.31) because $\mu^*P_{\pi^*} = \mu^*$. $\square$

4.E.2 Lemmas and proofs about subset Lyapunov functions

In this section, we consider two classes of functions, $\{h_W(x, D)\}_{D \subseteq [N]}$ and $\{h_{\text{ID}}(x, m)\}_{m \in [0,1]^N}$, defined in Section 4.6.1 and Section 4.7.1 respectively. Then we prove Lemmas 4.13 and 4.17, which verify that these two classes of functions satisfy the definition of subset Lyapunov functions.

For any system state x and subset $D \subseteq [N]$, recall that $h_W(x, D)$ is defined as

$$h_W(x, D) = \|x(D) - m(D)\mu^*\|_W,$$

where W is the matrix in Definition 4.11; $\|u\|_W = \sqrt{uWu^\top}$ for any row vector u .

The lemma restated below shows that $\{h_W(x, D)\}_{D \subseteq [N]}$ are subset Lyapunov functions.

Lemma 4.13. *The class of functions $\{h_W(\cdot, D)\}_{D \subseteq [N]}$ defined in (4.32) satisfies that for any system state x and any pair of subsets $D, D' \subseteq [N]$ with $D \subseteq D'$,*

$$\begin{aligned} \mathbb{E}[h_W(X_1, D) \mid X_0 = x, A_0(i) \sim \bar{\pi}^*(\cdot \mid S_0(i)) \forall i \in D] \\ \leq \left(1 - \frac{1}{2\lambda_W}\right) h_W(x, D) + \frac{2\lambda_W^{1/2}}{\sqrt{N}} \end{aligned} \quad (4.33)$$

$$h_W(x, D) \geq \frac{1}{|\mathbb{S}|^{1/2}} \|x(D) - m(D)\mu^*\|_1 \quad (4.34)$$

$$|h_W(x, D) - h_W(x, D')| \leq L_W(m(D') - m(D)), \quad (4.35)$$

where the Lipschitz constant $L_W = 2\lambda_W^{1/2}$. These inequalities imply the drift condition, distance dominance property, and Lipschitz continuity in Definition 4.8, respectively. Consequently, $\{h_W(x, D)\}_{D \subseteq [N]}$ is a class of subset Lyapunov functions for the optimal single-armed policy $\bar{\pi}^*$.

Proof of Lemma 4.13. We first prove the drift condition (4.33). Let X'_1 be the system state after one step of transition if $A_0(i)$ is an independent sample from $\bar{\pi}^*(\cdot | S_0(i))$ for each $i \in D$. Then

$$\begin{aligned} h_W(X'_1, D) - \left(1 - \frac{1}{2\lambda_W}\right) h_W(x, D) \\ &= \|X'_1(D) - m(D)\mu^*\|_W - \left(1 - \frac{1}{2\lambda_W}\right) \|x(D) - m(D)\mu^*\|_W \\ &\leq \|X'_1(D) - m(D)\mu^*\|_W - \|x(D)P_{\bar{\pi}^*} - m(D)\mu^*\|_W \\ &\leq \|X'_1(D) - x(D)P_{\bar{\pi}^*}\|_W. \end{aligned} \quad (4.105)$$

where the first inequality follows from applying Lemma 4.12 with $v = x(D)/m(D)$; the second inequality is due to the triangle inequality. To bound the expression in (4.105), we define the centered random vector $\xi(i) = X'_1(\{i\}) - x(\{i\})P_{\bar{\pi}^*}$ for each $i \in D$, and denote its s -th entry as $\xi(i, s)$ for $s \in \mathbb{S}$. This allows us to rewrite $\|X'_1(D) - x(D)P_{\bar{\pi}^*}\|_W$ as

$$\left\|X'_1(D) - x(D)P_{\bar{\pi}^*}\right\|_W = \left\|\sum_{i \in D} \xi(i)\right\|_W. \quad (4.106)$$

Because X'_1 is the system state after each arm in D independently follows $\bar{\pi}^*$, conditioned on x , $\xi(i)$'s are independent across i and have zero means on each entry, i.e., $\mathbb{E}[\xi(i, s) | X_0 = x] = 0$ for each $i \in D$ and $s \in \mathbb{S}$. We can thus bound the conditional expectation of $\|\sum_{i \in D} \xi(i)\|_W^2$ as follows:

$$\begin{aligned} &\mathbb{E}\left[\left\|\sum_{i \in D} \xi(i)\right\|_W^2 \middle| X_0 = x\right] \\ &\leq \lambda_W \mathbb{E}\left[\left\|\sum_{i \in D} \xi(i)\right\|_2^2 \middle| X_0 = x\right] \\ &= \lambda_W \mathbb{E}\left[\sum_{s \in \mathbb{S}} \left(\sum_{i \in D} \xi(i, s)^2 + 2 \sum_{0 \leq i < i' \leq Nm_d(x)-1} \xi(i, s)\xi(i', s)\right) \middle| X_0 = x\right] \end{aligned}$$

$$\begin{aligned}
&= \lambda_W \sum_{s \in \mathbb{S}} \sum_{i \in D} \mathbb{E} \left[\xi(i, s)^2 \mid X_0 = x \right] \\
&\leq \lambda_W \sum_{i \in D} \mathbb{E} \left[\left(\sum_{s \in \mathbb{S}} |\xi(i, s)| \right)^2 \mid X_0 = x \right] \\
&\leq \frac{4\lambda_W}{N},
\end{aligned} \tag{4.107}$$

where the first inequality uses from the fact that $\|v\|_W \leq \lambda_W^{1/2} \|v\|_2$ for any $v \in \mathbb{R}^{|\mathbb{S}|}$; the first equality is by the definition of $\|\cdot\|_2$ on $\mathbb{R}^{|\mathbb{S}|}$; the second equality is because $\xi(i, s)$'s are independent across $i \in D$ and have zero means; the last inequality uses the fact that $\sum_{s \in \mathbb{S}} |\xi(i, s)| = \|\xi(i)\|_1 \leq \|X'_1(\{i\})\|_1 + \|x(\{i\})P_{\bar{\pi}^*}\|_1 = 2/N$. By the Cauchy-Schwarz inequality, it follows from (4.107) that

$$\mathbb{E} \left[\left\| \sum_{i \in D} \xi(i) \right\|_W \mid X_0 = x \right] \leq \mathbb{E} \left[\left\| \sum_{i \in D} \xi(i) \right\|_W^2 \mid X_0 = x \right]^{1/2} \leq \frac{2\lambda_W^{1/2}}{\sqrt{N}}. \tag{4.108}$$

Therefore, by combining the above calculations, we get

$$\begin{aligned}
&\mathbb{E} \left[h_W(X'_1, D) - \left(1 - \frac{1}{2\lambda_W}\right) h_W(x, D) \mid X_0 = x \right] \\
&\leq \mathbb{E} \left[\|X'_1(D) - x(D)P_{\bar{\pi}^*}\|_W \mid X_0 = x \right] \\
&= \mathbb{E} \left[\left\| \sum_{i \in D} \xi(i) \right\|_W \mid X_0 = x \right] \\
&\leq \frac{2\lambda_W^{1/2}}{\sqrt{N}},
\end{aligned}$$

which implies the drift condition (4.33).

Next, we prove the distance dominance property (4.34). Because the eigenvalues of W are at least 1, we have

$$\begin{aligned}
h_W(x, D) &= \|x(D) - m(D)\mu^*\|_W \geq \|x(D) - m(D)\mu^*\|_2 \\
&\geq \frac{1}{|\mathbb{S}|^{1/2}} \|x(D) - m(D)\mu^*\|_1.
\end{aligned}$$

Finally, we show the Lipschitz continuity (4.35). Note that

$$|h_W(x, D) - h_W(x, D')|$$

$$\begin{aligned}
&= \left| \|x(D) - m(D)\mu^*\|_W - \|x(D') - m(D')\mu^*\|_W \right| \\
&\leq \|x(D) - m(D)\mu^* - x(D') + m(D')\mu^*\|_W \\
&= \|x(D' \setminus D) - m(D' \setminus D)\mu^*\|_W \\
&\leq \|x(D' \setminus D)\|_W + m(D' \setminus D) \|\mu^*\|_W.
\end{aligned}$$

Note that for any $v \in \mathbb{R}^{|\mathbb{S}|}$, $\|v\|_W \leq \lambda_W^{1/2} \|v\|_2 \leq \lambda_W^{1/2} \|v\|_1$. Because $\|x(D' \setminus D)\|_1 = m(D') - m(D)$, and $\|\mu^*\|_1 = 1$, we have

$$\|x(D' \setminus D)\|_W + m(D' \setminus D) \|\mu^*\|_W \leq 2\lambda_W^{1/2} (m(D') - m(D)).$$

This proves the Lipschitz continuity (4.35). $\square$

Recall that for any system state x and $m \in [0, 1]_N$, $h_{\text{ID}}(x, m)$ is given by

$$h_{\text{ID}}(x, m) = \max_{\substack{m' \in [0, 1]_N \\ m' \leq m}} h_W(x, m'),$$

where $h_W(x, m') = \|x([Nm']) - m'\mu^*\|_W$. Next, we restate and prove Lemma 4.17, which implies that $\{h_{\text{ID}}(x, m)\}_{m \in [0, 1]_N}$ are also subset Lyapunov functions satisfying Definition 4.8.

Lemma 4.17. *The class of functions $\{h_{\text{ID}}(\cdot, m)\}_{m \in [0, 1]_N}$ defined in (4.63) satisfies that for any system state x and any $m, m' \in [0, 1]_N$,*

$$\begin{aligned}
&\mathbb{E} \left[\left(h_{\text{ID}}(X_1, m) - \left(1 - \frac{1}{2\lambda_W}\right) h_{\text{ID}}(x, m) \right)^+ \right. \\
&\quad \left. \left| X_0 = x, A_0(i) \sim \bar{\pi}^*(\cdot | S_0(i)) \forall i \in [Nm] \right] \right. \\
&\quad \leq \frac{4\lambda_W^{1/2}}{\sqrt{N}}, \tag{4.64}
\end{aligned}$$

$$h_{\text{ID}}(x, m) \geq \frac{1}{|\mathbb{S}|^{1/2}} \|x([Nm]) - m\mu^*\|_1, \tag{4.65}$$

$$|h_{\text{ID}}(x, m) - h_{\text{ID}}(x, m')| \leq 2\lambda_W^{1/2} |m' - m|. \tag{4.66}$$

These inequalities imply the drift condition, distance dominance property, and Lipschitz continuity in Definition 4.8, respectively. Consequently, $\{h_{\text{ID}}(x, m)\}_{m \in [0, 1]_N}$ are subset Lyapunov functions for the optimal single-armed policy $\bar{\pi}^*$.

Proof. In this proof, the variable n is a non-negative integer unless otherwise stated. We first prove the drift condition (4.64). Let X'_1 be the system state after one step of transition if $A_0(i)$ is an independent sample from $\bar{\pi}^*(\cdot|S_0(i))$ for each $i \in [Nm]$. Then

$$\begin{aligned}
& h_{\text{ID}}(X'_1, m) - \left(1 - \frac{1}{2\lambda_W}\right) h_{\text{ID}}(x, m) \\
&= \max_{m' \in [0,1]_N, m' \leq m} h_W(X'_1, m') - \left(1 - \frac{1}{2\lambda_W}\right) \max_{m' \in [0,1]_N, m' \leq m} h_W(x, m') \\
&\leq \max_{m' \in [0,1]_N, m' \leq m} \left(h_W(X'_1, m') - \left(1 - \frac{1}{2\lambda_W}\right) h_W(x, m') \right) \\
&\leq \max_{m' \in [0,1]_N, m' \leq m} \|X'_1([Nm']) - x([Nm'])P_{\bar{\pi}^*}\|_W, \tag{4.109}
\end{aligned}$$

where the last inequality can be justified using the same argument as (4.105). Therefore,

$$\begin{aligned}
& \left(h_{\text{ID}}(X'_1, m) - \left(1 - \frac{1}{2\lambda_W}\right) h_{\text{ID}}(x, m) \right)^+ \\
&\leq \max_{m' \in [0,1]_N, m' \leq m} \|X'_1([Nm']) - x([Nm'])P_{\bar{\pi}^*}\|_W. \tag{4.110}
\end{aligned}$$

For any $i \in [Nm]$, we define the centered random vector $\xi(i) \in \mathbb{R}^{|\mathbb{S}|}$ as

$$\xi(i) = X'_1(\{i\}) - x(\{i\})P_{\bar{\pi}^*},$$

and denote its s -th entry as $\xi(i, s)$ for $s \in \mathbb{S}$. This allows us to rewrite the right-hand side of (4.110) as

$$\max_{m' \in [0,1]_N, m' \leq m} \|X'_1([Nm']) - x([Nm'])P_{\bar{\pi}^*}\|_W = \max_{n \leq Nm} \left\| \sum_{i \in [n]} \xi(i) \right\|_W. \tag{4.111}$$

Therefore, to prove the bound in (4.64), it suffices to show that

$$\mathbb{E} \left[\max_{n \leq Nm} \left\| \sum_{i \in [n]} \xi(i) \right\|_W \middle| X_0 = x \right] \leq \frac{4\lambda_W^{1/2}}{\sqrt{N}}. \tag{4.112}$$

Conditioned on $X_0 = x$, we argue that $\left\| \sum_{i \in [n]} \xi(i) \right\|_W$ is a submartingale in n so that we can invoke Doob's L_2 maximal inequality to bound the right-hand side of (4.111) (see, e.g., Theorem 4.4.4 of [86]). Observe that

- $\xi(i)$'s are independent across $i \in [Nm]$;
- For each $i \in [Nm]$ and $s \in \mathbb{S}$, $\mathbb{E}[\xi(i, s) \mid X_0 = x] = 0$.

Therefore, $\sum_{i \in [n]} \xi(i)$ is a martingale in n . Because $\|\cdot\|_W$ is a convex function, $\left\| \sum_{i \in [n]} \xi(i) \right\|_W$ is a submartingale in n . We apply Doob's L_2 maximal inequality to $\left\| \sum_{i \in [n]} \xi(i) \right\|_W$ to get

$$\mathbb{E} \left[\left(\max_{n \leq Nm} \left\| \sum_{i \in [n]} \xi(i) \right\|_W \right)^2 \mid X_0 = x \right] \leq 4 \mathbb{E} \left[\left\| \sum_{i \in [Nm]} \xi(i) \right\|_W^2 \mid X_0 = x \right]. \quad (4.113)$$

Applying Cauchy-Schwarz inequality to the LHS of (4.113), we get

$$\mathbb{E} \left[\max_{n \leq Nm} \left\| \sum_{i \in [n]} \xi(i) \right\|_W \mid X_0 = x \right] \leq \mathbb{E} \left[\left(\max_{n \leq Nm} \left\| \sum_{i \in [n]} \xi(i) \right\|_W \right)^2 \mid X_0 = x \right]^{1/2} \quad (4.114)$$

Using the same argument as (4.107) with $D = [Nm]$, we bound the right-hand side of (4.113) as

$$4 \mathbb{E} \left[\left\| \sum_{i \in [Nm]} \xi(i) \right\|_W^2 \mid X_0 = x \right] \leq \frac{16 \lambda_W}{N}. \quad (4.115)$$

Plugging (4.114) and (4.115) into two sides of (4.113), we get

$$\mathbb{E} \left[\max_{n \leq Nm} \left\| \sum_{i \in [n]} \xi(i) \right\|_W \mid X_0 = x \right] \leq \frac{4 \lambda_W^{1/2}}{\sqrt{N}}. \quad (4.116)$$

This proves the drift condition (4.64).

Next, we show the distance dominance property (4.65). By the definition of $h_{\text{ID}}(x, m)$ and the fact that the eigenvalues of W are at least 1, we have

$$\begin{aligned} h_{\text{ID}}(x, m) &\geq \|x([Nm]) - m\mu^*\|_W \geq \|x([Nm]) - m\mu^*\|_2 \\ &\geq \frac{1}{|\mathbb{S}|^{1/2}} \|x([Nm]) - m\mu^*\|_1. \end{aligned}$$

Finally, we show the Lipschitz continuity (4.66). For simplicity, we omit $m \in [0, 1]_N$ in the subscripts. Consider any $m, m' \in [0, 1]_N$. Without loss of

generality, we assume that $m \leq m'$. We expand the definitions of $h_{\text{ID}}(x, m)$ and $h_{\text{ID}}(x, m')$ to rewrite them as follows:

$$\begin{aligned} h_{\text{ID}}(x, m) &= \max \{h_{\text{ID}}(x, m), h_W(x, m)\} \\ h_{\text{ID}}(x, m') &= \max \left\{ h_{\text{ID}}(x, m), \max_{m'' \in [m, m']_N} h_W(x, m'') \right\}, \end{aligned}$$

where $[m, m']_N$ is a shorthand for $[m, m'] \cap [0, 1]_N$. Observe that for any $a, b, c \in \mathbb{R}$, we have $|\max\{a, b\} - \max\{a, c\}| \leq |b - c|$. Letting $a = h_{\text{ID}}(x, m)$, $b = h_W(x, m)$, and $c = \max_{m'' \in [m, m']_N} h_W(x, m'')$, we get

$$\left| h_{\text{ID}}(x, m) - h_{\text{ID}}(x, m') \right| \leq \left| \max_{m'' \in [m, m']_N} h_W(x, m'') - h_W(x, m) \right|. \quad (4.117)$$

We further bound the right-hand side of (4.117) as

$$\begin{aligned} & \left| \max_{m'' \in [m, m']_N} h_W(x, m'') - h_W(x, m) \right| \\ & \leq \max_{m'' \in [m, m']_N} |h_W(x, m'') - h_W(x, m)| \\ & \leq \max_{m'' \in [m, m']_N} 2\lambda_W^{1/2} |m'' - m| \\ & = 2\lambda_W^{1/2} |m' - m|, \end{aligned} \quad (4.118)$$

where in the second inequality we used (4.35), the Lipschitz continuity of $h_W(x, D)$ in D that we have proved in Lemma 4.13. Combining (4.117) and (4.118), we prove the Lipschitz continuity (4.66). $\square$

4.E.3 Lemmas and proofs about L_1 norm

In this subsection, we prove two lemmas about the L_1 norm that are useful for the analysis of the set-expansion policy, considering that the set-expansion policy chooses focus sets based on the slack $\delta(x, D)$ whose definition involves L_1 norm.

We first show that if the optimal single-armed policy $\bar{\pi}^*$ induces an aperiodic unichain, right-multiplying $P_{\bar{\pi}^*}$ is non-expansive under the L_1 norm. **Lemma 4.25** (Non-expansiveness of $P_{\bar{\pi}^*}$ under L_1 norm). *For any distribution $v \in \Delta(\mathbb{S})$,*

$$\|(v - \mu^*)P_{\bar{\pi}^*}\|_1 \leq \|v - \mu^*\|_1. \quad (4.119)$$

Proof. For any $v \in \Delta(\mathbb{S})$,

$$\begin{aligned}
\|(v - \mu^*)P_{\bar{\pi}^*}\|_1 &= \sum_{s' \in \mathbb{S}} \left| \sum_{s \in \mathbb{S}} (v(s) - \mu^*(s)) P_{\bar{\pi}^*}(s, s') \right| \\
&\leq \sum_{s' \in \mathbb{S}} \sum_{s \in \mathbb{S}} |v(s) - \mu^*(s)| P_{\bar{\pi}^*}(s, s') \\
&= \sum_{s \in \mathbb{S}} |v(s) - \mu^*(s)| \sum_{s' \in \mathbb{S}} P_{\bar{\pi}^*}(s, s') \\
&= \sum_{s \in \mathbb{S}} |v(s) - \mu^*(s)| \\
&= \|v - \mu^*\|_1. \quad \square
\end{aligned}$$

Next, we show that if all arms in a subset D follow $\bar{\pi}^*$, the L_1 distance between the scaled state-count vector $X_t(D)$ and the scaled optimal steady-state distribution $m(D)\mu^*$ does not significantly increase.

Lemma 4.26. *For any system state x and any subset $D \subseteq [N]$,*

$$\begin{aligned}
&\mathbb{E}[(\|X_1(D) - m(D)\mu^*\|_1 - \|x(D) - m(D)\mu^*\|_1)^+ \mid \\
&\quad X_0 = x, A_0(i) \sim \bar{\pi}^*(\cdot | S_0(i)) \forall i \in D] \leq \frac{2|\mathbb{S}|^{1/2}}{\sqrt{N}}. \quad (4.120)
\end{aligned}$$

Proof. Let X'_1 be the system state after one step of transition if $A_0(i) \sim \bar{\pi}^*(\cdot | S_0(i))$ for any $i \in D$. Then

$$\begin{aligned}
&\|X'_1(D) - m(D)\mu^*\|_1 - \|x(D) - m(D)\mu^*\|_1 \\
&\leq \|X'_1(D) - m(D)\mu^*\|_1 - \|x(D)P_{\bar{\pi}^*} - m(D)\mu^*\|_1 \\
&\leq \|X'_1(D) - x(D)P_{\bar{\pi}^*}\|_1, \quad (4.121)
\end{aligned}$$

where the first inequality follows from applying Lemma 4.25 with $v = x(D)/m(D)$; the second inequality is due to the triangular inequality. Therefore,

$$(\|X'_1(D) - m(D)\mu^*\|_1 - \|x(D) - m(D)\mu^*\|_1)^+ \leq \|X'_1(D) - x(D)P_{\bar{\pi}^*}\|_1. \quad (4.122)$$

For any $i \in [Nm_d(x)]$, we define the centered random vector $\xi(i) \in \mathbb{R}^{|\mathbb{S}|}$ as

$$\xi(i) = X'_1(\{i\}) - x(\{i\})P_{\bar{\pi}^*},$$

and denote its s -th entry as $\xi(i, s)$ for $s \in \mathbb{S}$. This allows us to rewrite $\|X'_1(D) - x(D)P_{\bar{\pi}^*}\|_1$ as

$$\left\|X'_1(D) - x(D)P_{\bar{\pi}^*}\right\|_1 = \left\|\sum_{i \in D} \xi(i)\right\|_1. \quad (4.123)$$

Observe that conditioned on $X_0 = x$, we have the following facts about these centered random vectors:

- $\xi(i)$'s are independent across $i \in D$;
- For each $i \in D$ and $s \in \mathbb{S}$, $\mathbb{E}[\xi(i, s) \mid X_0 = x] = 0$.

Therefore, we can bound the conditional expectation of $\|\sum_{i \in D} \xi(i)\|_1^2$ as follows:

$$\begin{aligned} & \mathbb{E}\left[\left\|\sum_{i \in D} \xi(i)\right\|_1^2 \middle| X_0 = x\right] \\ & \leq |\mathbb{S}| \mathbb{E}\left[\left\|\sum_{i \in D} \xi(i)\right\|_2^2 \middle| X_0 = x\right] \\ & = |\mathbb{S}| \mathbb{E}\left[\sum_{s \in \mathbb{S}} \left(\sum_{i \in D} \xi(i, s)^2 + 2 \sum_{0 \leq i < i' \leq Nm_d(x)-1} \xi(i, s)\xi(i', s)\right) \middle| X_0 = x\right] \\ & = |\mathbb{S}| \sum_{s \in \mathbb{S}} \sum_{i \in D} \mathbb{E}\left[\xi(i, s)^2 \middle| X_0 = x\right] \\ & \leq |\mathbb{S}| \sum_{i \in D} \mathbb{E}\left[\left(\sum_{s \in \mathbb{S}} |\xi(i, s)|\right)^2 \middle| X_0 = x\right] \\ & \leq \frac{4|\mathbb{S}|}{N}, \end{aligned} \quad (4.124)$$

where the first inequality uses from the fact that $\|v\|_1 \leq |\mathbb{S}|^{1/2} \|v\|_2$ for any $v \in \mathbb{R}^{|\mathbb{S}|}$; the first equality is by the definition of $\|\cdot\|_2$ on $\mathbb{R}^{|\mathbb{S}|}$; the second equality is because $\xi(i, s)$'s are independent across $i \in D$ and have zero means; the last inequality uses the fact that $\sum_{s \in \mathbb{S}} |\xi(i, s)| = \|\xi(i)\|_1 \leq \|X'_1(\{i\})\|_1 + \|x(\{i\})P_{\bar{\pi}^*}\|_1 = 2/N$. By the Cauchy-Schwarz inequality, it follows from (4.124) that

$$\mathbb{E}\left[\left\|\sum_{i \in D} \xi(i)\right\|_1 \middle| X_0 = x\right] \leq \mathbb{E}\left[\left\|\sum_{i \in D} \xi(i)\right\|_1^2 \middle| X_0 = x\right]^{1/2} \leq \frac{2|\mathbb{S}|^{1/2}}{\sqrt{N}}. \quad (4.125)$$

Combining the above calculations, we get

$$\begin{aligned}
& \mathbb{E}[(\|X'_1(D) - m(D)\mu^*\|_1 - \|x(D) - m(D)\mu^*\|_1)^+ \mid X_0 = x] \\
& \leq \mathbb{E}[\|X'_1(D) - x(D)P_{\pi^*}\|_1 \mid X_0 = x] \\
& = \mathbb{E}\left[\left\|\sum_{i \in D} \xi(i)\right\|_1 \mid X_0 = x\right] \\
& \leq \frac{2|S|^{1/2}}{\sqrt{N}},
\end{aligned}$$

which implies (4.120). $\square$

4.F Set-optimization policy

This appendix introduces the *set-optimization policy*, another focus-set policy that also achieves an $O(1/\sqrt{N})$ optimality gap. The key distinguishing feature of the set-optimization policy is that it is Markovian with respect to the scaled state-count vector $X_t([N])$, meaning that $X_t([N])$ forms a Markov chain under this policy — a property not shared by the set-expansion policy (Algorithm 4) or the ID policy (Algorithm 5). Although the advantages of the set-optimization policy over the other two focus-set policies are not evident as it does not perform better in simulations, we include it here for its potential theoretical interests.

4.F.1 Definition and asymptotic optimality of set-optimization policy

The set-optimization policy is similar to the set-expansion policy in that they both choose a focus set D_t in each time step and give priority to arms in D_t to follow their ideal actions. However, they differ in how D_t is chosen. In the set-optimization policy, D_t is updated by solving an optimization problem (4.126a)–(4.126b), where $h_W(x, D) = \|X_t(D) - m(D)\mu^*\|_W$ is the subset Lyapunov function that we have used in Section 4.6, $L_W = 2\lambda_W^{1/2}$, and the slack $\delta(x, D) = \beta(1 - m(D)) - 0.5\|x(D) - m(D)\mu^*\|_1$ is the same notion as in Section 4.4.2. Importantly, D_t is chosen to be a *maximal* optimal solution in the sense that there is no other optimal solution D' that contains D_t . When there are multiple maximal optimal solutions, D_t is picked uniformly

at random. The formal definition of the set-optimization policy is given in Algorithm 6.

We note that under the set-optimization policy, the number of arms corresponding to each state-action pair at time step t is determined by the state-count vector $X_t([N])$ rather than the full system state X_t , implying that $X_t([N])$ is a Markov chain. Although the subproblem in (4.126a)–(4.126b) requires evaluating $X_t(D)$ for a specific subset D , which seems to require full knowledge of the system state X_t , we observe that $h_W(X_t, D)$, $m(D)$, and $\delta(X_t, D)$ are determined solely by $X_t(D)$. Thus, solving the subproblem boils down to deciding $X_t(D_t)$, which only requires the knowledge of $X_t([N])$ rather than X_t .

Algorithm 6 Set-optimization policy

Input: number of arms N , budget αN , the optimal single-armed policy $\bar{\pi}^*$, initial system state X_0 , initial state vector $\mathbf{S}_0$

- 1: **for** $t = 0, 1, 2, \dots$ **do**
 $\triangleright$ *Set update*
 - 2: Let D_t be a maximal optimal solution to the problem below:
$$D_t \leftarrow \arg \min_{D \subseteq [N]} h_W(X_t, D) + L_W(1 - m(D)) \quad (4.126a)$$

$$\text{subject to } \delta(X_t, D) \geq 0 \quad (4.126b)$$
 - 3: Run the same action sampling and action rectification steps as in Lines 6–15 of Algorithm 4
 - 4: Apply $A_t(i)$ and observe $S_{t+1}(i)$ for each arm $i \in [N]$
-

Our next theorem, Theorem 4.27, shows that the set-optimization policy achieves an $O(1/\sqrt{N})$ optimality gap, which matches the order of optimality gaps of the set-expansion policy and the ID policy.

Theorem 4.27 (Optimality gap of set-optimization policy). *Consider an N -armed restless bandit problem with the single-armed MDP $(\mathbb{S}, \mathbb{A}, P, r)$ and budget αN for $0 < \alpha < 1$. Assume that the optimal single-armed policy induces an aperiodic unichain (Assumption 4.1). Let π be the set-optimization policy (Algorithm 6). Then for all N and initial state vector $\mathbf{S}_0$, the optimality gap of π is bounded as*

$$R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0) \leq \frac{C_{\text{so}}}{\sqrt{N}}, \quad (4.127)$$

where C_{SO} is a constant depending on $r_{\max}$, $|\mathbb{S}|$, $\beta \triangleq \min\{\alpha, 1 - \alpha\}$, and $P_{\bar{\pi}^*}$; the explicit expression of C_{SO} is given in the proof.

4.F.2 Proof of Theorem 4.27 (Optimality gap of set-optimization policy)

We will spend the remainder of this appendix proving Theorem 4.27. Unlike the ID policy and the set-expansion policy, the set-optimization policy does not satisfy Condition 2, so Theorem 4.27 can not be proved as a direct corollary of Theorem 4.9. However, the proof of Theorem 4.27 follows a similar structure as the framework established in Section 4.5. Specifically, in Appendix 4.F.2, we state and prove three lemmas; each lemma either verifies a condition or states a fact that modifies one of the three conditions in Section 4.5; in Appendix 4.F.2, we prove Theorem 4.27 using similar ideas as Theorem 4.9, with the subset Lyapunov functions being $\{h_W(x, D)\}_{D \subseteq [N]}$.

Lemmas and proofs

We first show that the set-optimization policy satisfies Condition 1.

Lemma 4.28 (Set-optimization policy satisfies Condition 1). *Consider the set-optimization policy defined in Algorithm 6. For any time step $t \geq 0$, there exists a subset $D'_t \subseteq D_t$ such that for all $i \in D'_t$, the policy chooses $A_t(i) = \hat{A}_t(i)$, and*

$$\mathbb{E}[m(D_t \setminus D'_t) \mid X_t, D_t] \leq \frac{1}{\sqrt{N}} + \frac{1}{N} \quad a.s. \quad (4.128)$$

Proof of Lemma 4.28. The whole proof is verbatim to the proof of Lemma 4.14, based on the fact that for both the set-optimization policy and the set-expansion policy, D_t satisfies $\delta(X_t, D_t) \geq 0$, and $\mathbf{A}_t$ is chosen such that the number of arms $i \in D_t$ with $A_t(i) = \hat{A}_t(i)$ is maximized. $\square$

Although the set-optimization policy does not satisfy the almost non-shrinking condition (Condition 2), we show that for each time step t , there is another subset D_{t+1}^{SE} such that the pair of subsets $(D_t, D_{t+1}^{\text{SE}})$ satisfies Condition 2, and D_{t+1}^{SE} is feasible to the optimization subproblem (4.126a)–(4.126b) in the $(t + 1)$ -th time step.

Lemma 4.29. *Consider the set-optimization policy defined in Algorithm 6. For any $t \geq 0$, there exists a subset $D_{t+1}^{\text{SE}} \subseteq [N]$ that depends on X_{t+1} such that*

1. $\delta(X_{t+1}, D_{t+1}^{\text{SE}}) \geq 0$;
 2. either $D_{t+1}^{\text{SE}} \supseteq D_t$ or $D_{t+1}^{\text{SE}} \subseteq D_t$;
 3. $\mathbb{E}[(m(D_t) - m(D_{t+1}^{\text{SE}}))^+ \mid X_t, D_t] \leq \frac{|\mathbb{S}|^{1/2} + 1}{\beta\sqrt{N}} + \frac{1 + (\beta + 1)|\mathbb{S}|}{\beta N} \quad a.s.$
- (4.129)

Proof of Lemma 4.29. We let the set D_{t+1}^{SE} be the next-time-step focus set if we apply the update rule of the set-expansion policy to (X_t, D_t) . By the definition of the set-expansion policy, we automatically get $\delta(X_{t+1}, D_{t+1}^{\text{SE}}) \geq 0$, and we also have $D_{t+1}^{\text{SE}} \supseteq D_t$ or $D_{t+1}^{\text{SE}} \subseteq D_t$.

The proof of (4.129) closely follows that of (4.37) in Lemma 4.15, with the only change being that D_{t+1} is replaced by D_{t+1}^{SE} . The logic of both proofs relies on two key facts that hold in this context:

- The definition of the set-expansion policy implies that when $D_{t+1}^{\text{SE}} \subseteq D_t$, D_{t+1}^{SE} is chosen to be the subset with the largest number of arms among all sets $\bar{D}$ such that $\delta(X_{t+1}, \bar{D}) \geq 0$.
- By Lemma 4.28, there exists a subset $D'_t \subseteq D_t$ such that for all $i \in D'_t$, the policy chooses $A_t(i) = \hat{A}_t(i)$, and $\mathbb{E}[m(D_t \setminus D'_t) \mid X_t, D_t] = O(1/\sqrt{N})$.

With these two properties established, the rest of the argument proceeds identically to the proof of Lemma 4.15 with D_{t+1} replaced by D_{t+1}^{SE} . $\square$

Finally, we show that the set-optimization policy satisfies Condition 3.

Lemma 4.30 (Set-optimization policy satisfies Condition 3). *Consider the set-optimization policy defined in Algorithm 6. For any $t \geq 0$,*

$$1 - m(D_t) \leq \frac{|\mathbb{S}|^{1/2}}{\beta} h_W(X_t, D_t) + \frac{2}{\beta N} \quad a.s., \quad (4.130)$$

Proof of Lemma 4.30. Recall that D_t is chosen to be maximal among the optimal solutions of

$$\min_{D \subseteq [N]} h_W(X_t, D) + L_W(1 - m(D)) \quad (4.126a)$$

$$\text{subject to } \delta(X_t, D) \geq 0. \quad (4.126b)$$

Because $h_W(X_t, D)$ is L_W -Lipschitz continuous in D according to Lemma 4.13, the objective $h_W(X_t, D) + L_W(1 - m(D))$ is non-increasing as D expands. Consequently, there is no subset D' strictly containing D_t that satisfies $\delta(X_t, D') \geq 0$, because otherwise D' would be an optimal solution to (4.126a)-(4.126b) that strictly contains D_t , violating the maximality of D_t . Then we must have

$$\beta(1 - m(D_t)) - \frac{1}{2} \|X_t(D_t) - m(D_t)\mu^*\|_1 \leq \frac{2}{N},$$

because otherwise, $m(D_t) < 1$, and we can pick any $i \notin D_t$ and show that $\delta(X_t, D_t \cup \{i\}) \geq 0$. Therefore,

$$\begin{aligned} 1 - m(D_t) &\leq \frac{1}{\beta} \|X_t(D_t) - m(D_t)\mu^*\|_1 + \frac{2}{\beta N} \\ &\leq \frac{|\mathbb{S}|^{1/2}}{\beta} h_W(X_t, D_t) + \frac{2}{\beta N} \end{aligned} \quad (4.131)$$

where (4.131) is by the distance domination property of $h_W(x, D)$ established in Lemma 4.13. $\square$

Proof of Theorem 4.27

Next, we prove Theorem 4.27. For simplicity, we focus on the case where the policy's induced Markov chain converges to a unique stationary distribution. The extension to the general case requires the same minor modifications as in the proof of Theorem 4.9 and is omitted to avoid repetition.

Proof of Theorem 4.27. Following the same steps that derive (4.21) in Theorem 4.9, we get

$$R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0) \leq r_{\max} \left(\frac{1}{K_{\text{dist}}} + \frac{2}{L_W} \right) \mathbb{E}[V(X_\infty, D_\infty)] + \frac{2r_{\max}K_{\text{conf}}}{\sqrt{N}}, \quad (4.132)$$

where

$$V(x, D) = h_W(x, D) + L_W(1 - m(D)).$$

Therefore, it suffices to bound $\mathbb{E}[V(X_\infty, D_\infty)]$.

We fix any $t \geq 0$. Recall that D_{t+1} is chosen to be the minimizer of $V(X_{t+1}, D)$ among sets D with $\delta(X_{t+1}, D) \geq 0$. Because D_{t+1}^{SE} defined in Lemma 4.29 satisfies $\delta(X_{t+1}, D_{t+1}^{\text{SE}}) \geq 0$, we must have

$$V(X_{t+1}, D_{t+1}) \leq V(X_{t+1}, D_{t+1}^{\text{SE}}). \quad (4.133)$$

Therefore,

$$\begin{aligned}
V(X_{t+1}, D_{t+1}) &\leq V(X_{t+1}, D_{t+1}^{\text{SE}}) \\
&= h_W(X_{t+1}, D_{t+1}^{\text{SE}}) + L_W(1 - m(D_{t+1}^{\text{SE}})) \\
&\leq h_W(X_{t+1}, D_t) + L_W|m(D_{t+1}^{\text{SE}}) - m(D_t)| \\
&\quad + L_W(1 - m(D_t)) + L_W(m(D_t) - m(D_{t+1}^{\text{SE}})) \tag{4.134}
\end{aligned}$$

$$\begin{aligned}
&= h_W(X_{t+1}, D_t) + L_W(1 - m(D_t)) \\
&\quad + 2L_W(m(D_t) - m(D_{t+1}^{\text{SE}}))^+, \tag{4.135}
\end{aligned}$$

where the second inequality utilizes the fact that $D_{t+1}^{\text{SE}} \supseteq D_t$ or $D_{t+1}^{\text{SE}} \subseteq D_t$ established in Lemma 4.29, and the Lipschitz continuity of $h_W(x, D)$ with respect to D established in Lemma 4.13.

Therefore, subtracting $V(X_t, D_t)$ and taking the expectation in (4.135) conditioned on X_t , we have

$$\begin{aligned}
&\mathbb{E}[V(X_{t+1}, D_{t+1}) - V(X_t, D_t) \mid X_t] \\
&\leq \mathbb{E}[h_W(X_{t+1}, D_t) - h_W(X_t, D_t) \mid X_t] \tag{4.136}
\end{aligned}$$

$$+ 2L_W\mathbb{E}[(m(D_t) - m(D_{t+1}^{\text{SE}}))^+ \mid X_t]. \tag{4.137}$$

We will bound each of the terms in (4.136) and (4.137) separately.

To bound the term in (4.136), notice that by Lemma 4.28, there exists $D'_t \subseteq D_t$ such that for any $i \in D'_t$, the policy chooses $A_t(i) = \hat{A}_t(i)$, and $\mathbb{E}[m(D_t \setminus D'_t) \mid X_t, D_t] = O(1/\sqrt{N})$. Let X'_{t+1} be the random element denoting the system state at time $t+1$ if $A_t(i) = \hat{A}_t(i)$ for all $i \in [N]$. We couple X_{t+1} with X'_{t+1} such that $X_{t+1}(D'_t) = X'_{t+1}(D'_t)$, then we have $h_W(X_{t+1}, D'_t) = h_W(X'_{t+1}, D'_t)$. Consequently,

$$\begin{aligned}
&\mathbb{E}[h_W(X_{t+1}, D_t) \mid X_t] \\
&= \mathbb{E}[h_W(X'_{t+1}, D_t) \mid X_t] + \mathbb{E}[h_W(X_{t+1}, D_t) - h_W(X'_{t+1}, D_t) \mid X_t] \\
&\leq \rho_2\mathbb{E}[h_W(X_t, D_t) \mid X_t] + \frac{K_{\text{drift}}}{\sqrt{N}} \tag{4.138}
\end{aligned}$$

$$\begin{aligned}
&\quad + \mathbb{E}[h_W(X_{t+1}, D_t) - h_W(X'_{t+1}, D_t) \mid X_t] \\
&\leq \rho_2\mathbb{E}[h_W(X_t, D_t) \mid X_t] + \frac{K_{\text{drift}}}{\sqrt{N}} \tag{4.139} \\
&\quad + 2L_W\mathbb{E}[m(D_t \setminus D'_t) \mid X_t]
\end{aligned}$$

$$\leq \rho_2 \mathbb{E}[h_W(X_t, D_t) \mid X_t] + \frac{K_{\text{drift}} + 2L_W K_{\text{conf}}}{\sqrt{N}}, \quad (4.140)$$

where $\rho_2 = 1 - 1/(2\lambda_W)$, $K_{\text{drift}} = 2\lambda_W^{1/2}$, $K_{\text{conf}} = 2$; the inequality in (4.138) follows from the drift condition of $h_W(x, D)$ established in Lemma 4.13; to get the inequality in (4.139), we use the argument that

$$\begin{aligned} & h_W(X_{t+1}, D_t) - h_W(X'_{t+1}, D_t) \\ &= h_W(X_{t+1}, D_t) - h_W(X_{t+1}, D'_t) + h_W(X'_{t+1}, D'_t) - h_W(X'_{t+1}, D_t) \\ &\leq 2L_W m(D_t \setminus D'_t); \end{aligned}$$

the inequality in (4.140) follows from the majority conformity of the set-optimization policy proved in Lemma 4.28. Therefore,

$$\begin{aligned} & \mathbb{E}[h_W(X_{t+1}, D_t) \mid X_t] - h_W(X_t, D_t) \\ &\leq -(1 - \rho_2) \mathbb{E}[h_W(X_t, D_t) \mid X_t] + \frac{K_{\text{drift}} + 2L_W K_{\text{conf}}}{\sqrt{N}}. \end{aligned}$$

To bound the term $2L_W \mathbb{E}[(m(D_t) - m(D_{t+1}^{\text{SE}}))^+ \mid X_t = x]$ in (4.137), we apply Lemma 4.29 to get

$$2L_W \mathbb{E}[(m(D_t) - m(D_{t+1}^{\text{SE}}))^+ \mid X_t] \leq \frac{2L_W K_{\text{mono}}}{\sqrt{N}},$$

where $K_{\text{mono}} = (2 + (\beta + 2)|\mathbb{S}|)/\beta$. Plugging the above bounds into (4.136) and (4.137), we get

$$\begin{aligned} & \mathbb{E}[V(X_{t+1}, D_{t+1}) - V(X_t, D_t) \mid X_t] \\ &\leq -(1 - \rho_2) \mathbb{E}[h_W(X_t, D_t) \mid X_t] + \frac{K_{\text{drift}} + 2L_W(K_{\text{conf}} + K_{\text{mono}})}{\sqrt{N}}. \end{aligned} \quad (4.141)$$

Note that by Lemma 4.30,

$$V(X_t, D_t) \leq (1 + L_W L_{\text{cov}}) h_W(X_t, D_t) + \frac{L_W K_{\text{cov}}}{\sqrt{N}}, \quad (4.142)$$

where $L_{\text{cov}} = |\mathbb{S}|^{1/2}/\beta$, $K_{\text{cov}} = 3/\beta$. Combining (4.142) with (4.141), we have proved that for any $t \geq 0$,

$$\mathbb{E}[V(X_{t+1}, D_{t+1}) \mid X_t] \leq \rho_1 \mathbb{E}[V(X_t, D_t) \mid X_t] + \frac{K_1}{\sqrt{N}}, \quad (4.143)$$

where $\rho_1 = 1 - (1 - \rho_2)/(1 + L_W L_{\text{cov}})$ and $K_1 = K_{\text{drift}} + 2L_W(K_{\text{conf}} + K_{\text{mono}}) + (1 - \rho_2)L_W K_{\text{cov}}/(1 + L_W L_{\text{cov}})$.

Now, with (4.143), $\mathbb{E}[V(X_\infty, D_\infty)]$ can be bounded as follows: We take the expectation on both sides of (4.143) conditioned on X_t , and let $t \rightarrow \infty$ to get

$$\mathbb{E}[V(X_\infty, D_\infty)] \leq \rho_1 \mathbb{E}[V(X_\infty, D_\infty)] + \frac{K_1}{\sqrt{N}},$$

which implies that

$$\mathbb{E}[V(X_\infty, D_\infty)] \leq \frac{K_1}{(1 - \rho_1)\sqrt{N}}. \quad (4.144)$$

We combine (4.144) with the bound of $R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0)$ in terms of $V(X_\infty, D_\infty)$ in (4.132), and substitute β , L_W , K_{drift} , K_{conf} , K_{mono} , L_{cov} , and K_{cov} with their values. This leads to the final bound

$$R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0) \leq \frac{256r_{\max}\lambda_W^2|\mathbb{S}|^2}{\beta^2\sqrt{N}}. \quad (4.145)$$

We omit the detailed calculations to obtain (4.145). $\square$

4.G Extending results to restless bandits with inequality budget constraint

In this appendix, we consider the restless bandit problem with an inequality budget constraint, i.e.,

$$\begin{aligned} & \underset{\text{policy } \pi}{\text{maximize}} && R^-(\pi, \mathbf{S}_0) \end{aligned} \quad (4.9a)$$

$$\begin{aligned} & \text{subject to} && \sum_{i \in [N]} A_t^\pi(i) \leq \alpha N, \quad \forall t \geq 0. \end{aligned} \quad (4.9b)$$

As stated in Section 4.4.5, our policies can be adapted to this setting and have the same $O(1/\sqrt{N})$ optimality gap bounds. In the following, we describe the necessary modifications to the policies and the proofs.

4.G.1 Policy modifications

We modify the set-expansion policy and the ID policy for the inequality-constraint setting. As in the equality-constraint setting, we first solve the LP relaxation of problem (4.9), as given in (4.147) below. This LP relaxation is almost identical to the LP relaxation in the equality-constraint setting (LP), except that the budget constraint (4.147b) is now an inequality.

$$\begin{aligned} & \underset{\{y(s,a)\}_{s \in \mathbb{S}, a \in \mathbb{A}}}{\text{maximize}} && \sum_{s \in \mathbb{S}, a \in \mathbb{A}} r(s, a) y(s, a) \end{aligned} \quad (4.147a)$$

$$\text{subject to} \quad \sum_{s \in \mathbb{S}} y(s, 1) \leq \alpha, \quad (4.147b)$$

$$\sum_{s' \in \mathbb{S}, a \in \mathbb{A}} y(s', a) P(s', a, s) = \sum_{a \in \mathbb{A}} y(s, a), \quad \forall s \in \mathbb{S}, \quad (4.147c)$$

$$\sum_{s' \in \mathbb{S}, a' \in \mathbb{A}} y(s', a') = 1, \quad y(s, a) \geq 0, \quad \forall s \in \mathbb{S}, a \in \mathbb{A}. \quad (4.147d)$$

Given the optimal solution $(y^*(s, a))_{s \in \mathbb{S}, a \in \mathbb{A}}$ of (4.147), the optimal single-armed policy $\bar{\pi}^*$ is defined using (4.3). We assume that $\bar{\pi}^*$ induces an aperiodic unichain in the single-armed MDP.

The set-expansion policy for the inequality-constraint setting is defined in Algorithm 7, where recall that $\delta(x, D)$ is a function of system state x and subset $D \subseteq [N]$, given by

$$\delta(x, D) = \beta(1 - m(D)) - \frac{1}{2} \|x(D) - m(D)\mu^*\|_1,$$

Compared with the equality-constraint setting, the only change of the set-expansion policy here is in the action rectification step: As indicated in Lines 7–12 of Algorithm 7, here we can simply take $A_t(i) = \hat{A}_t(i)$ for all $i \in D_t$ as long as $\sum_{i \in D_t} \hat{A}_t(i) \leq \alpha N$.

The ID policy for the inequality-constraint setting is defined in Algorithm 8, where the only change is again in the action rectification step. As indicated in Lines 6–8 of Algorithm 8, here we let $N_t^{\bar{\pi}^*} = N$ and $A_t(i) = \hat{A}_t(i)$ for all $i \in [N]$ if $\sum_{i \in [N]} \hat{A}_t(i) \leq \alpha N$.

4.G.2 Proof modifications

Next, we restate and prove Theorem 4.6 and Theorem 4.7, which show that the optimality gaps for the set-expansion policy and the ID policy are

Algorithm 7 Set-expansion policy (inequality constraint)

Input: number of arms N , budget αN , the optimal single-armed policy $\bar{\pi}^*$, initial system state X_0 , initial state vector $\mathbf{S}_0$, initial focus set $D_{-1} = \emptyset$

```

1: for  $t = 0, 1, 2, \dots$  do
     $\triangleright$  Set update
2:   if  $\delta(X_t, D_{t-1}) > 0$  then
3:     Let  $D_t$  be any set with the largest  $m(D_t)$  such that  $D_t \supseteq D_{t-1}$ 
       and  $\delta(X_t, D_t) \geq 0$ 
4:   else
5:     Let  $D_t$  be any set with the largest  $m(D_t)$  such that  $D_t \subseteq D_{t-1}$ 
       and  $\delta(X_t, D_t) \geq 0$ 
   .....
    $\triangleright$  Action sampling
6:   Independently sample  $\hat{A}_t(i) \sim \bar{\pi}^*(\cdot | S_t(i))$  for  $i \in [N]$ 
    $\triangleright$  Action rectification
7:   if  $\sum_{i \in D_t} \hat{A}_t(i) \geq \alpha N$  then
8:     Select  $\alpha N$  arms in  $D_t$  with  $\hat{A}_t(i) = 1$  uniformly at random, and
       set  $A_t(i) = 1$ 
9:     For the rest of  $i \in [N]$ , set  $A_t(i) = 0$ 
10:  else
11:    Set  $A_t(i) = \hat{A}_t(i)$  for  $i \in D_t$ 
12:    Set  $A_t(i) = \hat{A}_t(i)$  for as many  $i \notin D_t$  as possible; break ties
       uniformly at random
13:  Apply  $A_t(i)$  and observe  $S_{t+1}(i)$  for each arm  $i \in [N]$ 

```

$O(1/\sqrt{N})$. Since most components of the proofs are identical to those for the equality-constraint setting, below we outline the proof steps and focus on the necessary modifications, omitting the parts of the proofs that remain unchanged.

Theorem 4.6. *Consider the N -armed restless bandit problem with the inequality constraint (4.9) and assume that the optimal single-armed policy $\bar{\pi}^*$ induces an aperiodic unichain. Let π be the set-expansion policy defined in Algorithm 7. For any N and initial state vector $\mathbf{S}_0$, the optimality gap of π is bounded as*

$$R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0) \leq \frac{C'_{\text{SE}}}{\sqrt{N}}, \quad (4.10)$$

Algorithm 8 ID policy (inequality constraint)

Input: number of arms N , budget αN , the optimal single-armed policy $\bar{\pi}^*$, initial system state X_0 , initial state vector $\mathbf{S}_0$

```

1: for  $t = 0, 1, 2, \dots$  do
     $\triangleright$  Action sampling
2:   Independently sample  $\hat{A}_t(i) \sim \bar{\pi}^*(\cdot | S_t(i))$  for  $i \in [N]$ 
     $\triangleright$  Action rectification
3:   if  $\sum_{i \in [N]} \hat{A}_t(i) \geq \alpha N$  then
4:      $N_t^{\bar{\pi}^*} \leftarrow \max\{n \leq N : \sum_{i=1}^n \hat{A}_t(i) \leq \alpha N\}$ 
5:      $A_t(i) \leftarrow \hat{A}_t(i)$  for  $i \leq N_t^{\bar{\pi}^*}$ ,  $A_t(i) \leftarrow 0$  for  $i > N_t^{\bar{\pi}^*}$ 
6:   else
7:      $N_t^{\bar{\pi}^*} \leftarrow N$ 
8:      $A_t(i) \leftarrow \hat{A}_t(i)$  for  $i \in [N]$ 
9:   Apply  $A_t(i)$  and observe  $S_{t+1}(i)$  for each arm  $i \in [N]$ 

```

for some constant C'_{SE} depending on $r_{\max}$, $|\mathbb{S}|$, $\beta \triangleq \min\{\alpha, 1 - \alpha\}$, and $P_{\bar{\pi}^*}$.

Theorem 4.7. Consider the N -armed restless bandit problem with the inequality constraint (4.9) and assume that the optimal single-armed policy $\bar{\pi}^*$ induces an aperiodic unichain. Let π be the ID policy defined in Algorithm 8. For any N and initial state vector $\mathbf{S}_0$, the optimality gap of π is bounded as

$$R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0) \leq \frac{C'_{\text{ID}}}{\sqrt{N}}, \quad (4.11)$$

for some constant C'_{ID} depending on $r_{\max}$, $|\mathbb{S}|$, $\beta \triangleq \min\{\alpha, 1 - \alpha\}$, and $P_{\bar{\pi}^*}$.

Now we describe the proofs of Theorem 4.6 and Theorem 4.7. First, it is not hard to check that the meta-theorem Theorem 4.9 and its proof still hold in the inequality-constraint setting. Intuitively, the proof of Theorem 4.9 does not directly utilize the specific form of the budget constraint. Instead, it is based on the properties of the subset Lyapunov functions and Conditions 1–3 that characterize the focus set.

Now we apply Theorem 4.9 to prove Theorem 4.6 for the set-expansion policy. We consider the same subset Lyapunov function as in the equality-constraint setting, recalled below:

$$h_W(x, D) = \|x(D) - m(D)\mu^*\|_W.$$

The fact that $\{h_W(x, D)\}_{D \subseteq [N]}$ satisfies the definition of subset Lyapunov functions does not rely on the budget constraint of the RB problem, as one can see by inspecting the proof of Lemma 4.13 in Appendix 4.E.2.

Next, it remains to verify the Conditions 1–3 for the set-expansion policy. We restate these conditions below for ease of reference.

Condition 1 (Majority conformity). *Let $K_{\text{conf}} > 0$ be a constant. For any $t \geq 0$, there exists $D'_t \subseteq D_t$ such that for any $i \in D'_t$, we have $A_t(i) = \hat{A}_t(i)$ and*

$$\mathbb{E}[m(D_t \setminus D'_t) \mid X_t, D_t] \leq \frac{K_{\text{conf}}}{\sqrt{N}} \quad a.s. \quad (4.148)$$

Condition 2 (Almost non-shrinking). *For any $t \geq 0$, either $D_{t+1} \supseteq D_t$ or $D_{t+1} \subseteq D_t$. Moreover, there exists a constant $K_{\text{mono}} > 0$ such that for any $t \geq 0$,*

$$\mathbb{E}[(m(D_t) - m(D_{t+1}))^+ \mid X_t, D_t] \leq \frac{K_{\text{mono}}}{\sqrt{N}} \quad a.s. \quad (4.149)$$

Condition 3 (Sufficient coverage). *There exists a class of subset Lyapunov functions $\{h(\cdot, D) : D \in \mathcal{D}\}$ and constants $L_{\text{cov}} > 0, K_{\text{cov}} > 0$ such that for any $t \geq 0$,*

$$1 - m(D_t) \leq L_{\text{cov}} h(X_t, D_t) + \frac{K_{\text{cov}}}{\sqrt{N}} \quad a.s. \quad (4.150)$$

Inspecting the proofs in Section 4.6, we can see that only the proof of Condition 1 could be affected by the form of the budget constraint. Intuitively, only Condition 1 is directly associated with the feasibility of actions, whereas the other two conditions are more closely related to the dynamics of the focus set.

In Lemma 4.31 stated below, we show Condition 1 for the set-expansion policy with the inequality constraint. The proof of this lemma is modified from the proof of Lemma 4.14.

Lemma 4.31 (Set-expansion policy with inequality constraint satisfies Condition 1). *Consider the set-expansion policy for restless bandits with the inequality constraint (Algorithm 7). For any $t \geq 0$, there exists a subset $D'_t \subseteq D_t$ such that for any $i \in D'_t$, $A_t(i) = \hat{A}_t(i)$, and*

$$\mathbb{E}[m(D_t \setminus D'_t) \mid X_t, D_t] \leq \frac{1}{\sqrt{N}} + \frac{1}{N} \quad a.s. \quad (4.151)$$

Proof of Lemma 4.14. Recall that under the set-expansion policy, the actions in the focus set $(A_t(i))_{i \in D_t}$ are chosen according to the ideal actions $(\hat{A}_t(i))_{i \in D_t}$ based on the following rules:

- When $\sum_{i \in D_t} \hat{A}_t(i) > \alpha N$, the set-expansion policy chooses αN arms in D_t with $\hat{A}_t(i) = 1$ and sets $A_t(i) = \hat{A}_t(i)$.
- Otherwise, the set-expansion policy sets $A_t(i) = \hat{A}_t(i)$ for all $i \in D_t$.

Let $D'_t = \{i \in D_t : A_t(i) = \hat{A}_t(i)\}$. Then we have

$$|D_t \setminus D'_t| = \left(\sum_{i \in D_t} \hat{A}_t(i) - \alpha N \right)^+. \quad (4.152)$$

Therefore, if we can show that for each $t \geq 0$,

$$\mathbb{E} \left[\left(\sum_{i \in D_t} \hat{A}_t(i) - \alpha N \right)^+ \mid X_t, D_t \right] \leq \sqrt{N}, \quad (4.153)$$

we will have $\mathbb{E}[m(D_t \setminus D'_t) \mid X_t, D_t] \leq 1/\sqrt{N}$, which will complete the proof. The remainder of the proof is dedicated to proving (4.153).

Observe that given X_t and D_t , the $\hat{A}_t(i)$'s are independent for $i \in D_t$. Consider the *scaled expected budget requirement* for arms in a set D , defined as

$$C_{\bar{\pi}^*}(x, D) \triangleq \frac{1}{N} \mathbb{E} \left[\sum_{i \in D} \hat{A}_t(i) \mid X_t = x \right] = \sum_{s \in \mathbb{S}} x(D, s) \bar{\pi}^*(1|s), \quad (4.154)$$

Then $\mathbb{E}[\sum_{i \in D_t} \hat{A}_t(i) \mid X_t, D_t] = NC_{\bar{\pi}^*}(X_t, D_t)$. By the Cauchy-Schwarz inequality,

$$\begin{aligned} & \mathbb{E} \left[\left| \sum_{i \in D_t} \hat{A}_t(i) - NC_{\bar{\pi}^*}(X_t, D_t) \right| \mid X_t, D_t \right] \\ & \leq \mathbb{E} \left[\left(\sum_{i \in D_t} \hat{A}_t(i) - NC_{\bar{\pi}^*}(X_t, D_t) \right)^2 \mid X_t, D_t \right]^{\frac{1}{2}} \\ & = \left(\sum_{i \in D_t} \text{Var}[\hat{A}_t(i) \mid X_t, D_t] \right)^{\frac{1}{2}} \\ & \leq \sqrt{N}. \end{aligned} \quad (4.155)$$

We next prove (4.153) utilizing (4.155). Recall that D_t satisfies $\delta(X_t, D_t) \geq 0$, i.e., $\|X_t(D_t) - m(D_t)\mu^*\|_1/2 \leq \beta(1 - m(D_t))$. Also, note that

$$\sum_{s \in \mathbb{S}} \mu^*(s) \bar{\pi}^*(1|s) = \sum_{s \in \mathbb{S}} y^*(s, 1) \leq \alpha. \quad (4.156)$$

Then $(C_{\bar{\pi}^*}(X_t, D_t) - \alpha m(D_t))^+$ can be bounded as

$$\begin{aligned} & \left(C_{\bar{\pi}^*}(X_t, D_t) - \alpha m(D_t) \right)^+ \\ & \leq \left(\sum_{s \in \mathbb{S}} (X_t(D_t, s) - m(D_t)\mu^*(s)) \bar{\pi}^*(1|s) \right)^+ \end{aligned} \quad (4.157)$$

$$\leq \left(\sum_{s \in \mathbb{S}} \left(X_t(D_t, s) - m(D_t)\mu^*(s) \right) \left(\bar{\pi}^*(1|s) - \frac{1}{2} \right) \right)^+ \quad (4.158)$$

$$\leq \frac{1}{2} \|X_t(D_t) - m(D_t)\mu^*\|_1 \quad (4.159)$$

$$\leq \beta(1 - m(D_t)), \quad (4.160)$$

where (4.158) is because $\sum_{s \in \mathbb{S}} (X_t(D_t, s) - m(D_t)\mu^*(s)) = 0$, and (4.159) is because $|\bar{\pi}^*(1|s) - 1/2| \leq 1/2$ for all $s \in \mathbb{S}$. Thus,

$$C_{\bar{\pi}^*}(X_t, D_t) \leq \alpha m(D_t) + \beta(1 - m(D_t)) \leq \alpha m(D_t) + \alpha(1 - m(D_t)) = \alpha.$$

Therefore, we have

$$\begin{aligned} & \mathbb{E} \left[\left(\sum_{i \in D_t} \hat{A}_t(i) - \alpha N \right)^+ \middle| X_t, D_t \right] \\ & = \mathbb{E} \left[\left(\sum_{i \in D_t} \hat{A}_t(i) - \alpha N \right)^+ \middle| X_t, D_t \right] \\ & \leq \mathbb{E} \left[\left| \sum_{i \in D_t} \hat{A}_t(i) - NC_{\bar{\pi}^*}(X_t, D_t) \right| \middle| X_t, D_t \right] \\ & \leq \sqrt{N}. \end{aligned} \quad (4.161)$$

This proves (4.153), concluding the proof of Lemma 4.31. $\square$

Now we bound the optimality gap for the ID policy in the inequality-constraint setting. As in the case of set-expansion policy, we only need to establish Condition 1. The proof of Condition 1 follows almost identical argument as Lemma 4.19.

Lemma 4.32 (ID policy with the inequality constraint satisfies Condition 1). *Consider the ID policy for restless bandits with inequality constraint (Algorithm 8). For any $t \geq 0$. Let $D'_t = [\min(N_t^{\bar{\pi}^*}, Nm_d(X_t))]$. Then*

$$\mathbb{E}[m(D_t \setminus D'_t) \mid X_t, D_t] = \frac{1}{N} \mathbb{E}[(Nm_d(X_t) - N_t^{\bar{\pi}^*})^+ \mid X_t] \leq \frac{2}{\beta\sqrt{N}} + \frac{1}{N} \quad a.s. \quad (4.162)$$

Proof of Lemma 4.32. In this proof, the variables n and n' are by default non-negative integers. We fix a time step $t \geq 0$, and condition on $X_t = x$ for a fixed system state realization x . We first prove a lower bound for $N_t^{\bar{\pi}^*}$ in terms of the *budget requirements* of the arms in $[n]$, $\sum_{i \in [n]} \hat{A}_t(i)$, for each $n \leq N$. Observe that the arms in $[n]$ can follow their ideal actions if and only if

$$\sum_{i \in [n]} \hat{A}_t(i) \leq \alpha N. \quad (4.163)$$

Consequently, a sufficient condition for the arms in $[n]$ to follow their ideal actions is

$$\left(\sum_{i \in [n]} \hat{A}_t(i) - \alpha n \right)^+ \leq \beta(N - n), \quad (4.164)$$

where recall that $\beta = \min\{\alpha, 1 - \alpha\}$. A further sufficient condition for (4.164) is

$$\max_{n' \leq n} \left(\sum_{i \in [n']} \hat{A}_t(i) - \alpha n' \right)^+ \leq \beta(N - n). \quad (4.165)$$

Therefore, by the definition of $N_t^{\bar{\pi}^*}$,

$$\begin{aligned} N_t^{\bar{\pi}^*} &\geq \max \left\{ n \leq N : \max_{n' \leq n} \left(\sum_{i \in [n']} \hat{A}_t(i) - \alpha n' \right)^+ \leq \beta(N - n) \right\} \\ &= N \max \left\{ m \in [0, 1]_N : \right. \\ &\quad \left. \max_{\substack{m' \in [0, 1]_N \\ m' \leq m}} \left(\frac{1}{N} \sum_{i \in [Nm']} \hat{A}_t(i) - \alpha m' \right)^+ \leq \beta(1 - m) \right\}. \end{aligned} \quad (4.166)$$

Our next step is to bound the quantity

$$\max_{m' \in [0, 1]_N, m' \leq m} \left(\frac{1}{N} \sum_{i \in [Nm']} \hat{A}_t(i) - \alpha m' \right)^+.$$

To do this, we relate it to the subset Lyapunov function $h_{\text{ID}}(x, m)$ and the scaled expected budget requirement, $C_{\bar{\pi}^*}(x, D)$, which is defined as:

$$C_{\bar{\pi}^*}(x, D) \triangleq \frac{1}{N} \mathbb{E} \left[\sum_{i \in D} \hat{A}_t(i) \middle| X_t = x \right] = x(D) c_{\bar{\pi}^*}^\top, \quad (4.167)$$

where $c_{\bar{\pi}^*}$ is the row vector $(\bar{\pi}^*(1|s))_{s \in \mathbb{S}}$. We first decompose the target quantity using the triangle inequality: for any $m \in [0, 1]_N$, we have

$$\begin{aligned} & \max_{\substack{m' \in [0, 1]_N \\ m' \leq m}} \left(\frac{1}{N} \sum_{i \in [Nm']} \hat{A}_t(i) - \alpha m' \right)^+ \\ & \leq \max_{\substack{m' \in [0, 1]_N \\ m' \leq m}} \left(\left(C_{\bar{\pi}^*}(x, [Nm']) - \alpha m' \right)^+ \right. \\ & \quad \left. + \left| \frac{1}{N} \sum_{i \in [Nm']} \hat{A}_t(i) - C_{\bar{\pi}^*}(x, [Nm']) \right| \right) \\ & \leq \max_{\substack{m' \in [0, 1]_N \\ m' \leq m}} \left(C_{\bar{\pi}^*}(x, [Nm']) - \alpha m' \right)^+ \\ & \quad + \max_{m' \in [0, 1]_N} \left| \frac{1}{N} \sum_{i \in [Nm']} \hat{A}_t(i) - C_{\bar{\pi}^*}(x, [Nm']) \right|, \end{aligned} \quad (4.168)$$

The second term on the right-hand side is a noise term that we will bound later. We now focus on bounding the first term, which captures the deviation of the instantaneous expected budget requirement from its steady-state expectation. Using the fact that $\alpha \geq \mu^* c_{\bar{\pi}^*}^\top$, we have:

$$\begin{aligned} (C_{\bar{\pi}^*}(x, [Nm']) - \alpha m')^+ & \leq ((x([Nm']) - m' \mu^*) c_{\bar{\pi}^*}^\top)^+ \\ & \leq |(x([Nm']) - m' \mu^*) c_{\bar{\pi}^*}^\top| \\ & = |(x([Nm']) - m' \mu^*) W^{1/2} W^{-1/2} c_{\bar{\pi}^*}^\top| \\ & \leq \|x([Nm']) - m' \mu^*\|_W \|c_{\bar{\pi}^*}\|_{W^{-1}} \\ & = K_{c/h} h_W(x, m'). \end{aligned} \quad (4.169)$$

Consequently, we have

$$\max_{\substack{m' \in [0, 1]_N \\ m' \leq m}} \left(C_{\bar{\pi}^*}(x, [Nm']) - \alpha m' \right)^+ \leq K_{c/h} \max_{\substack{m' \in [0, 1]_N \\ m' \leq m}} h_W(x, m') = K_{c/h} h_{\text{ID}}(x, m).$$

Using the monotonicity of $h_{\text{ID}}(x, m)$ in m and the definition of $m_d(x)$, we have that for any $m \leq m_d(x)$,

$$K_{c/h} h_{\text{ID}}(x, m) \leq K_{c/h} h_{\text{ID}}(x, m_d(x)) \leq \beta(1 - m_d(x)). \quad (4.170)$$

Substituting these bounds back into the decomposition in (4.168) implies the following bound for any $m \in [0, 1]_N$ such that $m \leq m_d(x)$:

$$\begin{aligned} & \max_{\substack{m' \in [0, 1]_N \\ m' \leq m}} \left(\frac{1}{N} \sum_{i \in [Nm']} \hat{A}_t(i) - \alpha m' \right)^+ \\ & \leq \beta(1 - m_d(x)) + \max_{m' \in [0, 1]_N} \left| \frac{1}{N} \sum_{i \in [Nm']} \hat{A}_t(i) - C_{\bar{\pi}^*}(x, [Nm']) \right|. \end{aligned} \quad (4.171)$$

To lower bound $N_t^{\bar{\pi}^*}$ utilizing (4.166) and (4.171), we want to find $\bar{m} \in [0, 1]_N$ such that $\bar{m} \leq m_d(x)$ and the right-hand side of (4.171) is at most $\beta(1 - \bar{m})$. For any such $\bar{m}$, (4.166) implies that

$$\bar{m} \in \left\{ m \in [0, 1]_N : \max_{\substack{m' \in [0, 1]_N \\ m' \leq m}} \left| \frac{1}{N} \sum_{i \in [Nm']} \hat{A}_t(i) - \alpha m' \right| \leq \beta(1 - m) \right\},$$

and then (4.171) implies that $N_t^{\bar{\pi}^*} \geq N \bar{m}$. Taking the largest such $\bar{m}$ yields

$$\begin{aligned} N_t^{\bar{\pi}^*} & \geq \min \left\{ Nm_d(x), \left\lfloor N - \frac{N}{\beta} \left(\beta(1 - m_d(x)) \right. \right. \right. \\ & \quad \left. \left. \left. + \max_{m' \in [0, 1]_N} \left| \frac{1}{N} \sum_{i \in [Nm']} \hat{A}_t(i) - C_{\bar{\pi}^*}(x, [Nm']) \right| \right) \right\rfloor \right\} \\ & \geq \min \left\{ Nm_d(x), N - \frac{N}{\beta} \left(\beta(1 - m_d(x)) \right. \right. \\ & \quad \left. \left. + \max_{m' \in [0, 1]_N} \left| \frac{1}{N} \sum_{i \in [Nm']} \hat{A}_t(i) - C_{\bar{\pi}^*}(x, [Nm']) \right| \right) - 1 \right\} \\ & = \min \left\{ Nm_d(x), Nm_d(x) - 1 \right. \\ & \quad \left. - \frac{1}{\beta} \max_{m' \in [0, 1]_N} \left| \sum_{i \in [Nm']} \hat{A}_t(i) - NC_{\bar{\pi}^*}(x, [Nm']) \right| \right\} \\ & = Nm_d(x) - 1 - \frac{1}{\beta} \max_{n' \leq N} \left| \sum_{i \in [n']} \hat{A}_t(i) - NC_{\bar{\pi}^*}(x, [n']) \right|. \end{aligned}$$

Rearranging the terms and taking the expectation, we get

$$\begin{aligned} & \mathbb{E} \left[(Nm_d(x) - N_t^{\pi^*})^+ \mid X_t = x \right] \\ & \leq 1 + \frac{1}{\beta} \mathbb{E} \left[\max_{n' \leq N} \left| \sum_{i \in [n']} \hat{A}_t(i) - NC_{\pi^*}(x, [n']) \right| \mid X_t = x \right]. \end{aligned} \quad (4.172)$$

With (4.172), it remains to prove

$$\mathbb{E} \left[\max_{n \leq N} \left| \sum_{i \in [n]} \hat{A}_t(i) - NC_{\pi^*}(x, [n]) \right| \mid X_t = x \right] \leq 2\sqrt{N}. \quad (4.173)$$

Our proof uses Doob's maximum inequality. First, we simplify the expression by defining the centered random variables $\xi(i) = \hat{A}_t(i) - \mathbb{E}[\hat{A}_t(i) \mid X_t = x]$. This allows us to rewrite the left-hand side of (4.173) as:

$$\mathbb{E} \left[\max_{n \leq N} \left| \sum_{i \in [n]} \hat{A}_t(i) - NC_{\pi^*}(x, [n]) \right| \mid X_t = x \right] = \mathbb{E} \left[\max_{n \leq N} \left| \sum_{i \in [n]} \xi(i) \right| \mid X_t = x \right]. \quad (4.174)$$

The sequence of partial sums $(\sum_{i \in [n]} \xi(i))_{n \in [N]}$ forms a martingale when conditioned on $X_t = x$. This holds because the ideal actions $\hat{A}_t(i)$ are sampled independently, making the $\xi(i)$ terms independent random variables, each with a conditional expectation of zero. Applying Cauchy-Schwarz and Doob's L_2 maximum inequality [see, e.g., 86, Theorem 4.4.4], we bound this term as:

$$\begin{aligned} \mathbb{E} \left[\max_{n \leq N} \left| \sum_{i \in [n]} \xi(i) \right| \mid X_t = x \right] & \leq \mathbb{E} \left[\max_{n \leq N} \left| \sum_{i \in [n]} \xi(i) \right|^2 \mid X_t = x \right]^{1/2} \\ & \leq \left(4 \mathbb{E} \left[\left| \sum_{i \in [N]} \xi(i) \right|^2 \mid X_t = x \right] \right)^{1/2} \\ & = \left(4 \sum_{i \in [N]} \mathbb{E} \left[\xi(i)^2 \mid X_t = x \right] \right)^{1/2} \\ & \leq 2\sqrt{N}. \end{aligned}$$

Here, the first inequality follows from Cauchy-Schwarz, the second applies Doob's maximal L_2 inequality, and the equality follows from the independence of the $\xi(i)$ terms. The final inequality holds because $|\xi(i)| = |\hat{A}_t(i) - \mathbb{E}[\hat{A}_t(i) \mid X_t = x]| \leq 1$. This completes the proof. $\square$

4.H Experimental details

In this appendix, we provide details for the experiments in Section 4.8. In Appendix 4.H.1, we discuss the implementation details of the set-expansion policy. Then in Appendix 4.H.2, we provide additional details of some performance comparison experiments in Section 4.8, including the simulation settings and the definitions of the MDPs. Next, in Appendix 4.H.3, we comment on the details of the experiments in Section 4.8.2 which investigate the probability that a random RB instance violates GAP. In Appendix 4.H.4, we investigate the effect of dense reward functions on the commonness of non-GAP instances and the performance of the index policies on the non-GAP instances. Finally, in Appendix 4.H.5, we conduct an additional experiment to investigate the differences between the set-expansion policy and the ID policy.

4.H.1 Implementation details of set-expansion policy

In this subsection, we discuss the implementation details of the set-expansion policy (Algorithm 4) that we use in our experiments. We first discuss the set-update step on Lines 2–5, where the focus-set D_t is updated. Then we discuss the action rectification step on Lines 7–15 which determines the actions based on the focus set and the ideal actions.

Set update step. Recall that $\delta(x, D)$ is given by $\delta(x, D) = \beta(1 - m(D)) - 0.5 \|x(D) - m(D)\mu^*\|_1$, and the set-expansion policy chooses D_t with the maximal cardinality such that

- $\delta(X_t, D_t) \geq 0$, and
- $D_t \supseteq D_{t-1}$ if $\delta(X_t, D_{t-1}) > 0$, or $D_t \subseteq D_{t-1}$ if $\delta(X_t, D_{t-1}) \leq 0$.

Due to the complexity of directly optimizing over the discrete variable D_t , we will first decide $X_t(D_t)$, and then find D_t based on $X_t(D_t)$. Specifically, when $\delta(X_t, D_{t-1}) \geq 0$, consider the following optimization problem, whose optimal solution (z^*, m^*) gives $(X_t(D_t), m(D_t))$ if we omit the integer effect:

$$\begin{aligned} & \underset{z \in \mathbb{R}^{|S|}}{\text{maximize}} && \sum_{s \in S} z(s) && (4.175a) \end{aligned}$$

$$\text{subject to} \quad X_t(D_{t-1}, s) \leq z(s) \leq X_t([N], s) \quad \forall s \in S \quad (4.175b)$$

$$\sum_{s \in S} z(s) = m \quad (4.175c)$$

$$\frac{1}{2} \sum_{s \in \mathbb{S}} |z(s) - \mu^*(s)m| \leq \beta(1 - m). \quad (4.175d)$$

Here, the constraints (4.175b) and (4.175c) ensure that each feasible solution (z, m) corresponds to a $(X_t(D), m(D))$ for some $D \supseteq D_{t-1}$, modulo the integer effect; the constraint (4.175d) ensures that $\delta(X_t, D) \geq 0$. To solve the above optimization problem, we can equivalently convert it to the following LP:

$$\begin{array}{ll} \text{maximize} & m \\ \mathbf{z}, \mathbf{d} \in \mathbb{R}^{|\mathbb{S}|}, m \in \mathbb{R} & \end{array} \quad (4.176a)$$

$$\text{subject to} \quad X_t(D_{t-1}, s) \leq z(s) \leq X_t([N], s) \quad \forall s \in \mathbb{S} \quad (4.176b)$$

$$\sum_{s \in \mathbb{S}} z(s) = m \quad (4.176c)$$

$$z(s) - \mu^*(s)m \leq d(s) \quad \forall s \in \mathbb{S} \quad (4.176d)$$

$$-z(s) + \mu^*(s)m \leq d(s) \quad \forall s \in \mathbb{S} \quad (4.176e)$$

$$\frac{1}{2} \sum_{s \in \mathbb{S}} d(s) \leq \beta(1 - m). \quad (4.176f)$$

Similarly, when $\delta(X_t, D_{t-1}) \leq 0$, we consider the optimization problem

$$\begin{array}{ll} \text{maximize} & \sum_{s \in \mathbb{S}} z(s) \\ \mathbf{z} \in \mathbb{R}^{|\mathbb{S}|} & \end{array} \quad (4.177a)$$

$$\text{subject to} \quad 0 \leq z(s) \leq X_t(D_{t-1}, s) \quad \forall s \in \mathbb{S} \quad (4.177b)$$

$$\sum_{s \in \mathbb{S}} z(s) = m \quad (4.177c)$$

$$\frac{1}{2} \sum_{s \in \mathbb{S}} |z(s) - \mu^*(s)m| \leq \beta(1 - m), \quad (4.177d)$$

where the only change is (4.177b), which ensures that each feasible solution (z, m) corresponds to a $(X_t(D), m(D))$ for some $D \subseteq D_{t-1}$. This optimization problem is equivalent to the LP given by

$$\begin{array}{ll} \text{maximize} & m \\ \mathbf{z}, \mathbf{d} \in \mathbb{R}^{|\mathbb{S}|}, m \in \mathbb{R} & \end{array} \quad (4.178a)$$

$$\text{subject to} \quad 0 \leq z(s) \leq X_t(D_{t-1}, s) \quad \forall s \in \mathbb{S} \quad (4.178b)$$

$$\sum_{s \in \mathbb{S}} z(s) = m \quad (4.178c)$$

$$z(s) - \mu^*(s)m \leq d(s) \quad \forall s \in \mathbb{S} \quad (4.178d)$$

$$-z(s) + \mu^*(s)m \leq d(s) \quad \forall s \in \mathbb{S} \quad (4.178e)$$

$$\frac{1}{2} \sum_{s \in \mathbb{S}} d(s) \leq \beta(1 - m). \quad (4.178f)$$

Note that each of the two LPs has $2|\mathbb{S}| + 1$ variables and $4|\mathbb{S}| + 2$ constraints, so the complexities of solving them are polynomials in $|\mathbb{S}|$ and independent of N .

After obtaining the optimal solution $(\mathbf{z}^*, m^*)$ of either of the LPs above, we pick $\lfloor N\mathbf{z}^*(s) \rfloor$ arms in state s for each $s \in \mathbb{S}$ to form the subset D_t .

Note that because we first solve the LP and then perform the rounding, the resulting D_t may not be the exact optimal solution that maximizes the cardinality as required by Lines 2–5 of Algorithm 4. A more rigorous alternative is to include the integer constraints $N\mathbf{z}^*(s) \in \mathbb{Z}$ for $s \in \mathbb{S}$ in the LPs. Nevertheless, our current implementation is more efficient and has demonstrated good performance in the simulations. One can also verify that an $O(1/\sqrt{N})$ optimality gap can still be achieved using this approximate implementation, since intuitively, the maximality of $m(D_t)$ is still approximately preserved up to a negligible $O(1/N)$ error.

Action rectification step. For the vanilla version of the set-expansion policy, the action rectification step has been completely specified in Algorithm 4, where we prioritize the arms in D_t over those in D_t^c to follow the ideal actions, breaking ties uniformly at random.

For the version of the set-expansion policy that utilizes the LP index policy for tie-breaking, its pseudocode is given in Algorithm 9. To summarize the differences, if not all arm in D_t can follow the ideal actions, this version of the set-expansion policy breaks ties using LP indices instead of uniformly at random; if all arms in D_t can follow the ideal actions, this version of the set-expansion policy invokes the LP index policy to allocate the remaining budget to the arms not in D_t .

4.H.2 Additional details of the performance comparison experiments in Section 4.8

Next, we provide some additional details of the performance comparison experiments in Section 4.8. We first talk about the simulation settings and

Algorithm 9 Set-expansion policy (with LP index)

Input: number of arms N , budget αN , the optimal single-armed policy $\bar{\pi}^*$, initial system state X_0 , initial state vector $\mathbf{S}_0$, initial focus set $D_{-1} = \emptyset$

- 1: **for** $t = 0, 1, 2, \dots$ **do**
 - ▷ *Set update*
- 2: **if** $\delta(X_t, D_{t-1}) > 0$ **then**
- 3: Let D_t be any set with the largest $m(D_t)$ such that $D_t \supseteq D_{t-1}$ and $\delta(X_t, D_t) \geq 0$
- 4: **else**
- 5: Let D_t be any set with the largest $m(D_t)$ such that $D_t \subseteq D_{t-1}$ and $\delta(X_t, D_t) \geq 0$

▷ *Lines below implement Lines 2–6 of Algorithm 3*

▷ *Action sampling*

- 6: Independently sample $\hat{A}_t(i) \sim \bar{\pi}^*(\cdot | S_t(i))$ for $i \in [N]$
 - ▷ *Action rectification*
 - 7: **if** $\sum_{i \in D_t} \hat{A}_t(i) \geq \alpha N$ **then**
 - 8: Select αN arms in D_t with $\hat{A}_t(i) = 1$ to set $A_t(i) = 1$; break ties favoring larger LP indices
 - 9: For the rest of $i \in [N]$, set $A_t(i) = 0$
 - 10: **else if** $\sum_{i \in D_t} \hat{A}_t(i) \leq \alpha N - (N - |D_t|)$ **then**
 - 11: Select $(1 - \alpha)N$ arms in D_t with $\hat{A}_t(i) = 0$ to set $A_t(i) = 0$; break ties favoring smaller LP indices
 - 12: For the rest of $i \in [N]$, set $A_t(i) = 1$
 - 13: **else**
 - 14: Set $A_t(i) = \hat{A}_t(i)$ for $i \in D_t$
 - 15: Apply the LP index policy to the arms in D_t^c with $(\alpha N - \sum_{i \in D_t} \hat{A}_t(i))$ units of budget
 - 16: Apply $A_t(i)$ and observe $S_{t+1}(i)$ for each arm $i \in [N]$
-

the computation of the confidence interval. Then we include the definitions of the two RB instances in Section 4.8.1.

Simulation setting and output analysis. When simulating most of the RB problems and the policies, we run 5 independent replications for each N . The initial state of each arms in each replication is independently sampled from the uniform distribution over the state space. Each replication runs for 2×10^4 time steps. The exceptions are the simulations of the FTVA policy on the two non-SA examples in Figure 4.9, where we run the simulations for 1.6×10^5 time steps in each replication.

We compute the confidence interval of the average reward using the batch means method, a common method for computing the confidence intervals in steady-state simulations. Specifically, for each RB problem, policy, and number of arms N , we divide the sample path of each replication into 4 intervals of equal lengths and compute the sample mean of the rewards within each interval. As a result, we get 20 sample means from the 5 replications. Then we further average the 20 sample means and use it as the estimation of the long-run average reward. The confidence interval for the estimate is calculated using the variance of the 20 sample means, under the assumption that each interval is long enough for the system to mix to the steady state, resulting in nearly independent sample means. We refer the reader to the textbook [17] for more backgrounds on the batch means method.

Definition of the example in Figure 4.5a. The RB instance in Figure 4.5a has been given in Example 2 in Appendix E.2 of [105] and Appendix G.1 of [149]. Nevertheless, we include it here for completeness. The single-armed MDP of this RB instance has three states, whose transition kernel is given by

$$P(\cdot, 0, \cdot) = \begin{bmatrix} 0.02232142 & 0.10229283 & 0.87538575 \\ 0.03426605 & 0.17175704 & 0.79397691 \\ 0.52324756 & 0.45523298 & 0.02151947 \end{bmatrix},$$

$$P(\cdot, 1, \cdot) = \begin{bmatrix} 0.14874601 & 0.30435809 & 0.54689589 \\ 0.56845754 & 0.41117331 & 0.02036915 \\ 0.25265570 & 0.27310439 & 0.4742399 \end{bmatrix},$$

where the s -th row and s' -th column in each matrix above represents $P(s, 0, s')$ or $P(s, 1, s')$ for each pair of $s, s' \in \mathbb{S}$. The reward function of the single-armed MDP is given by

$$\begin{aligned} r(\cdot, 0) &= \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}, \\ r(\cdot, 1) &= \begin{bmatrix} 0.37401552 & 0.11740814 & 0.07866135 \end{bmatrix}. \end{aligned}$$

The budget parameter α is 0.4 for this RB instance, that is, $0.4N$ arms are activated in each time step.

Definition of the example in Figure 4.5b. The RB instance in Figure 4.5b is a modification of the example provided in Section 3.3 of [149]. Specifically, let the single-armed MDP have the state space $\mathbb{S} = \{0, 1, \dots, 7\}$. Each state is associated with a *preferred action*, which is action 1 for states in $\{0, 1, 2, 3\}$, and action 0 for the other states. If an arm is in state s and takes the preferred action, it moves to state $(s + 1) \bmod 8$ with probability $p_{s,R}$, and stays in state s otherwise; if it does not take the preferred action, it moves to state $(s - 1)^+$ with probability $p_{s,L}$. Here, the subscript L (resp., R) represents “left” (resp., “right”), and we are imagining the states of the single-armed MDP being lined up in a row from state 0 to state 7; by taking the preferred actions, the arm moves to the right and loops back to the state 0 after passing the state 7. The probabilities $p_{\cdot,R}$ and $p_{\cdot,L}$ are given by

$$\begin{aligned} p_{\cdot,R} &= \begin{bmatrix} 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \end{bmatrix}, \\ p_{\cdot,L} &= \begin{bmatrix} 1.0 & 1.0 & 0.48 & 0.47 & 0.46 & 0.45 & 0.44 & 0.43 \end{bmatrix}. \end{aligned}$$

The reward function is given by $r(7, 0) = 0.1$, $r(0, 1) = 1/300$, and $r(s, a) = 0$ for all other $s \in \mathbb{S}, a \in \mathbb{A}$. The budget parameter $\alpha = 1/2$, so $N/2$ arms are activated in each time step.

Compared to the RB instance in Section 3.3. of [149], the only modification in this RB instance is changing $r(0, 1)$ from 0 to $1/300$. This small modification makes this instance non-indexable for the Whittle index policy, while keeping the performances of the other policies almost unchanged.

4.H.3 Generating and identifying non-GAP examples in Section 4.8.2

In this subsection, we first provide some details on generating random RB instances following the symmetric Dirichlet distribution in Section 4.8.2. Then

we comment on how we identify the non-GAP instances through the notion of local instability.

Generating Dirichlet random examples. Each of the random RB examples in Figure 4.6 is generated as follows. First, for each state-action pair (s, a) , we sample the vector $(P(s, a, s'))_{s' \in \mathbb{S}}$ independently from a symmetric Dirichlet distribution. Similarly, for each action a , we sample the vector $(r(s, a))_{s \in \mathbb{S}}$ independently from the same Dirichlet distribution. To highlight the sparsity, we then set to zero any entries of $(P(s, a, s'))_{s' \in \mathbb{S}}$ and $(r(s, a))_{s \in \mathbb{S}}$ that are less than 10^{-7} . Finally, we sample the budget parameter α from the uniform distribution over the interval $[0.1, 0.9]$ and round it down to the nearest multiple of 0.01.

Identifying local instability. To identify a non-GAP example, we use the notation of *local instability*. To avoid ambiguity, we only consider local instability for the RB problems that satisfy the three conditions:

1. The optimal solution to the LP relaxation (LP), y^* , is unique;
2. There are no transient states for y^* , that is, $y^*(s, 1) + y^*(s, 0) > 0$ for all $s \in \mathbb{S}$;
3. The RB problem is non-degenerate, that is, there exists $\tilde{s} \in \mathbb{S}$ with $y^*(\tilde{s}, 1) > 0$ and $y^*(\tilde{s}, 0) > 0$.

Note that the state $\tilde{s}$ in the third condition must be unique, due to properties of this LP's basic feasible solutions [108, Proposition 2]. Under these assumptions, the mean-field dynamics around μ^* is locally linear and is the same for all LP-Priority policies. Specifically, we have

$$\mathbb{E}[X_{t+1}([N]) - \mu^* \mid X_t([N])] = (X_t([N]) - \mu^*)\Phi, \quad (4.179)$$

given that $X_t([N])$ is sufficiently close to μ^* , and an LP-Priority policy is used; the matrix Φ is defined as

$$\Phi \triangleq P_{\bar{\pi}^*} - \mathbf{1}^\top \mu^* - (c_{\bar{\pi}^*} - \alpha \mathbf{1})^\top (P_1(\tilde{s}) - P_0(\tilde{s})), \quad (4.180)$$

where $c_{\bar{\pi}^*} \triangleq (\bar{\pi}^*(1|s))_{s \in \mathbb{S}}$ and $P_a(\tilde{s}) \triangleq (P(\tilde{s}, a, s))_{s \in \mathbb{S}}$ are both row vectors; $\mathbf{1}$ is the all-one row vector. We refer the readers to Appendix B of [107] for a detailed derivation of the dynamics under LP-Priority policies.

In the context of our experiments, we say an RB instance is *locally unstable* if the matrix Φ defined in (4.180) is unstable, i.e., its spectral radius is strictly

larger than 1. This local instability of an RB instance implies that under any LP-Priority policy, the system's mean-field dynamics will drift away from the optimal state distribution μ^* whenever the system gets close to μ^* . This causes the GAP assumption to be violated under any LP-Priority policy.

4.H.4 Commonness of non-GAP examples when the reward functions are not sparse

In Section 4.8.2, we have shown that a certain type of non-GAP RB instances, i.e., the locally unstable instances, are common when the transition kernel and the reward functions are generated from some sparse distribution. We have also investigated the performance of the LP-index policy and the Whittle index policy on the locally unstable instances. Here, we include further experimental results to show that the sparsity of the reward function may not be essential to the phenomenon observed in Section 4.8.2.

Specifically, we consider random RB instances with $|\mathbb{S}| = 10$, where the rows of the transition kernel $P(s, a, \cdot)$ for $s \in \mathbb{S}, a \in \mathbb{A}$ are independently sampled from the symmetric Dirichlet distribution. Different from Section 4.8.2, we independently sample each entry of the reward functions from the uniform distribution over $[0, 1]$, so that the non-zero entries in the reward functions are more dense. We make two scatter plots in Figure 4.12, each of which visualizes 10^4 such random examples with a different parameter for the Dirichlet distribution. Each point in the scatter plots represents an RB instance, whose y -coordinate is the spectral radius of the matrix Φ (defined in (4.180)) and whose x -coordinate is the second-largest absolute value of P_{π^*} 's eigenvalues.

As we can see from Figure 4.12, the scatter plots look almost the same as those in Figure 4.6. Moreover, in Figure 4.12b, among 8965 instances that are aperiodic unichain whose x -coordinates are less than 0.95 (i.e., have a decently large spectral gap), 1828 instances are also locally unstable. So the ratio of locally unstable instances is approximately 20.4%, very close to the ratio 20.2% in Figure 4.6c. These phenomena suggest that changing the distribution for generating the reward function may not have a significant effect on the fraction of locally unstable instances.

In Figure 4.13, we plot some CDF curves representing the optimality ratios of the LP index policy, the Whittle index policy, and their maximal performances when $N = 500$, among 2113 locally unstable instances with transition kernels sampled from Dirichlet(0.05) and reward functions sampled

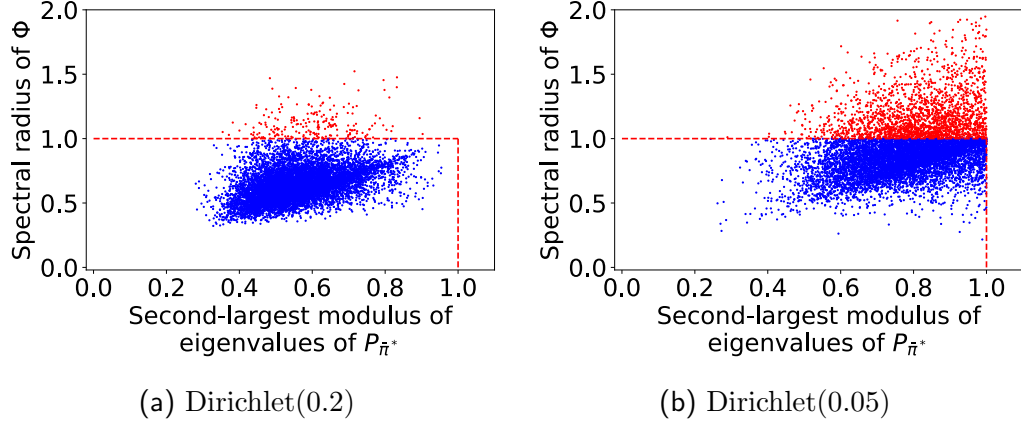


Figure 4.12: Scatter plots illustrating the eigenvalues of about 10^4 random RB problems, each with $|\mathbb{S}| = 10$. The transition probabilities of the instances follow the symmetric Dirichlet distribution with different parameters in each subplot. The entries of the reward functions are sampled from *the uniform distribution over $[0, 1]$* . The x - and y -axis values and the colors are interpreted in the same way as in Figure 4.6.

from the uniform distribution. Each policy is simulated for 2×10^4 time steps for each RB instance.

We compare the CDF curves in Figure 4.13 with those in Figure 4.7 to see the effect of the dense reward functions on the performances of the index policies. The shapes of the CDF curves look similar, but the optimality gap ratios are slightly better with dense reward functions. In particular, there are about 3.9% locally unstable instances where the average rewards of both policies are less than 90% of the LP upper bound, which is less than the 6.7% in Figure 4.7. Intuitively, the consequences of suboptimal decisions are less severe when the reward function is dense than when the reward function is sparse.

4.H.5 Comparing set-expansion policy with ID policy

In this subsection, we make an additional observation about different behaviors of the set-expansion policy and the ID policy: under the ID policy, a larger subset of arms can persistently follow the optimal single-armed policy $\bar{\pi}^*$ for a long period of time than under the set-expansion policy. Although this

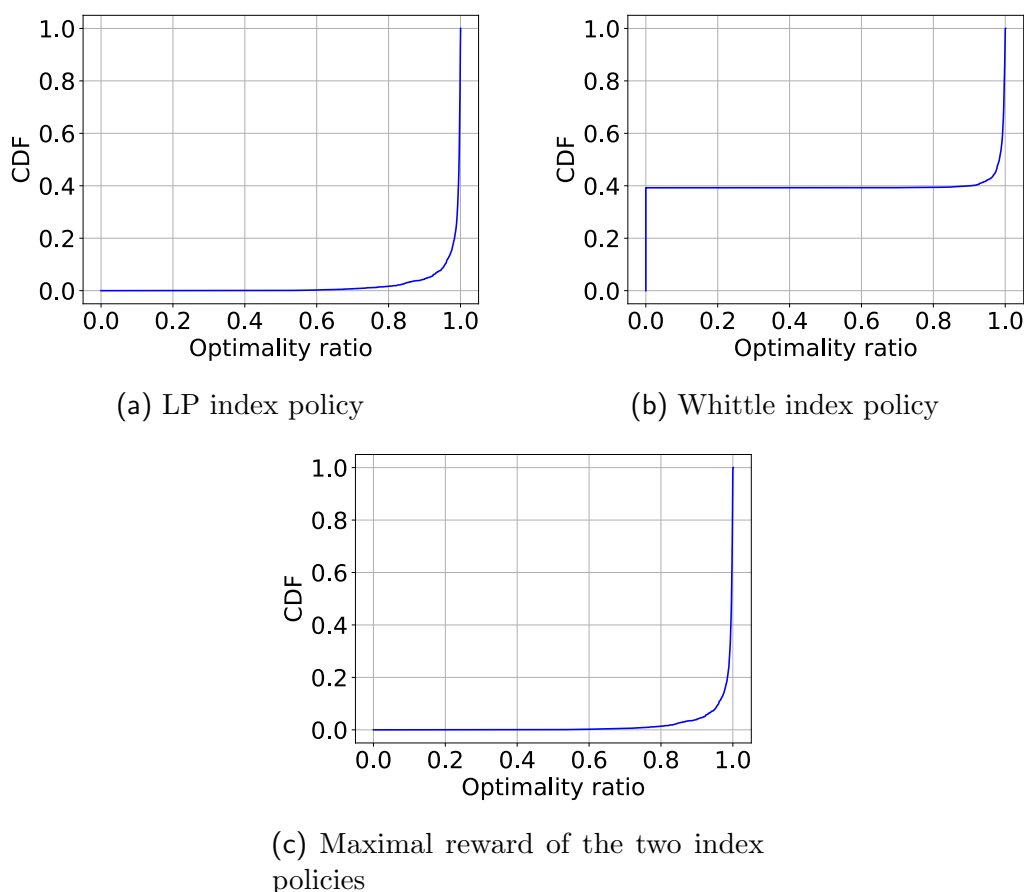


Figure 4.13: CDF of the optimality ratios of LP-Priority policies when $N = 500$, among 2113 non-GAP instances with transition kernels following Dirichlet(0.05) and reward functions following the uniform distribution. We regard the average reward of Whittle index policy as 0 if it is not well-defined.

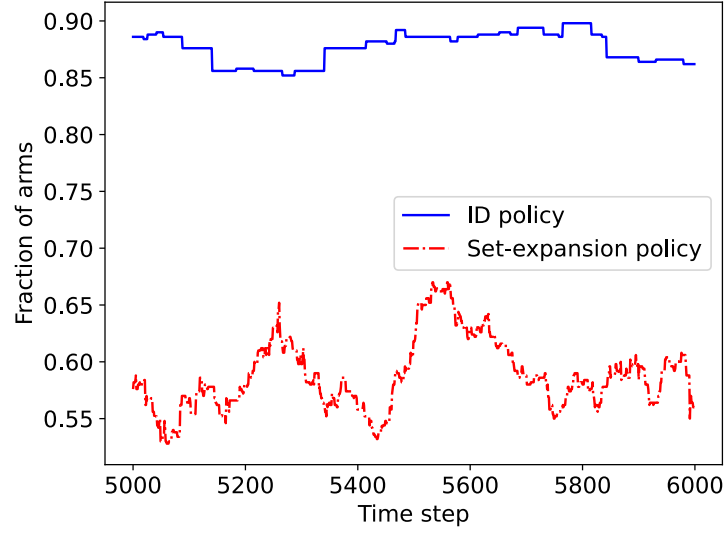
phenomenon does not necessarily imply that the ID policy performs better than the set-expansion policy, we include it here for its potential theoretical interests.

The observation comes from the following experiment. We run the ID policy and the set-expansion policy on each of the six RB problems simulated in Section 4.8 with $N = 500$. For each of the two policies, we first let them run 5000 time steps to mix to the steady state. Then for each of the next 1000 time step t , we plot the fraction of arms that whose actions agree with

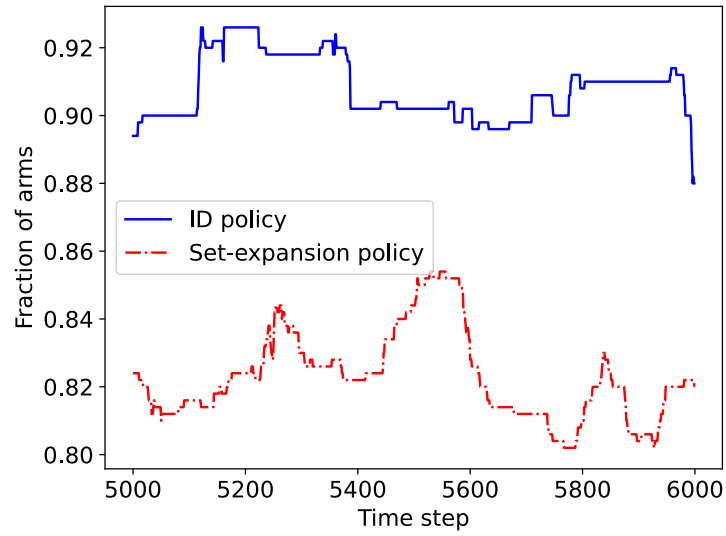
ideal actions in the time interval $[t, t + 199]$; these are the arms that will persistently follow $\bar{\pi}^*$ for 200 time steps starting from time t . We choose the look-ahead window to be 200 because it is large enough for an arm following $\bar{\pi}^*$ to converge to the stationary distribution: notice that by Lemma 4.12, the W -weighted L_2 distance between any distribution on $\mathbb{S}$ and the optimal stationary distribution μ^* reduces by a ratio of at most $1 - 1/(2\lambda_W)$ every time step under $\bar{\pi}^*$, and $2\lambda_W$ ranges from 2.82 to 84.29 in each of the examples simulated in Section 4.8.

The simulation results are shown in Figures 4.14 to 4.16. As we can see, the numbers of arms that can persistently follow $\bar{\pi}^*$ for 200 time steps under the ID policy are clearly larger than those under the set-expansion policy in all the examples.

Here is a plausible explanation for this phenomenon. Because the set-expansion policy randomly selects arms outside the focus set to follow $\bar{\pi}^*$, only those arms in the focus set could persistently follow $\bar{\pi}^*$. Moreover, the focus set of the set-expansion policy is determined according to the L_1 norm constraint, which is somewhat conservatively designed to facilitate the analysis. In contrast, the ID policy relies on a fixed set of IDs rather than the explicitly computed focus sets to decide the subset of arms that follow $\bar{\pi}^*$. As a result, some arms outside the focus set $[Nm_d(X_t)]$ that are considered in the analysis may also follow $\bar{\pi}^*$ for long periods of times under the ID policy. Consequently, more arms could follow $\bar{\pi}^*$ persistently under the ID policy than under the set-expansion policy.

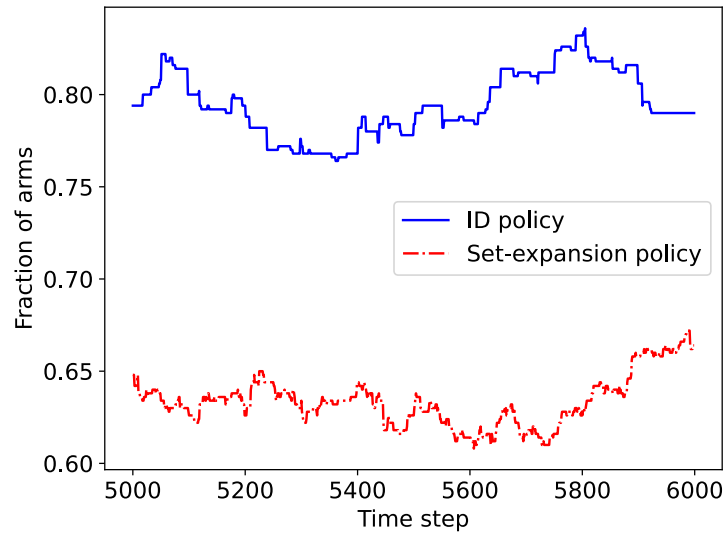


(a) The example in Figure 4.5a

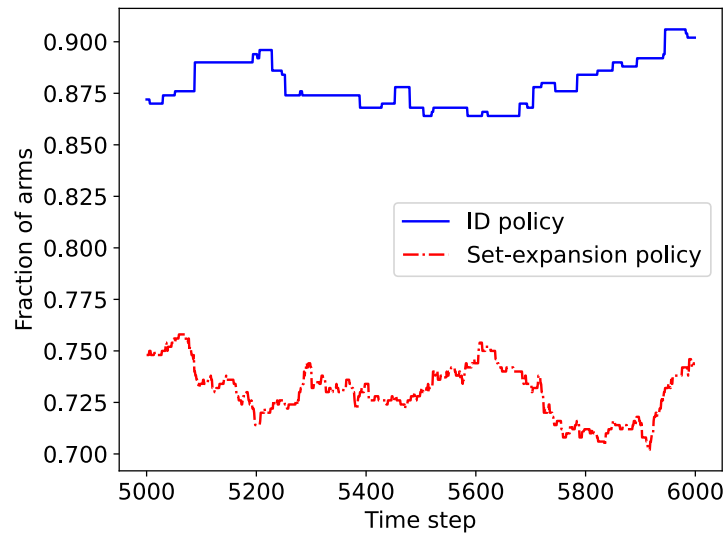


(b) The example in Figure 4.5b

Figure 4.14: Illustration of the fractions of arms that persistently follow $\bar{\pi}^*$ for the two RB problems simulated in Section 4.8.1.

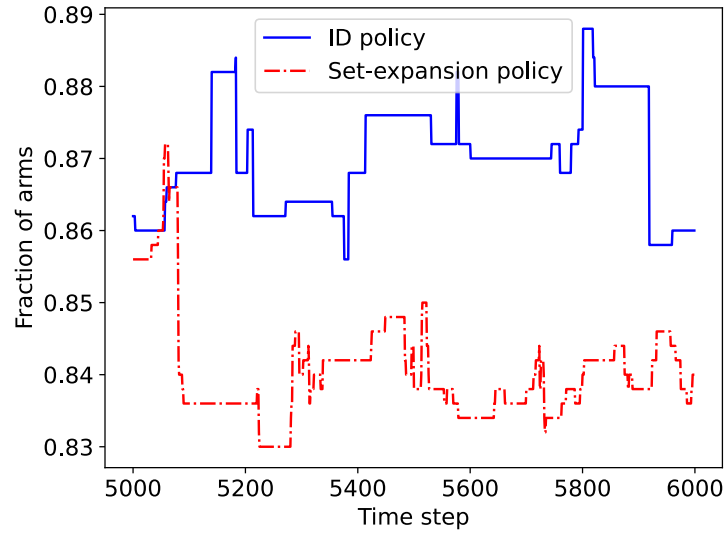


(a) The example in Figure 4.8a

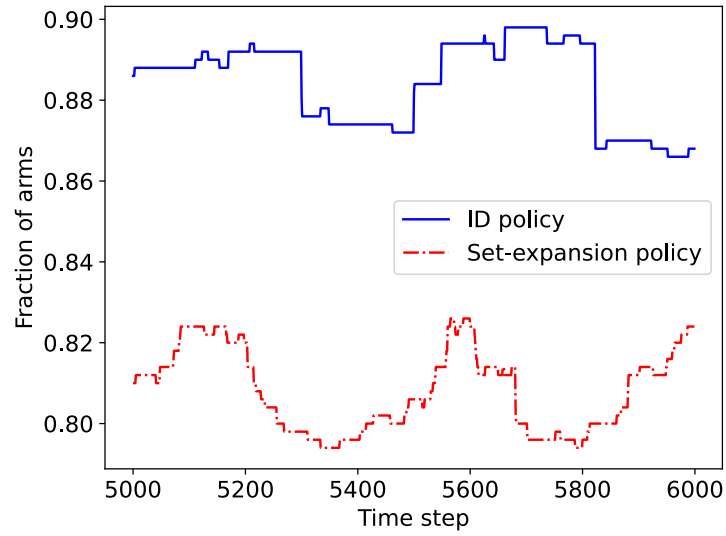


(b) The example in Figure 4.8b

Figure 4.15: Illustration of the fractions of arms that persistently follow $\bar{\pi}^*$ for the two RB problems simulated in Section 4.8.2.



(a) The example in Figure 4.9a



(b) The example in Figure 4.9b

Figure 4.16: Illustration of the fractions of arms that persistently follow $\bar{\pi}^*$ for the two RB problems simulated in Section 4.8.3.

Chapter 5

Achieving Exponential Asymptotic Optimality in Average-Reward Restless Bandits without Global Attractor Assumption

5.1 Introduction

In the previous two chapters, we have weakened the conditions for achieving asymptotic optimality. In this chapter, we seek to strengthen the notion of asymptotic optimality. A recent milestone on this direction is the *exponential asymptotic optimality* established in [107, 108], where a class of policies called LP-Priority is proved to achieve an optimality gap $O(\exp(-CN))$ for a constant C , under some assumptions. This exponential gap is particularly noteworthy because it “beats” the Central Limit Theorem (CLT), an intriguing theoretical property that is not a priori obvious to be possible. In particular, at a high level, all asymptotic optimality results in the RB literature are based on the concentration of the empirical distribution of the arms’ states around a certain optimal distribution. Therefore, an $O(1/\sqrt{N})$ bound owing to CLT was believed to be fundamental. An important observation made in [107, 108] is that if the system has a certain notion of *local linearity*, the

optimality gap only depends on the distance between the optimal distribution and the *expected* empirical state distribution, allowing one to go beyond CLT.

Despite the above intuition, it remains unclear what is the fundamental mechanism leading to exponential asymptotic optimality. Although a simple assumption called non-degeneracy (or non-singularity) suffices to ensure the local linearity property described above, non-degeneracy alone is not enough to guarantee exponential optimality of the LP-Priority policies considered in [107, 108]. In particular, another crucial assumption is needed, namely the Uniform Global Attractor Property (UGAP). UGAP is a stronger version of the global attractor property (GAP) — the latter is assumed in all asymptotic analyses of LP-Priority policies, without which an LP-Priority may even have a constant optimality gap [see 105, 149, 274, for concrete examples]. Unfortunately, UGAP is hard to interpret and verify, as it concerns the global convergence of a certain non-linear difference equation.

In light of the difficulty caused by UGAP, very recent work has studied new policies that are asymptotically optimal without assuming UGAP [149, 153, 294]. Unlike LP-Priority and Whittle index policies, these new policies actively control the empirical distribution of the states of the arms, driving the distribution to *globally converge* towards a certain optimal distribution, without relying on extraneous assumptions. However, while these results conclude that UGAP is not needed for asymptotic optimality, it remains unclear whether UGAP is fundamental for exponential asymptotic optimality — the best optimality gap proved in this line of work is $O(1/\sqrt{N})$.

Given these developments, it is natural to ask if removing UGAP necessarily comes at the cost of degrading the optimality gap. Is it possible to efficiently find policies that achieve *exponential asymptotic optimality* without assuming UGAP?

Our contributions

This setting continues to focus on the discrete-time setting. The primary contribution of this chapter is to design the first policy, the *two-set policy*, that achieves exponential asymptotic optimality without UGAP. The two-set policy maintains two dynamic subsets of arms and applies two different subroutines to the subsets: one subset of arms has a nearly optimal state distribution and takes actions according to a certain LP-Priority; the other subset of arms is driven towards the optimal state distribution and gradually merged into the first subset. Intuitively, the LP-Priority in the first subset induces local

linearity, while the second subset enables global convergence. While the ideas in each subroutine are not completely new, the two-set policy leverages them in a novel way to achieve the best of both worlds. The theoretical guarantee of the two-set policy is informally summarized in the theorem below.

Theorem (Informal version of Theorem 5.12). *Suppose each arm is unichain, aperiodic, non-degenerate, and locally stable. Then the two-set policy π satisfies*

$$R^{\text{rel}} - R(\pi, \mathbf{S}_0) = O(\exp(-CN)), \quad (5.1)$$

where R^{rel} is an upper bound on the optimal long-run average reward, and $R(\pi, \mathbf{S}_0)$ is the long-run average reward under the policy π given the vector of initial states $\mathbf{S}_0$.

We will make precise the definitions of non-degeneracy and local stability. We emphasize that we only need aperiodic unichain under a particular policy rather than all policies. The assumptions of Theorem 5.12 are a subset of, and hence strictly weaker than, the assumptions made in prior work [107, 108] for exponential asymptotic optimality. Moreover, all our assumptions pertain to a linear program defined by problem primitives and are thus much easier to verify than UGAP.

The algorithmic ideas of the two-set policy share some common traits with those in the “set-expansion” policy in Chapter 4 [153] and the “align and steer” policy in [294]. However, these two policies have only been shown to have $O(1/\sqrt{N})$ and $o(1)$ optimality gaps, respectively.

The proof of Theorem 5.12 employs a novel multivariate Lyapunov function technique, which generalizes the focus-set approach in Chapter 4 [153]. This multivariate Lyapunov function enables us to decouple the complex dynamics under the two-set policy, where the states of the arms and the two dynamic subsets are coupled and change simultaneously.

We next elucidate the roles of our assumptions in achieving exponential optimality. The importance of unichain, aperiodicity, and non-degeneracy is well recognized in the literature; we will complement the discussion with concrete examples. We further establish the fundamental importance of local stability by proving the following theorem.

Theorem (Informal version of Theorem 5.15). *For regular unstable RBs, every policy π satisfies*

$$R^{\text{rel}} - R(\pi, \mathbf{S}_0) = \Omega(1/\sqrt{N}). \quad (5.2)$$

We will formally define the class of regular unstable RBs and provide concrete examples. Note that all known efficient policies rely on the upper bound R^{rel} . For these policies, Theorem 5.15 shows that local stability is crucial for achieving exponential asymptotic optimality for a natural class of RB instances. The proof of Theorem 5.15 uses Lyapunov analysis to reveal an interesting trade-off between two competing needs when local stability fails: to keep the states of the arms near the optimal distribution, and to take the optimal actions when the states are near the optimal distribution.

Chapter organization

This chapter is organized as follows. We first set up the restless bandit problem in Section 5.2. Then we introduce our assumptions and the two-set policy in Section 5.3. Next, in Section 5.4.1, we prove that the two-set policy achieves exponential asymptotic optimality under our assumptions. Subsequently, in Section 5.4.2, we show that these assumptions are necessary for any policy to achieve exponential asymptotic optimality. Then we provide a proof sketch for the performance guarantee of the two-set policy in Section 5.5. Finally, we present numerical experiments in Section 5.6 and conclude in Section 5.7. The details of the proofs, examples, and experiments are provided in the appendices.

5.2 Problem setup

We restate the restless bandit problem, with the set of notation specific to this chapter. Consider the discrete-time average-reward restless bandit problem with N homogeneous arms, referred to as the N -armed problem. Index the arms by $[N] \triangleq \{1, 2, \dots, N\}$. To avoid confusion with the Whittle index, we refer to i as the *ID* of Arm i . Each arm is associated with a Markov decision process (MDP) defined by $(\mathbb{S}, \mathbb{A}, P, r)$, which is called the single-armed MDP. Here $\mathbb{S}$ is a finite state space; $\mathbb{A} = \{0, 1\}$ is the action space, where action 1 is activating/pulling the arm; $P : \mathbb{S} \times \mathbb{A} \times \mathbb{S} \rightarrow [0, 1]$ is the transition kernel; $r : \mathbb{S} \times \mathbb{A} \rightarrow \mathbb{R}$ is the reward function. Let $r_{\max} = \max_{s \in \mathbb{S}, a \in \mathbb{A}} |r(s, a)|$. The RB problem has a *budget constraint*, given by a constant $\alpha \in (0, 1)$, which requires that α fraction of arms must be pulled at every time step. We assume that αN is an integer. We focus on the setting where all the model parameters, $\mathbb{S}, \mathbb{A}, P, r, \alpha$, are known.

A policy π for the N -armed problem chooses the action for each of the N arms in each time step. The policy can be randomized and history-dependent. Under π , the *state vector* $\mathbf{S}_t^\pi \triangleq (S_t^\pi(i))_{i \in [N]} \in \mathbb{S}^N$ represents the states of all arms at time t , and the *action vector* $\mathbf{A}_t^\pi \triangleq (A_t^\pi(i))_{i \in [N]} \in \mathbb{A}^N$ denotes the actions applied to each arm. For each policy π , consider the following two quantities: the limsup average reward $R^+(\pi, \mathbf{S}_0) \triangleq \limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i \in [N]} \mathbb{E}[r(S_t^\pi(i), A_t^\pi(i))]$; and the liminf average reward $R^-(\pi, \mathbf{S}_0) \triangleq \liminf_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i \in [N]} \mathbb{E}[r(S_t^\pi(i), A_t^\pi(i))]$. When the limsup and the liminf average rewards are equal, the *long-run average reward* is defined as:

$$R(\pi, \mathbf{S}_0) \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i \in [N]} \mathbb{E}[r(S_t^\pi(i), A_t^\pi(i))].$$

Our goal is solving the optimization problem:

$$\begin{aligned} & \underset{\text{policy } \pi}{\text{maximize}} && R^-(\pi, \mathbf{S}_0) && \text{subject to} && \sum_{i \in [N]} A_t^\pi(i) = \alpha N, \quad \forall t \geq 0. \quad (\text{RB}) \end{aligned}$$

Let $R^*(N, \mathbf{S}_0)$ be the optimal value of (RB), termed as the *optimal reward*. Note that $R^*(N, \mathbf{S}_0) = \sup_{\pi'} R^-(\pi', \mathbf{S}_0) = \sup_{\pi'} R^+(\pi', \mathbf{S}_0)$ because (RB) is an MDP with finite state and action spaces [230, Theorem 9.1.6]. For any policy π , we define its optimality gap as $R^*(N, \mathbf{S}_0) - R^-(\pi, \mathbf{S}_0)$. Consistent with the literature, we call a policy π *asymptotically optimal* if its optimality gap vanishes as $N \rightarrow \infty$, i.e., $R^*(N, \mathbf{S}_0) - R^-(\pi, \mathbf{S}_0) = o(1)$ [266, Definition 4.11].

In the rest of the chapter, we restrict ourselves to policies whose long-run average reward $R(\pi, \mathbf{S}_0)$ is well-defined. Such policies include all stationary Markovian policies [230, Proposition 8.1.1], as well as any stationary policy that makes decisions based on an augmented system state with a finite state space, because they all induce finite-state Markov chains. Note that focusing on such policies is sufficient because there always exists a stationary Markovian policy that achieves the optimal reward [230, Theorem 9.1.8]. Consequently, we can refer to $R(\pi, \mathbf{S}_0)$ as the objective function of (RB) and write the optimality gap as $R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0)$.

Scaled state-count vector. We introduce an alternative representation of the N -armed system state, which is used extensively throughout the chapter. For each subset $D \subseteq [N]$, we define the *scaled state-count vector*

on D as $X_t^\pi(D) = (X_t^\pi(D, s))_{s \in \mathbb{S}}$, where $X_t^\pi(D, s) \triangleq \frac{1}{N} \sum_{i \in D} \mathbb{1}\{S_t^\pi(i) = s\}$ is the number of arms in D in state s scaled by $1/N$. When $D = [N]$ is the set of all arms, we simply call $X_t^\pi([N])$ the *scaled state-count vector*. Sometimes we view $X_t^\pi(D)$ as a vector-valued function of $D \subseteq [N]$. We refer to this function X_t as the *system state* at time t . Note that the system state X_t^π contains the same information as the state vector $\mathbf{S}_t^\pi$. We also define the *scaled state-action-count vector* as $Y_t^\pi = (Y_t^\pi(s, a))_{s \in \mathbb{S}, a \in \mathbb{A}}$, where $Y_t^\pi(s, a) \triangleq \frac{1}{N} \sum_{i \in [N]} \mathbb{1}\{S_t^\pi(i) = s, A_t^\pi(i) = a\}$ is the fraction of arms with state s and action a at time t .

Additional notation. Let $[0, 1]_N = \{0, 1/N, 2/N, \dots, 1\}$. For each subset $D \subseteq [N]$, $m(D) = |D|/N$ denotes the fraction of arms contained in D , thus $m(D) \in [0, 1]_N$. Let $\Delta(\mathbb{S})$ be probability simplex on $\mathbb{S}$. For an integer $k > 0$, let I_k be the k -by- k identity matrix. We treat each distribution $v \in \Delta(\mathbb{S})$ as a row vector. Recall that π denotes a policy for the N -armed problem. When the context is clear, we drop superscript π from $\mathbf{S}_t^\pi$, $\mathbf{A}_t^\pi$, and X_t^π .

5.3 Two-Set Policy

In this section, we describe our main policy, the two-set policy. We first discuss a linear programming (LP) relaxation of (RB) in Section 5.3.1 and state assumptions. The optimal solution of the LP is then used to define two subroutines, $\bar{\pi}^*$ -Fixed-Ratio Control in Section 5.3.3, and LP-Priority in Section 5.3.4. We then use these two subroutines to construct the two-set policy in Section 5.3.5. The ideas in the two subroutines are not new; our novelty lies in how we build them into the two-set policy.

5.3.1 LP relaxation

We consider the linear programming (LP) relaxation of the N -armed problem (RB), which is crucial for the design and analysis of RB policies.

$$\begin{aligned} & \underset{\{y(s, a)\}_{s \in \mathbb{S}, a \in \mathbb{A}}}{\text{maximize}} && \sum_{s \in \mathbb{S}, a \in \mathbb{A}} r(s, a) y(s, a) && (\text{LP}) \\ & \text{subject to} && \sum_{s \in \mathbb{S}} y(s, 1) = \alpha, && (5.3) \end{aligned}$$

$$\sum_{s' \in \mathbb{S}, a \in \mathbb{A}} y(s', a) P(s', a, s) = \sum_{a \in \mathbb{A}} y(s, a), \forall s; \quad (5.4)$$

$$\sum_{s \in \mathbb{S}, a \in \mathbb{A}} y(s, a) = 1; \quad y(s, a) \geq 0, \forall s, a. \quad (5.5)$$

To see why (LP) is a relaxation of (RB), for any stationary Markovian policy π , consider

$$y^\pi(s, a) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{1}{N} \sum_{i \in [N]} \mathbf{1}\{S_t^\pi(i) = s, A_t^\pi(i) = a\} \right], \forall s \in \mathbb{S}, a \in \mathbb{A}.$$

It is not hard to see that $R(\pi, \mathbf{S}_0) = \sum_{s \in \mathbb{S}, a \in \mathbb{A}} r(s, a) y^\pi(s, a)$, and $(y^\pi(s, a))_{s \in \mathbb{S}, a \in \mathbb{A}}$ satisfies the constraints (5.3)-(5.5). Therefore, letting R^{rel} be the optimal value of (LP), one can show that $R^{\text{rel}} \geq R^*(N, \mathbf{S}_0)$ [153]. Note that the LP relaxation can be efficiently solved. We can bound the optimality gap of any policy π using the inequality $R^*(N, \mathbf{S}_0) - R^-(\pi, \mathbf{S}_0) \leq R^{\text{rel}} - R^-(\pi, \mathbf{S}_0)$. This bound has been widely used in prior work [107, 108, 149, 266, 274].

Later in this section, we will make assumptions and construct our policy based on an optimal solution to (LP), $y^* = \{y^*(s, a)\}_{s \in \mathbb{S}, a \in \mathbb{A}}$. When there are multiple optimal solutions, we stick to an arbitrary one that satisfies all the assumptions.

Optimal single-armed policy. For a fixed optimal solution to (LP), $y^* = \{y^*(s, a)\}_{s \in \mathbb{S}, a \in \mathbb{A}}$, we consider the conditional distribution $\bar{\pi}^* = (\bar{\pi}^*(a|s))_{s \in \mathbb{S}, a \in \mathbb{A}}$ given by:

$$\bar{\pi}^*(a|s) = \begin{cases} y^*(s, a) / (y^*(s, 0) + y^*(s, 1)), & \text{if } y^*(s, 0) + y^*(s, 1) > 0, \\ 1/2, & \text{if } y^*(s, 0) + y^*(s, 1) = 0. \end{cases} \quad (5.6)$$

for $s \in \mathbb{S}$ and $a \in \mathbb{A}$. We call $\bar{\pi}^*$ the *optimal single-armed policy* associated with y^* , since it can be viewed as the optimal policy for the single-armed MDP $(\mathbb{S}, \mathbb{A}, P, r)$ with a relaxed budget constraint [153].

Let $P_{\bar{\pi}^*}$ be the transition matrix induced by the policy $\bar{\pi}^*$ on the single-armed MDP, i.e., $P_{\bar{\pi}^*}(s, s') = \sum_{a \in \mathbb{A}} \bar{\pi}^*(a|s) P(s, a, s')$. Since $P_{\bar{\pi}^*}$ is a stochastic matrix, its largest eigenvalue is equal to 1. For the other eigenvalues of $P_{\bar{\pi}^*}$, we consider the following assumption.

Assumption 5.1 (Aperiodic unichain). The transition probability matrix $P_{\bar{\pi}^*}$ has a simple eigenvalue 1; the other eigenvalues all have a modulus strictly smaller than 1. In other words, $P_{\bar{\pi}^*}$ defines an aperiodic unichain on $\mathbb{S}$.¹

Most literature assume certain versions of the aperiodic unichain condition [107, 108, 149, 153, 266, 274], which are either stronger than or equivalent to Assumption 5.1 (See the appendices of Chapter 4 [153] for a detailed discussion). With Assumption 5.1, the transition matrix $P_{\bar{\pi}^*}$ induces a unique stationary distribution, $\mu^* \triangleq (\mu^*(s))_{s \in \mathbb{S}}$, where $\mu^*(s) = y^*(s, 1) + y^*(s, 0)$, as one can verify using the definition of $P_{\bar{\pi}^*}$. We call μ^* the *optimal stationary distribution*. We define $S^\emptyset \triangleq \{s \in \mathbb{S} : y^*(s, 1) = y^*(s, 0) = 0\}$, the set of transient states under policy $\bar{\pi}^*$.

Our next assumption imposes a non-degenerate condition on the optimal solution to (LP).

Assumption 5.2 (Non-degenerate). For the fixed optimal solution to (LP), y^* , there exists a unique *fluid neutral state* (or simply *neutral state*), i.e., a state $\tilde{s} \in \mathbb{S}$ s.t. $y^*(\tilde{s}, 1) > 0$ and $y^*(\tilde{s}, 0) > 0$.

Assumption 5.2 is considered in all prior analysis of RB policies that achieves exponential asymptotic optimality [107, 108]. Note that the restrictive aspect of Assumption 5.2 is the existence rather than the uniqueness. In fact, one can always find an optimal solution of (LP) such that there is at most one neutral state. The existence of such an LP solution has been argued for the $S^\emptyset = \emptyset$ case in Gast et al. [108, Section 6.1], and the general case can be found in Altman [10, Theorem 3.8].

We note that this notion of non-degeneracy is different from y^* being a non-degenerate basic feasible solution of the LP relaxation (LP): the two notions coincide when $S^\emptyset = \emptyset$ but are mutually exclusive when $S^\emptyset \neq \emptyset$. See Appendix 5.A.1 for the comparison.

Assumption 5.3 (Local stability). Given the non-degenerate condition, define Φ as a $|\mathbb{S}|$ -by- $|\mathbb{S}|$ matrix:

$$\Phi \triangleq P_{\bar{\pi}^*} - (c_{\bar{\pi}^*} - \alpha \mathbf{1})^\top (P_1(\tilde{s}) - P_0(\tilde{s})) - \mathbf{1}^\top \mu^*, \quad (5.7)$$

where $c_{\bar{\pi}^*} \triangleq (\bar{\pi}^*(1|s))_{s \in \mathbb{S}}$ and $P_a(\tilde{s}) \triangleq (P(\tilde{s}, a, s))_{s \in \mathbb{S}}$ are both row vectors; $\mathbf{1}$ is the all-one row vector. We assume that the moduli of all eigenvalues of Φ are strictly less than 1.

¹A finite-state Markov chain is a unichain if it has a single-recurrent class, with possibly some transient states.

The matrix Φ captures the transition dynamics induced by the subroutine LP-Priority (Lemma 5.8). Assumption 5.3 is equivalent to the statement that right-multiplication by $\Phi_0 \triangleq P_{\bar{\pi}^*} - (c_{\bar{\pi}^*} - \alpha \mathbb{1})^\top (P_1(\tilde{s}) - P_0(\tilde{s}))$ is a contraction on the probability simplex. Intuitively, Φ and Φ_0 agree on the probability simplex up to translation; subtracting $\mathbb{1}^\top \mu^*$ from Φ_0 replaces the trivial eigenvalue 1 of Φ_0 with 0 without affecting the other eigenvalues, which lets us state local stability as the single spectral condition that all eigenvalues of Φ have modulus less than 1. See Proposition 5.21 for a formal statement and proof. We will give a more detailed interpretation of Φ_0 in the remark after Lemma 5.8.

Remark 5.4 (Local stability and UGAP). Local stability replaces UGAP assumed in [107, 108]. UGAP is concerned with the deterministic iteration given by $x_{t+1} = \phi(x_t)$, where $\phi(x) \triangleq \mathbb{E}[X_1([N]) \mid X_0([N]) = x]$ denotes the mean transition map on $\Delta(\mathbb{S})$ under an LP-Priority policy, which we introduce in Section 5.3.4; UGAP requires x_t to converge to μ^* uniformly for any initial point x_0 . We can show that UGAP implies local stability, using the fact that the matrix Φ in (5.7) is the local linearization of ϕ at μ^* (Lemma 5.8); the formal statement and proof are provided in Appendix 5.A.1. Moreover, UGAP is strictly stronger than local stability since it rules out other global failure modes, such as having multiple attractors or limit cycles; it is known to fail on instances where local stability still holds [see, e.g. 153, Section 8].

Remark 5.5 (UGAP and ergodicity). UGAP should not be confused with ergodicity of the N -armed Markov chain: ergodicity regulates the long-run behavior of the stochastic chain $X_t([N])$, whereas UGAP is about the mean-field dynamics x_t . Ergodicity is needed for the steady-state to be well-defined, whereas UGAP is used to show the tight concentration of $\{X_\infty([N])\}_t$ around μ^* [107, Appendix C].

Remark 5.6. Assumptions 5.1, 5.2 and 5.3 are easy to verify as they pertain to either the *single-armed* MDP primitives (P, r, α) or a linear program define by these problem primitives. In contrast, verifying UGAP is significantly harder due to the dependency on the *non-linear* N -armed dynamics.

5.3.2 Overview of our main policy

Before starting to introduce elements for constructing our main policy, the two-set policy (Algorithm 10), we first give an overview. The two-set policy utilizes two subroutines, $\bar{\pi}^*$ -Fixed-Ratio Control (Subroutine 1) and LP-Priority (Subroutine 2); the latter is a local version of LP-Priority policies.

At each time step, the two-set policy partitions N arms into three subsets: the first subset D_t^{LP} has nearly optimal state distribution and takes actions according to the LP-Priority; the second subset $D_t^{\bar{\pi}^*}$ is driven towards the optimal state distribution by the $\bar{\pi}^*$ -Fixed-Ratio Control and is gradually merged into D_t^{LP} ; the third subset $(D_t^{\text{LP}} \cup D_t^{\bar{\pi}^*})^c$ chooses actions to meet the budget constraint $\sum_{i \in [N]} A_t(i) = \alpha N$, and will be gradually merged into $D_t^{\bar{\pi}^*}$. In the long run, D_t^{LP} will contain all arms in the system and remain so for the majority of the time; the LP-Priority guarantees that the average reward is close to R^{rel} .

Subroutine 1 $\bar{\pi}^*$ -Fixed-Ratio Control

Input: A set of arms with number of arms in each state $(z(s))_{s \in \mathbb{S}}$,
optimal single-armed policy $\bar{\pi}^*$

- 1: **for** $s \in \mathbb{S}$ **do**
 - 2: Activate $\bar{\pi}^*(1|s)z(s)$ expected number of arms in state s through randomized rounding
-

Subroutine 2 LP-Priority

Input: A set of n arms with number of arms in each state $(z(s))_{s \in \mathbb{S}}$,
budget parameter α , optimal single-armed policy $\bar{\pi}^*$, neutral state $\tilde{s}$

Assert Assumption 5.2 and conditions in (5.8)

- 1: Let the target budget $B = \lfloor \alpha n \rfloor$ or $\lceil \alpha n \rceil$ with suitable probability such that $\mathbb{E}[B] = \alpha n$
 - 2: Activate all arms in state $s \in S^+$
 - 3: Activate no arms in state $s \in S^-$
 - 4: Activate $z(s)/2$ expected number of arms in state s through randomized rounding for $s \in S^\emptyset$
 - 5: Activate the arms with state $\tilde{s}$ using the remaining budget out of B
-

5.3.3 Subroutine 1: $\bar{\pi}^*$ -Fixed-Ratio Control

We now describe the $\bar{\pi}^*$ -Fixed-Ratio Control, as presented in Subroutine 1. The subroutine takes the optimal single-armed policy $\bar{\pi}^*$ defined in (5.6) and a subset of arms D from the RB system as the input. The subroutine decides the actions for all arms in D to match with $\bar{\pi}^*$. Specifically, let the number of arms from D in state $s \in \mathbb{S}$ be $z(s)$. The subroutine aims to activate

$\bar{\pi}^*(1|s)$ fraction of the $z(s)$ arms. To handle the integer effect, it performs a *randomized rounding*: it activates $\lfloor \bar{\pi}^*(1|s)z(s) \rfloor$ or $\lceil \bar{\pi}^*(1|s)z(s) \rceil$ arms with properly chosen probability such that the expected proportion is $\bar{\pi}^*(1|s)$.

As shown in Lemma 5.24, if all N arms in the RB system follow the $\bar{\pi}^*$ -Fixed-Ratio Control, $\mathbb{E}[X_{t+1}([N])] = \mathbb{E}[X_t([N])] P_{\bar{\pi}^*}$, so $\mathbb{E}[X_t([N])]$ converges to μ^* as $t \rightarrow \infty$. As a result, the expected fraction of arms in state s taking action a converges to $y^*(s, a)$; the *expected budget usage* of the arms converges to αN ; the long-run average reward achieves the optimum of (LP).

Remark 5.7 (Alternative to fixed-ratio control). An alternative to $\bar{\pi}^*$ -Fixed-Ratio Control is independently sampling each arm's action using $\bar{\pi}^*$, an approach taken by [149, 153]. Both approaches activate the same expected number of arms in each state, but $\bar{\pi}^*$ -Fixed-Ratio Control is cleaner for analysis as its activation count deviates from the mean by at most 1.

5.3.4 Subroutine 2: LP-Priority

Next, we describe Subroutine 2, LP-Priority, another building block for constructing our policy. In the literature, LP-Priority refers to a *class of* priority policies [see, e.g. 108] satisfying certain a consistency condition with the optimal LP solution y^* . This consistency condition makes each instance of LP-Priority act identically when the empirical state distribution of the arms is sufficiently close to μ^* . Below we define this identical behavior in the neighborhood of μ^* , and simply use “LP-Priority” to refer to a generic instance of LP-Priority restricted to this neighborhood.

Given a set of n arms, the LP-Priority starts with a target budget B generated using the randomized rounding satisfying $\mathbb{E}[B] = \alpha n$. Then it partitions the state space $\mathbb{S}$ into four subsets based on y^* : $S^+ = \{s \in \mathbb{S}: y^*(s, 1) > 0, y^*(s, 0) = 0\}$, $S^- = \{s \in \mathbb{S}: y^*(s, 1) = 0, y^*(s, 0) > 0\}$, $S^\emptyset = \{s \in \mathbb{S}: y^*(s, 1) = 0, y^*(s, 0) = 0\}$, and $S^0 = \{s \in \mathbb{S}: y^*(s, 1) > 0, y^*(s, 0) > 0\}$. Note that by Assumption 5.3, $S^0 = \{\tilde{s}\}$. LP-Priority tries to activate all arms in S^+ , keep arms in S^- passive, activate half of the arms in state s for each $s \in S^\emptyset$, and activate a suitable number of arms in $\tilde{s}$ to meet the target budget B . Equivalently, the subroutine aims to activate a $\bar{\pi}^*(1|s)$ fraction of the $z(s)$ arms from D in state s for each non-neutral state $s \in \mathbb{S} \setminus \{\tilde{s}\}$, and activate arms in $\tilde{s}$ according to B . To ensure that the target

budget can be met, we consider the following conditions for the input subset D :

$$\begin{aligned} \sum_{s \neq \tilde{s}} \bar{\pi}^*(1|s)z(s) &\leq \alpha n - |S^\emptyset| - 1, \quad \text{and} \\ \sum_{s \neq \tilde{s}} \bar{\pi}^*(0|s)z(s) &\leq (1 - \alpha)n - |S^\emptyset| - 1. \end{aligned} \quad (5.8)$$

By the definition of $\bar{\pi}^*$ in (5.6), it is not hard to see that given (5.8), there hold $\sum_{s \neq \tilde{s}} \lceil \bar{\pi}^*(1|s)z(s) \rceil \leq \lfloor \alpha n \rfloor$ and $z(\tilde{s}) + \sum_{s \neq \tilde{s}} \lfloor \bar{\pi}^*(1|s)z(s) \rfloor \geq \lceil \alpha n \rceil$. Therefore, the budget requested by non-neural states $\mathbb{S} \setminus \{\tilde{s}\}$ does not exceed B , and the neural state ensures that exactly B unit of budget is used.

If all arms in the RB system follow Subroutine 2, the expected scaled state-count vector $\mathbb{E}[X_t([N])]$ has a linear dynamics, as stated by following the lemma (proof provided in Appendix 5.B.1).

Lemma 5.8. *Under Assumption 5.2, suppose (5.8) holds for the N -armed system at time step t . We have*

$$\mathbb{E}[X_{t+1}([N]) - \mu^* \mid X_t, \text{ all arms follow Subroutine 2}] = (X_t([N]) - \mu^*)\Phi, \quad (5.9)$$

where Φ is the $|\mathbb{S}|$ -by- $|\mathbb{S}|$ matrix defined in (5.7).

Thus under Assumption 5.3, Subroutine 2 induces locally stable dynamics. Specifically, if the RB system satisfies (5.8) for a sufficiently long time and we let all N arms follow Subroutine 2, then the scaled state-count vector $X_t([N])$ would converge to μ^* . Consequently, the expected fraction of arms in state s taking action a would be close to $y^*(s, a)$ and the time-average of expected reward would be close to the upper bound R^{rel} —hence the name “LP-Priority” for Subroutine 2.

Remark 5.9 (Additional comment on $P_{\bar{\pi}^*}$ and Φ_0). LP-Priority can be viewed as an *adaptive* modification of $\bar{\pi}^*$ -Fixed-Ratio Control: it keeps activating a $\bar{\pi}^*(1|s)$ fraction of arms at every state $s \neq \tilde{s}$, but adjusts this fraction at $\tilde{s}$ in an $X_t([N])$ -dependent way so that the per-step budget constraint is met exactly. This explains the relationship between $P_{\bar{\pi}^*}$ and $\Phi_0 = P_{\bar{\pi}^*} - (c_{\bar{\pi}^*} - \alpha \mathbf{1})^\top (P_1(\tilde{s}) - P_0(\tilde{s}))$, where the additional rank-one term in Φ_0 performs this adjustment.

When is the condition in (5.8) true?

We argue that the condition in (5.8) is true when the empirical distribution on D , $x(D)/m(D)$, is sufficiently close to μ^* , and $m(D)$ is not too close

to zero. Specifically, we will measure the distance by U -weighted L_2 norm $\|v\|_U = \sqrt{vUv^\top}$ for $v \in \mathbb{R}^{|\mathbb{S}|}$, with U given by $U \triangleq \sum_{k=0}^{\infty} \Phi^k(\Phi^\top)^k$. The role of the matrix U will be clear when we introduce our main policy in Section 5.3.5; the well-definedness and properties of U are given in Lemmas 5.26 and 5.27 of Appendix 5.B.2. In Lemma 5.10 stated below, we show that there exists $\eta = \Theta(1)$ and $\epsilon_N^{\text{fe}} = \Theta(1/N)$ such that the condition in (5.8) is true if

$$\|x(D) - m(D)\mu^*\|_U \leq \eta m(D) - \epsilon_N^{\text{fe}}. \quad (5.10)$$

Lemma 5.10 (Feasibility-ensuring pair). *For any system state x and $D \subseteq [N]$, if (5.10) holds with $\eta = |\mathbb{S}|^{-1/2} \min\{y^*(\tilde{s}, 0), y^*(\tilde{s}, 1)\}$ and $\epsilon_N^{\text{fe}} = |\mathbb{S}|^{-1/2}(|S^\emptyset| + 1)/N$, then (5.8) also holds.*

We prove Lemma 5.10 in Appendix 5.B.3. For a pair of $(\eta, \epsilon_N^{\text{fe}})$ such that (5.10) implies the assertion, we refer to it as *feasibility ensuring*.

We can equivalently write (5.10) as the non-negativity of a *slack* function, $\delta(x, D)$, given by

$$\delta(x, D) = \eta m(D) - \|x(D) - m(D)\mu^*\|_U - \epsilon_N^{\text{fe}}. \quad (5.11)$$

In our main policy, we also need a notion of approximate maximality for subsets satisfying (5.10).

Definition 5.11 (ϵ_N^{rd} -maximal feasible set). Given the current system state x , a set of arms $D \subseteq [N]$ is ϵ_N^{rd} -maximal feasible if the two conditions hold: (1) $\delta(x, D) \geq 0$ or $D = \emptyset$; (2) for any D' such that $D \subseteq D' \subseteq [N]$ and $\delta(x, D') \geq \epsilon_N^{\text{rd}}$, we have $m(D') \leq m(D) + \epsilon_N^{\text{rd}}$.

To understand this definition, note that when $\epsilon_N^{\text{rd}} = 0$, a ϵ_N^{rd} -maximal feasible set is an element of the collection of sets $\mathcal{F}(x) = \{D' \subseteq [N] : \delta(x, D') \geq 0\} \cup \{\emptyset\}$ whose supersets are not in $\mathcal{F}(x)$. Thus, this definition captures the ideas of letting as many arms follow the LP-Priority as possible. By letting ϵ_N^{rd} to be a small positive number, we tolerate some errors in the maximality, which allows a ϵ_N^{rd} -maximal feasible set to be identified efficiently for any system state x . In Appendix 5.E.1, we provide concrete algorithms to efficiently determine the ϵ_N^{rd} -maximal feasible sets for some $\epsilon_N^{\text{rd}} = O(1/N)$ when discussing the implementations of our main algorithm.

5.3.5 Constructing the Two-Set Policy

We now define the *two-set policy* for the N -armed system, whose pseudocode is given in Algorithm 10. On a high level, the two-set policy works by

picking out two disjoint subsets of arms among the N arms in each time step, and letting them follow two subroutines, LP-Priority and $\bar{\pi}^*$ -Fixed-Ratio Control, respectively. When the two sets are picked properly, the policy pushes the scaled state-count vector $X_t([N])$ of the N -armed system towards μ^* . Ultimately, the system will stabilize in the region where all arms can follow Subroutine 2 LP-Priority.

Specifically, at time step t , the two-set policy first picks a subset D_t^{LP} , where the arms will follow LP-Priority. The policy chooses D_t^{LP} to be a ϵ_N^{rd} -maximal feasible set for some predetermined $\epsilon_N^{\text{rd}} = O(1/N)$, with some additional requirements as specified in Lines 2–7 of Algorithm 10. Roughly speaking, the two-set policy intends to maximize the cardinality of D_t^{LP} and let D_t^{LP} be a superset of D_{t-1}^{LP} whenever possible, under the constraint $\delta(X_t, D_t^{\text{LP}}) \geq 0$. The constraint $\delta(X_t, D_t^{\text{LP}}) \geq 0$ ensures that arms in D_t^{LP} can follow the LP-Priority. Moreover, due to nice properties of the U -weighted norm, the choice of D_t^{LP} is “closed” under the LP-Priority, in the sense that $\delta(X_{t+1}, D_t^{\text{LP}})$ remains non-negative with high probability, and D_{t+1}^{LP} can usually be a superset of D_t^{LP} (See Lemma 5.30).

After choosing D_t^{LP} , the two-set policy picks a subset $D_t^{\bar{\pi}^*}$ that contains $\beta \triangleq \min(\alpha, 1 - \alpha)$ fraction of the remaining arms, lower-rounded to the nearest integer. The set $D_t^{\bar{\pi}^*}$ should be such that $D_t^{\bar{\pi}^*} \cap D_t^{\text{LP}} = \emptyset$ and $D_t^{\bar{\pi}^*} \supseteq D_{t-1}^{\bar{\pi}^*} \setminus D_{t-1}^{\text{LP}}$.

With D_t^{LP} and $D_t^{\bar{\pi}^*}$ chosen, the two-set policy let arms in D_t^{LP} follow Subroutine 2 LP-Priority, let the arms in $D_t^{\bar{\pi}^*}$ follow Subroutine 1 $\bar{\pi}^*$ -Fixed-Ratio Control, and chooses the actions of remaining arms to meet the budget constraint. We will show that these steps are always feasible in Lemma 5.28 in Appendix 5.C.1.

5.4 Main results

In this section, we present our main results.

5.4.1 Optimality guarantee on Two-Set Policy

Our first main theorem states an upper bound on the optimality gap of the two-set policy.

Theorem 5.12. *Suppose Assumptions 5.1, 5.2 and 5.3 hold. Let π be the two-set policy in Algorithm 10 with a pair of feasibility-ensuring $(\eta, \epsilon_N^{\text{fe}})$ such*

Algorithm 10 Two-Set Policy

Input: number of arms N , budget αN , optimal LP solution y^* ,
 $\beta \triangleq \min(\alpha, 1 - \alpha)$, a feasibility-ensuring pair of (η, ϵ_N^{fe}) ,
error tolerance $\epsilon_N^{rd} = O(1/N)$, initial system state X_0 ,
initial state vector $\mathbf{S}_0$, initial subsets $D_{-1}^{LP} = D_{-1}^{\bar{\pi}^*} = \emptyset$

- 1: **for** $t = 0, 1, \dots$ **do**
- 2: **if** $\delta(X_t, [N]) \geq 0$ **then**
- 3: Let $D_t^{LP} = [N]$
- 4: **else if** $\delta(X_t, D_{t-1}^{LP}) \geq 0$ **then**
- 5: Let D_t^{LP} be any ϵ_N^{rd} -maximal feasible subset such that $D_t^{LP} \supseteq D_{t-1}^{LP}$
- 6: **else**
- 7: Let D_t^{LP} be any ϵ_N^{rd} -maximal feasible subset
- 8: Let $D_t^{\bar{\pi}^*}$ be a subset of $[N] \setminus D_t^{LP}$ with $|D_t^{\bar{\pi}^*}| = \lfloor \beta(N - |D_t^{LP}|) \rfloor$, s.t.
either $D_t^{\bar{\pi}^*} \supseteq D_{t-1}^{\bar{\pi}^*} \setminus D_t^{LP}$ or $D_t^{\bar{\pi}^*} \subseteq D_{t-1}^{\bar{\pi}^*} \setminus D_t^{LP}$
- 9: Set $A_t(i)$ for $i \in D_t^{LP}$ using Subroutine 2 LP-Priority
- 10: Set $A_t(i)$ for $i \in D_t^{\bar{\pi}^*}$ using Subroutine 1 $\bar{\pi}^*$ -Fixed-Ratio Control
- 11: Set $A_t(i)$ for $i \notin D_t^{LP} \cup D_t^{\bar{\pi}^*}$ such that $\sum_{i \in [N]} A_t(i) = \alpha N$
- 12: Apply $A_t(i)$ for $i \in [N]$ and observe the new state vector $\mathbf{S}_{t+1}$

that $\eta = \Theta(1)$ and $\epsilon_N^{fe} = O(1/N)$, and with error tolerance $\epsilon_N^{rd} = O(1/N)$.
Then for some $C > 0$ independent of N :

$$R^{\text{rel}} - R(\pi, \mathbf{S}_0) = O(\exp(-CN)). \quad (5.12)$$

Theorem 5.12 establishes the *first* exponential asymptotic optimality without UGAP. In contrast, prior work either requires UGAP to achieve exponential asymptotic optimality or only has an $O(1/\sqrt{N})$ optimality gap. We prove Theorem 5.12 in Appendix 5.C.

Remark 5.13. We discuss the relationship between the two-set policy in this chapter and a policy from prior work named the set-expansion policy of Chapter 4 [153]. The set-expansion policy achieves an $O(1/\sqrt{N})$ optimality gap under the aperiodic unichain assumption (Assumption 5.1); it works by letting arms in a specific subset follow $\bar{\pi}^*$ and gradually expanding the subset. The two-set policy in this chapter refines this approach by maintaining two judiciously chosen sets instead of one set. Furthermore, we argue that the two-set policy can be implemented in a way such that when Assumption 5.2 or 5.3 fails, it achieves $O(1/\sqrt{N})$ optimality gap, because it becomes almost the

same as the set-expansion policy. To see this, note that when Assumption 5.2 (non-degeneracy) fails, it holds that $\eta = 0$; when Assumption 5.3 (local stability) fails, the matrix U becomes infinity; in either case, $D_t^{\text{LP}} = \emptyset$. By taking suitable actions for the arms in $(D_t^{\text{LP}} \cup D_t^{\bar{\pi}^*})^c$, we can maintain a subset of arms following the $\bar{\pi}^*$ -Fixed-Ratio Control that gradually expands in the same way as the set-expansion policy. Using similar arguments as those in Chapter 4 [153], one can prove an $O(1/\sqrt{N})$ bound for the two-set policy.

5.4.2 Discussion on the necessities of the assumptions

In this subsection, we discuss the necessities of Assumptions 5.1–5.3 in Theorem 5.12. In particular, we establish lower bounds of the following flavor: when any of Assumptions 5.1–5.3 does not hold, there exist RB problem instances such that for every policy π and certain initial state $\mathbf{S}_0$, it holds that

$$R^{\text{rel}} - R(\pi, \mathbf{S}_0) = \Omega\left(\frac{1}{\sqrt{N}}\right). \quad (5.13)$$

It is important to note that $R^{\text{rel}} - R(\pi, \mathbf{S}_0)$ is an upper bound on the optimality gap $R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0)$, so a lower bound on $R^{\text{rel}} - R(\pi, \mathbf{S}_0)$ does not necessarily imply a lower bound on the optimality gap. Nevertheless, *all* existing bounds on the optimality gap reply on R^{rel} or its variants, and our lower bounds reveal insights on when it is impossible to get exponentially close to R^{rel} .

When Assumption 5.1 (aperiodic unichain) or Assumption 5.2 (non-degeneracy) is violated.

We construct RB instances where (5.13) holds for every policy π and certain initial states $\mathbf{S}_0$; the details are provided in Appendix 5.D.1.

When Assumption 5.3 (local stability) is violated.

Instead of finding one counterexample, we take a step further to establish the lower bound (5.13) for a large class of instances.

Definition 5.14 (Regular unstable RBs). For an RB instance satisfying Assumption 5.1 and Assumption 5.2, we call it regular unstable if it has the following properties:

- (a) (LP) has a unique optimal solution y^* , and $y^*(s, 1) + y^*(s, 0) > 0$ for all $s \in \mathbb{S}$.
- (b) The matrix Φ defined in (5.7) has at least one eigenvalue whose modulus is strictly larger than 1, and the modulus of any eigenvalue of Φ does not equal 1.
- (c) For each $\xi \in \mathbb{R}^{|\mathbb{S}|}$ with $\|\xi\|_2 = 1$ and $\mu^* \xi^\top = 0$, for each large-enough N and each time step t , if $\|X_t([N]) - \mu^*\|_1 \leq \eta'$,

$$\mathbb{E}[\|(X_{t+1}([N]) - \mu^* - (X_t([N]) - \mu^*)\Phi)\xi^\top\| \mid X_t, \mathbf{A}_t] \geq \frac{K_{\text{noise}}}{\sqrt{N}}$$

where $\eta', K_{\text{noise}} > 0$ are some constants; $\mathbf{A}_t$ follows the LP-Priority (Subroutine 2), which is always feasible if $\|X_t([N]) - \mu^*\|_1 \leq \min\{y^*(\tilde{s}, 0), y^*(\tilde{s}, 1)\} - 1/N$.

Our next main theorem is a lower bound for regular unstable RBs; it is proved in Appendix 5.D.2.

Theorem 5.15. *For every regular unstable RB, policy π and initial state vector $\mathbf{S}_0$, we have*

$$R^{\text{rel}} - R(\pi, \mathbf{S}_0) = \Omega\left(\frac{1}{\sqrt{N}}\right).$$

Theorem 5.15 implies that an exponential optimality gap cannot be achieved for a significant subset of RB instances violating local stability (Assumption 5.3). Therefore, in addition to the widely accepted non-degenerate condition, local stability is also fundamental for exponential optimality.

We briefly comment on Property (c) of regular unstable RB: this property requires the state-count vector to have sufficient randomness in most directions under the LP-Priority when the system is near the optimal state distribution μ^* . In Appendix 5.D.3, we provide an intuitive sufficient condition for (c), and a weaker version of the condition under which Theorem 5.15 remains valid.

The proof of Theorem 5.15 is based on the following idea: By Lemma 5.32, arms need to closely follow the LP-Priority to maintain a small optimality gap; however, if the LP-Priority is locally unstable, the system will quickly drift away from the region where the LP-Priority is applicable. In the proof, we use Lyapunov functions to quantify how much the system drifts away from μ^* in each step, based on how closely it follows the LP-Priority.

We prove the following in Appendix 5.D.4 by giving a concrete example of regular unstable RB.

Corollary 5.16. *There exists an RB satisfying Assumptions 5.1 and 5.2 and violating Assumption 5.3 such that $R^{\text{rel}} - R(\pi, \mathbf{S}_0) \geq \Omega\left(\frac{1}{\sqrt{N}}\right)$.*

5.5 Analysis of two-set policy (Proof of Theorem 5.12)

The two-set policy induces a finite-state Markov chain on the aggregate state $\Sigma_t \triangleq (X_t, D_t^{\text{LP}}, D_t^{\bar{\pi}^*})$, whose state space is a finite set $\mathbb{X}$, the set of all $(X_t, D_t^{\text{LP}}, D_t^{\bar{\pi}^*})$ -tuples attained with positive probability under the policy π . We write $\sigma = (x, D^{\text{LP}}, D^{\bar{\pi}^*})$ for a generic element of $\mathbb{X}$. For the ease of presentation, we assume the Markov chain σ_t is aperiodic, then fixing any initial state $(X_0, D_0^{\text{LP}}, D_0^{\bar{\pi}^*})$, the limiting distribution is well-defined. We denote $\Sigma_\infty = (X_\infty, D_\infty^{\text{LP}}, D_\infty^{\bar{\pi}^*})$ as the random element distributed according to this limiting distribution. The proof for the general case requires only notational change, where expectations with respect to the limiting distribution is replaced with long-run time averages. We skip the general case for simplicity.

We define the one-step drift operator

$$\Delta f(\sigma) \triangleq \mathbb{E}[f(\Sigma_{t+1}) \mid \Sigma_t = \sigma] - f(\sigma)$$

for any function $f: \mathbb{X} \rightarrow \mathbb{R}$. Taking $\sigma = \Sigma_\infty$, by the stationarity, we have $\mathbb{E}[\Delta f(\Sigma_\infty)] = 0$.

We define the expected instantaneous reward under the two-set policy as

$$r^\pi(\sigma) \triangleq \sum_{s \in \mathbb{S}, a \in \mathbb{A}} r(s, a) \mathbb{E}[Y_t(s, a) \mid \Sigma_t = \sigma].$$

The long-run average reward of the policy then satisfies $R(\pi, \mathbf{S}_0) = \mathbb{E}[r^\pi(\Sigma_\infty)]$.

5.5.1 Lyapunov framework

We seek a pair of functions $V, u: \mathbb{X} \rightarrow \mathbb{R}$ and a scalar $O_N \geq 0$ satisfying the following two conditions:

1. (*Drift condition*) $\Delta V(\sigma) \leq -u(\sigma) + O_N$ for all $\sigma \in \mathbb{X}$;
2. (*Dominance condition*) $u(\sigma) \geq R^{\text{rel}} - r^\pi(\sigma)$ for all $\sigma \in \mathbb{X}$.

This pair of conditions is equivalent to the standard Lyapunov drift condition [206, Ch. 14]. Here V is the *Lyapunov function*, whose value at σ quantifies the

system's distance to convergence, and u bounds V 's expected rate of decrease modulo the residual O_N . The drift condition requires that V decreases at a rate of at least $u(\sigma) - O_N$; the dominance condition ensures that this rate is in turn large enough to bound the reward gap $R^{\text{rel}} - r^\pi(\sigma)$. Taking expectations in both conditions under Σ_∞ and using $\mathbb{E}[\Delta V(\Sigma_\infty)] = 0$ then gives

$$R^{\text{rel}} - R(\pi, \mathbf{S}_0) = R^{\text{rel}} - \mathbb{E}[r^\pi(\Sigma_\infty)] \leq O_N, \quad (5.14)$$

so the optimality gap is controlled by the residual scalar O_N .

5.5.2 Warm-up: proving $O(1/\sqrt{N})$ optimality gap

As a warm-up, we first bound the optimality gap by $O(1/\sqrt{N})$. The argument will be subsequently strengthened to yield an $O(\exp(-CN))$ bound.

We take $u_1(\sigma) = K_0 \|\mu^* - x([N])\|_U$, for a constant $K_0 > 0$ chosen so that the dominance condition

$$R^{\text{rel}} - r^\pi(\sigma) \leq K_0 \|\mu^* - x([N])\|_U$$

holds for every $\sigma \in \mathbb{X}$. Intuitively, such a K_0 exists because $r^\pi(\sigma) = R^{\text{rel}}$ whenever $x([N]) = \mu^*$, and r^π changes at a bounded rate as $x([N])$ deviates from μ^* . A formal derivation can be read off from Lemma 5.18, which we state and use in the $O(\exp(-CN))$ analysis; for the warm-up we simply take the dominance condition as given.

It remains to find a Lyapunov function $V_1: \mathbb{X} \rightarrow \mathbb{R}$ that satisfies the drift condition with $u_1(\sigma) = K_0 \|\mu^* - x([N])\|_U$. Equivalently, up to constants, we seek V_1 such that

$$\Delta V_1(\sigma) \leq -\|\mu^* - x([N])\|_U + O\left(\frac{1}{\sqrt{N}}\right). \quad (5.15)$$

To construct such a Lyapunov function, the key challenge lies in the complex dynamics of the two-set policy, which maintains two subsets, each running a different subroutine and could expand or shrink over time. To tackle this challenge, we use a novel *multivariate Lyapunov function* that takes not only the system state X_t but also the subsets $D_t^{\text{LP}}, D_t^{\pi^*}$ as variables.

Specifically, we first define two functions, $h_W(x, D)$ and $h_U(x, D)$, for any system state x and subset $D \subseteq [N]$:

$$h_W(x, D) = \|x(D) - m(D)\mu^*\|_W \quad (5.16)$$

$$h_U(x, D) = \|x(D) - m(D)\mu^*\|_U. \quad (5.17)$$

These two functions are Lyapunov functions that witness convergence of $X_t(D)$ to $m(D)\mu^*$ if the arms in D follow the $\bar{\pi}^*$ -Fixed-Ratio Control or the LP-Priority, and we refer to them as *subset Lyapunov functions*.

Now we define V_1 , using $h_W(x, D)$ and $h_U(x, D)$ as building blocks:

$$V_1(\sigma) = h_U(x, D^{\text{LP}}) + h_W(x, D^{\bar{\pi}^*}) + L_1(1 - m(D^{\text{LP}})), \quad (5.18)$$

where $L_1 \triangleq 2\lambda_U^{1/2} + 4\lambda_W^{1/2} - 2\lambda_W^{1/2}\beta$ is a large constant. Intuitively, the first two terms of $V_1(\sigma)$ handles the state transitions of the arms in D^{LP} and $D^{\bar{\pi}^*}$, whereas the third term handles the effect of changing D^{LP} and $D^{\bar{\pi}^*}$. We have the following lemma for properties of V_1 :

Lemma 5.17 (Properties of V_1). *Assume Assumptions 5.1 to 5.3 hold and let $\bar{\eta} = \eta - \epsilon_N^{\text{fe}}$ for a feasibility-ensuring pair $(\eta, \epsilon_N^{\text{fe}})$. There exist $\rho_1 \in (0, 1)$ and $K_1, K_2, C > 0$ independent of N , such that for all $t \geq 0$ and all $\sigma = (x, D^{\text{LP}}, D^{\bar{\pi}^*}) \in \mathbb{X}$,*

$$\mathbb{E} \left[\left(V_1(\Sigma_{t+1}) - \rho_1 V_1(\Sigma_t) \right)^+ \middle| \Sigma_t = \sigma \right] \leq \frac{K_1}{\sqrt{N}} \quad (5.19)$$

$$\mathbb{E} \left[\left(V_1(\Sigma_{t+1}) - \rho_1 V_1(\Sigma_t) - \frac{(1-\rho_1)\bar{\eta}}{2} \right)^+ \middle| \Sigma_t = \sigma \right] \leq K_2 \exp(-CN), \quad (5.20)$$

and

$$V_1(\sigma) \geq \|x([N]) - \mu^*\|_U. \quad (5.21)$$

Lemma 5.17 is proved in Appendix 5.C.2. Equation (5.21) shows that V_1 dominates the target distance $\|x([N]) - \mu^*\|_U$. Equations (5.19) and (5.20) both establish a geometric decay of V_1 by bounding $V_1(\Sigma_{t+1}) - \rho_1 V_1(\Sigma_t)$ in a suitable probabilistic sense: (5.19) controls the mean and (5.20) controls the tail. Together, (5.21) and (5.19) imply

$$\Delta V_1(\sigma) \leq -(1 - \rho_1) \|x([N]) - \mu^*\|_U + \frac{K_1}{\sqrt{N}},$$

which is (5.15) up to constants and yields the $O(1/\sqrt{N})$ optimality gap. The tail bound (5.20) will be used when we establish the stronger $O(\exp(-CN))$ bound for the optimality gap.

5.5.3 Proving $O(\exp(-CN))$ optimality gap

To get a stronger optimality-gap bound, we construct a tighter upper bound $u_2(\sigma)$ on $R^{\text{rel}} - r^\pi(\sigma)$ that exploits the *local linearity* of the instantaneous reward.

Specifically, when $x([N])$ is sufficiently close to μ^* , namely, when $\|x([N]) - \mu^*\|_U \leq \bar{\eta} \triangleq \eta - \epsilon_N^{\text{fe}}$, $r^\pi(\sigma)$ is fully determined by the LP-Priority. The following lemma, proved in Appendix 5.C.3, shows that r^π is then an affine function of $x([N])$.

Lemma 5.18 (Instantaneous reward under LP-Priority). *Under the two-set policy, for any $\sigma \triangleq (x, D^{\text{LP}}, D^{\bar{\pi}^*}) \in \mathbb{X}$, if $\|x([N]) - \mu^*\|_U \leq \bar{\eta}$, we have*

$$r^\pi(\sigma) = \hat{r}(x([N])),$$

where $\hat{r}: \Delta(\mathbb{S}) \rightarrow \mathbb{R}$ is the affine function

$$\hat{r}(v) \triangleq \sum_{s \neq \tilde{s}, a \in \mathbb{A}} (r(s, a) - r(\tilde{s}, a)) \bar{\pi}^*(a|s) v(s) + \alpha r(\tilde{s}, 1) + (1 - \alpha) r(\tilde{s}, 0), \quad (5.22)$$

which can be written in vector form as $\hat{r}(v) = v \cdot \mathbf{g} + r_0$ for some $\mathbf{g} \in \mathbb{R}^{|\mathbb{S}|}$ and $r_0 \in \mathbb{R}$.

Lemma 5.18 implies $\hat{r}(\mu^*) = R^{\text{rel}}$ and

$$R^{\text{rel}} - \hat{r}(v) = (\mu^* - v) \cdot \mathbf{g} \quad \text{for every } v \in \Delta(\mathbb{S}).$$

Combining this with the trivial bound $R^{\text{rel}} - r^\pi(\sigma) \leq 2r_{\max}$ when $D^{\text{LP}} \neq [N]$, we obtain $R^{\text{rel}} - r^\pi(\sigma) \leq u_2(\sigma)$ for every $\sigma \in \mathbb{X}$, with

$$u_2(\sigma) \triangleq \begin{cases} (\mu^* - x([N])) \cdot \mathbf{g} & \text{if } \|x([N]) - \mu^*\|_U \leq \bar{\eta}, \\ 2r_{\max} & \text{otherwise.} \end{cases} \quad (5.23)$$

This two-piece u_2 is tighter than the warm-up choice: when $x([N])$ is close to μ^* , $u_2(\sigma)$ is exactly equal to $R^{\text{rel}} - r^\pi(\sigma)$.

We now construct a second Lyapunov function V_2 whose drift mirrors the two-piece structure of $u_2(\sigma)$ in (5.23). The key new ingredient is the *local-approximation* term

$$\Psi(\sigma) \triangleq (\mu^* - x([N])) Q \mathbf{g}, \quad (5.24)$$

where $Q \triangleq (I - \Phi)^{-1}$ is well-defined since $\rho(\Phi) < 1$ by Assumption 5.3. Let

$$\lambda_Q \triangleq \max \left\{ \max_{v \in \Delta(\mathbb{S})} \|(v - \mu^*)Q\|_1, \max_{v \in \Delta(\mathbb{S})} \|(v - \mu^*)\Phi Q\|_1 \right\}. \quad (5.25)$$

We define

$$V_2(\sigma) = \left(V_1(\sigma) - \frac{\bar{\eta}}{2}\right)^+ + L_2 \Psi(\sigma), \quad (5.26)$$

where $L_2 \triangleq (1 - \rho_1)\bar{\eta} / (8r_{\max}\lambda_Q + 4r_{\max})$. The two summands serve complementary roles: $L_2 \Psi(\sigma)$ tracks the reward gap $R^{\text{rel}} - r^\pi(\sigma)$ when $x([N])$ is close to μ^* , while $(V_1(\sigma) - \bar{\eta}/2)^+$ supplies a large negative drift when $x([N])$ is far from μ^* . The following two lemmas, proved in Appendices 5.C.4 and 5.C.5, make this precise.

Lemma 5.19 (Drift of truncated V_1). *For any $\sigma, \sigma' \in \mathbb{X}$,*

$$\begin{aligned} (V_1(\sigma') - \frac{\bar{\eta}}{2})^+ - (V_1(\sigma) - \frac{\bar{\eta}}{2})^+ &\leq -\frac{(1 - \rho_1)\bar{\eta}}{2} \mathbf{1}\{V_1(\sigma) > \bar{\eta}\} \\ &\quad + \left(V_1(\sigma') - \rho_1 V_1(\sigma) - \frac{(1 - \rho_1)\bar{\eta}}{2}\right)^+. \end{aligned} \quad (5.27)$$

Lemma 5.20 (Drift of Ψ). *For any $\sigma = (x, D^{\text{LP}}, D^{\bar{\pi}^*}) \in \mathbb{X}$,*

$$\Delta \Psi(\sigma) \leq -(R^{\text{rel}} - r^\pi(\sigma)) + (4r_{\max}\lambda_Q + 2r_{\max}) \mathbf{1}\{\|x([N]) - \mu^*\|_U > \bar{\eta}\}, \quad (5.28)$$

where $\Psi(\sigma) = (\mu^* - x([N])) Q \mathbf{g}$, $Q = (I - \Phi)^{-1}$, and $\bar{\eta} = \eta - \epsilon_N^e$.

5.5.4 Proof of Theorem 5.12

With Lemmas 5.17, 5.19 and 5.20 in hand, we are ready to derive the exponential optimality-gap bound.

Proof. Setting $\sigma = \Sigma_t$, $\sigma' = \Sigma_{t+1}$ in Lemma 5.19 and taking the conditional expectation given $\Sigma_t = \sigma$,

$$\begin{aligned} \Delta(V_1(\sigma) - \frac{\bar{\eta}}{2})^+ &\leq -\frac{(1 - \rho_1)\bar{\eta}}{2} \mathbf{1}\{V_1(\sigma) > \bar{\eta}\} \\ &\quad + \mathbb{E} \left[\left(V_1(\Sigma_{t+1}) - \rho_1 V_1(\Sigma_t) - \frac{(1 - \rho_1)\bar{\eta}}{2} \right)^+ \middle| \Sigma_t = \sigma \right] \\ &\leq -\frac{(1 - \rho_1)\bar{\eta}}{2} \mathbf{1}\{\|x([N]) - \mu^*\|_U > \bar{\eta}\} \\ &\quad + K_2 \exp(-CN), \end{aligned} \quad (5.29)$$

where the expectation term is bounded by $K_2 \exp(-CN)$ via (5.20) of Lemma 5.17, and the indicator is relaxed using $\mathbf{1}\{V_1(\sigma) > \bar{\eta}\} \geq$

$\mathbb{1}\{\|x([N]) - \mu^*\|_U > \bar{\eta}\}$, which holds since $V_1(\sigma) \geq \|x([N]) - \mu^*\|_U$ by (5.21). Adding L_2 times (5.28) to (5.29),

$$\begin{aligned} \Delta V_2(\sigma) &\leq \left(-\frac{(1 - \rho_1)\bar{\eta}}{2} + L_2(4r_{\max}\lambda_Q + 2r_{\max}) \right) \mathbb{1}\{\|x([N]) - \mu^*\|_U > \bar{\eta}\} \\ &\quad - L_2(R^{\text{rel}} - r^\pi(\sigma)) + K_2 \exp(-CN) \\ &= -L_2(R^{\text{rel}} - r^\pi(\sigma)) + K_2 \exp(-CN), \end{aligned} \tag{5.30}$$

where the bracket in the first line is zero by the choice $L_2 = (1 - \rho_1)\bar{\eta} / (8r_{\max}\lambda_Q + 4r_{\max})$. Taking the expectation in (5.30) with $\sigma = \Sigma_\infty$, and using the fact that $\mathbb{E}[\Delta V_2(\Sigma_\infty)] = 0$, we get

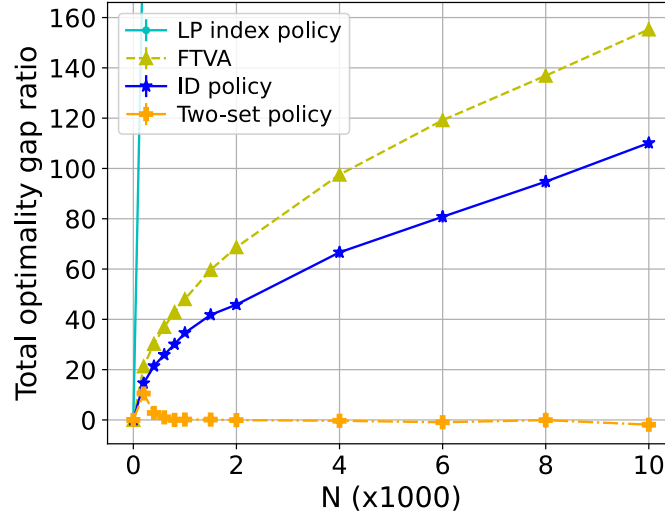
$$R^{\text{rel}} - R(\pi, \mathbf{S}_0) \leq \frac{K_2}{L_2} \exp(-CN) = O(\exp(-CN)),$$

which completes the proof. $\square$

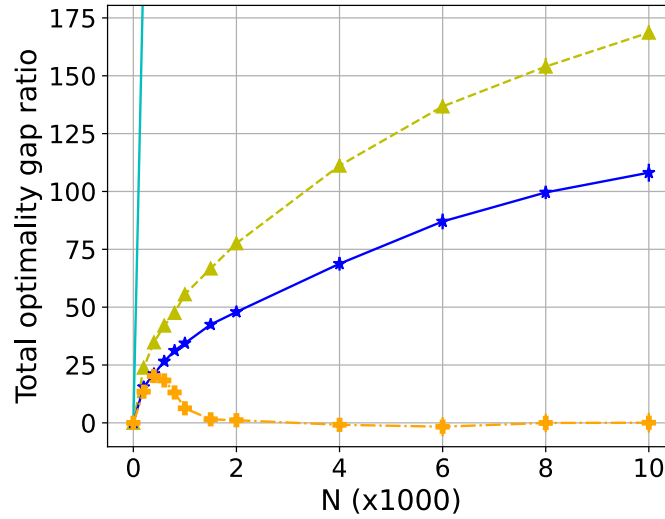
5.6 Experiments

In this section, we compare the performance of the two-set policy for RBs with those of the policies from prior work; these policies include two popular versions of LP-Priority policies, the Whittle index policy [283] and the LP index policy [108], as well as two more recent policies, Follow-the-Virtual-Advice [149], and the ID policy [153].

We make the comparisons on two types of instances of the RB problem. The first type of RB instances satisfy all of our assumptions, namely, Assumptions 5.1, 5.2, and 5.3, but violate UGAP. On this type of instances, no previous policies can achieve exponential optimality gaps. The second type of RB instances are random instances generated from the uniform distribution (See Appendix 5.E.2 for details on how they are generated). On this type of instances, simulations in prior work have shown that the Whittle index and LP index policies often perform quite well [see, e.g. 108]; here we want to see if the two-set policy also performs reasonably well. In addition to the two types of instances mentioned above, there also exist RB instances where one of our three assumptions fails. In particular, when Assumptions 5.2 or 5.3 fails, the two-set policy could preserve an $O(1/\sqrt{N})$ optimality gap if the actions of the arms in $(D_t^{\text{LP}} \cup D_t^{\bar{\pi}^*})^c$ are chosen appropriately, as argued in Remark 5.13 (See Appendix 5.E.1 for implementation details). Moreover, in



(a) First non-UGAP instance



(b) Second non-UGAP instance

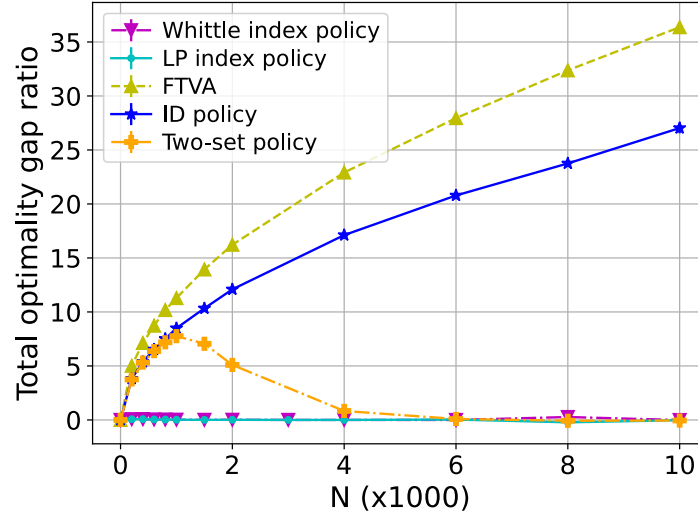
Figure 5.1: RB instances satisfying Assumptions 5.1, 5.2, and 5.3, but violating UGAP. Both subfigures plot the same four policies in the same styles, and the legend is shown only in Figure 5.1a. The Whittle index policy does not appear because it is non-indexable on both instances.

this case, the empirical performances of the two-set policy and the ID policy turn out to be indistinguishable, which we omit displaying here.

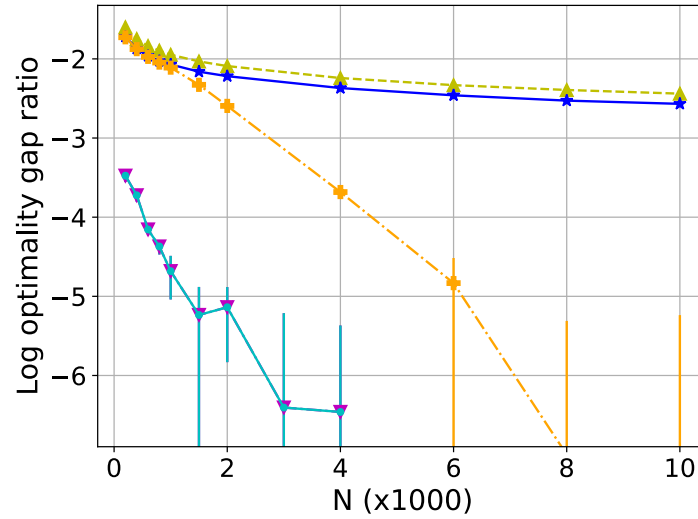
For the first type of instances, we consider two examples, which are adapted from Chapter 3 [149]. The definitions of these two examples are given in Appendix 5.E.2, where we also discuss potential methods to construct more instances of this type. In Figure 5.1, we consider the *total optimality gap ratio* as the performance metric, which is defined as $N(1 - R(\pi, \mathbf{S}_0)/R^{\text{rel}})$. Under this metric, a horizontal line indicates an $O(1/N)$ optimality gap. In both examples, we find the Whittle index policy to be non-indexable. The curves for the LP index policy grow linearly with a sharp slope, indicating its asymptotic suboptimality. The curves for FTVA and the ID policy grow with N , consistent with their $O(1/\sqrt{N})$ theoretical bounds. The curves for the two-set policy quickly converge to zero, demonstrating better-than- $O(1/N)$ optimality gaps. On both examples, the two-set policy performs significantly better than the other policies simulated here.

For the second type of instances, we consider a few uniformly randomly generated instances, each with an eight-state single-armed MDP. The simulation result of one such instance is displayed in Figure 5.2. In Figure 5.2a, we compare the total optimality gap ratios of different policies. As shown in the figure, the Whittle index policy and the LP index policy perform the best, whereas FTVA and the ID policy perform visibly worse. The total optimality gap ratios of the two-set policy match those of the ID policy for small N , but quickly decrease after $N > 2000$ and converge to zero. To further distinguish the orders of optimality gaps, in Figure 5.2b, we also plot the *log optimality gap ratios*, defined as $\log_{10}(1 - R(\pi, \mathbf{S}_0)/R^{\text{rel}})$, for the same RB instance. In Figure 5.2b, one can observe rapid decrease of the curves corresponding to the Whittle index policy, the LP index policy, and the two-set policy, which supports their exponential asymptotic optimality as predicted by the theory.

We find that the trends in Figure 5.2 are representative and can be observed in many other random instances. For example, Figure 5.3a shows the total optimality gap ratios in another random instance. Although the Whittle index and LP index policies still perform better than the two-set policy, the two-set policy demonstrates better-than- $O(1/N)$ optimality gaps and consistently performs better than those policies with $O(1/\sqrt{N})$ optimality gaps. Note that there is also a smaller fraction of random instances that shows different trends, where the two-set policy starts displaying better-than- $O(1/N)$ optimality gaps only for large N ; the simulation result on one such instance is shown in Figure 5.3b.

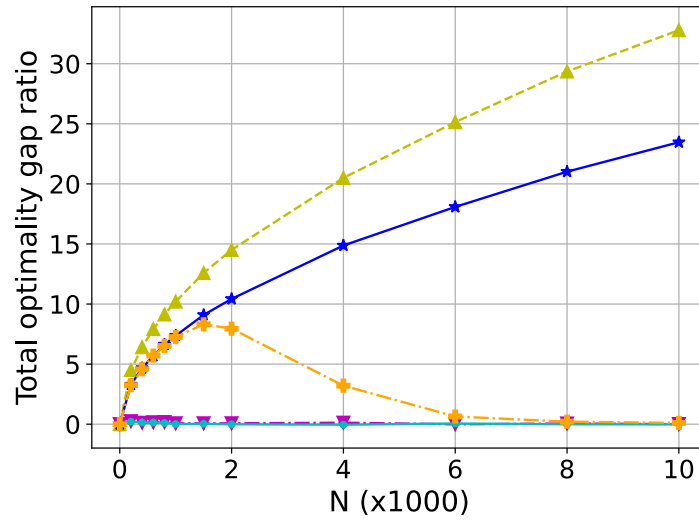


(a) Total optimality gap ratios

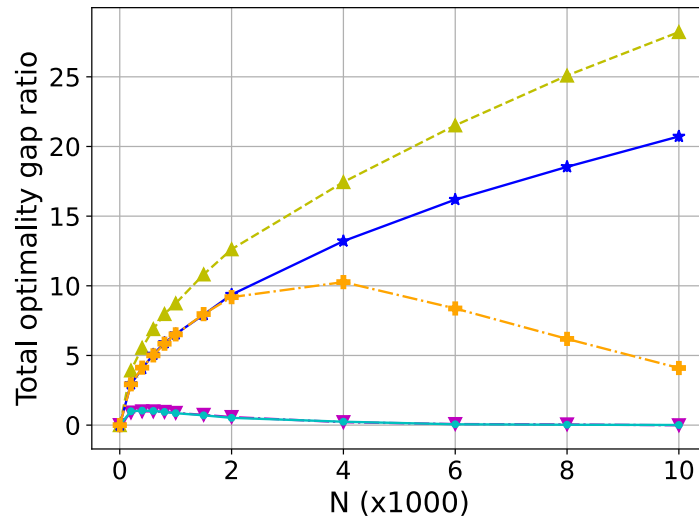


(b) Log optimality gap ratios

Figure 5.2: First random instance sampled from uniform distribution. In Figure 5.2b, we do not include data for the Whittle index and LP index policies when $N > 4000$ due to the difficulty of high-accuracy simulations. The legend of Figure 5.2a applies to both subfigures.



(a) Second random instance



(b) Third random instance

Figure 5.3: Total optimality gap ratios of two additional uniformly-generated random instances. Both subfigures plot the same five policies in the same styles as Figure 5.2a, whose legend applies here as well.

5.7 Conclusion

In this chapter, we have studied the conditions for achieving exponential asymptotic optimality in discrete-time RBs with the infinite-horizon average-reward criterion. We propose a novel two-set policy that achieves exponential asymptotic optimality under three assumptions: aperiodic unichain (Assumption 5.1), non-degeneracy (Assumption 5.2), and local stability (Assumption 5.3). Our assumptions are strictly weaker than those in the prior work, where a global attractor assumption is also required for their policies to achieve exponential asymptotic optimality.

To complement our positive result, we also discuss the necessities of our three assumptions. In particular, we prove a lower bound on $R^{\text{rel}} - R^*(N, \mathbf{S}_0)$ for a large class of RB instances that do not satisfy local stability, illustrating the fundamental role of local stability for exponential asymptotic optimality.

Finally, we numerically evaluate the performance of the two-set policy. Our numerical results show that the two-set policy achieves the best performance among the policies simulated here when our three assumptions are satisfied but UGAP is violated; the two-set policy also performs reasonably well on RB instances generated uniformly at random.

For future work, an interesting direction is to extend our results to RBs with *heterogeneous* arms, where the arms could have distinct transition dynamics and reward functions. In this case, the best known optimality gap is $O(1/\sqrt{N})$ [299] for the long-run average-reward setting. In the finite-horizon setting, Zhang [297] has achieved better optimality gaps than $O(1/\sqrt{N})$, and it remains open if their machinery can be adapted to our setting. The second interesting direction is the learning setting, where the transition kernels and reward functions of the arms are unknown. Some initial steps have been taken in [289].

5.A Auxiliary facts about the assumptions for restless bandits

This appendix collects auxiliary facts about the assumptions for restless bandits, Assumptions 5.2 and 5.3. Appendix 5.A.1 compares Assumption 5.2 (having a unique neutral state) to the notion of y^* being a non-degenerate basic feasible solution for the LP relaxation, showing that the two notions coincide

when $S^\emptyset = \emptyset$ but are mutually exclusive when $S^\emptyset \neq \emptyset$. Appendix 5.A.2 establishes two facts about Assumption 5.3: an equivalent characterization in terms of Φ_0 , and that the Uniform Global Attractor Property (UGAP) [107, 108] implies Assumption 5.3. None of these results is used in the proofs of our main results, but they provide intuition and context for the assumptions.

Below, we use the notation $\rho(M)$ to refer to the spectral radius (maximal moduli of eigenvalues) for any matrix M .

5.A.1 Comparison of Assumption 5.2 with LP non-degeneracy

We note that the non-degeneracy in Assumption 5.2 (having a unique neutral state at y^*) is distinct from y^* being a non-degenerate basic feasible solution (BFS) of the LP relaxation (LP). To compare the two, we assume that y^* is a BFS. We partition $\mathbb{S} = S^0 \cup S^+ \cup S^- \cup S^\emptyset$, where $S^+ = \{s: y^*(s, 1) > 0, y^*(s, 0) = 0\}$, $S^- = \{s: y^*(s, 1) = 0, y^*(s, 0) > 0\}$, and $S^0 = \{s: y^*(s, 1) > 0, y^*(s, 0) > 0\}$ is the set of neutral states. After dropping one redundant flow equation from (5.4), the LP relaxation has $|\mathbb{S}| + 1$ effective equality constraints. The relationship between the two notions of non-degeneracy splits into two cases:

- When $|S^\emptyset| = 0$, the two notions coincide: y^* satisfies Assumption 5.2 if and only if it is a non-degenerate BFS of the LP relaxation.
- When $|S^\emptyset| > 0$, the two notions are mutually exclusive.

These claims follow from a counting argument: When y^* is a non-degenerate BFS of the LP relaxation, it has exactly $|\mathbb{S}| + 1$ non-zero entries, which implies that $2|S^0| + |S^+| + |S^-| = |\mathbb{S}| + 1$. Because $|S^0| + |S^+| + |S^-| + |S^\emptyset| = |\mathbb{S}|$, we get $|S^0| = 1 + |S^\emptyset|$. Therefore, when $|S^\emptyset| = 0$, we have $|S^0| = 1$, so Assumption 5.2 holds. When $|S^\emptyset| > 0$, we have $|S^0| \geq 2$, contradicting Assumption 5.2.

It remains to verify the converse direction of the first bullet, i.e., that when $S^\emptyset = \emptyset$, $|S^0| = 1$ forces y^* to be a non-degenerate BFS. Suppose for contradiction y^* is a degenerate BFS; then it has strictly less than $|\mathbb{S}| + 1$ non-zero entries, giving $|S^0| < 1 + |S^\emptyset| = 1$. Hence $|S^0| = 0$, contradicting Assumption 5.2.

5.A.2 Validity of Assumption 5.3

This subsection presents two facts about Assumption 5.3. The first, Proposition 5.21, establishes an equivalent characterization of Assumption 5.3 in terms of Φ_0 . Specifically, recall that the matrices Φ and Φ_0 are defined as

$$\Phi = P_{\bar{\pi}^*} - (c_{\bar{\pi}^*} - \alpha \mathbf{1})^\top (P_1(\tilde{s}) - P_0(\tilde{s})) - \mathbf{1}^\top \mu^* \quad (5.31)$$

$$\Phi_0 = P_{\bar{\pi}^*} - (c_{\bar{\pi}^*} - \alpha \mathbf{1})^\top (P_1(\tilde{s}) - P_0(\tilde{s})). \quad (5.32)$$

Proposition 5.21 below implies that $\rho(\Phi) < 1$ (Assumption 5.3) is equivalent to the assumption that Φ_0 is contractive when right-multiplying the vectors in $\Delta(\mathbb{S})$.

Proposition 5.21 (Equivalent characterization of local stability). *Let Φ and Φ_0 be as in (5.31) and (5.32). We have the following:*

- (i) *Φ and Φ_0 coincide on the hyperplane $H \triangleq \{v \in \mathbb{R}^{|\mathbb{S}|} : v\mathbf{1}^\top = 0\}$, i.e., $v\Phi = v\Phi_0$ for all $v \in H$.*
- (ii) *The eigenvalues of Φ are exactly the eigenvalues of Φ_0 with one occurrence of the eigenvalue 1 (associated with the right eigenvector $\mathbf{1}^\top$) replaced by 0, leaving all other eigenvalues unchanged.*

Proof. For (i): for any $v \in H$, $v(\Phi - \Phi_0) = -v\mathbf{1}^\top \mu^* = -(v\mathbf{1}^\top)\mu^* = 0$, so $v\Phi = v\Phi_0$.

For (ii): Brauer's theorem on rank-one spectral perturbations [43] states that if $Ax = \lambda x$ for some right eigenvector x of a square matrix A at an eigenvalue λ , then for any column vector y , the perturbed matrix $A + xy^\top$ has the same eigenvalues as A except that one occurrence of λ is replaced by $\lambda + y^\top x$ (see 154, Example 1.2.8 for a textbook treatment, and 201 for a short and elementary proof).

We will apply Brauer's theorem with $A = \Phi_0$, $\lambda = 1$, $x = \mathbf{1}^\top$, and $y^\top = -\mu^*$. To this end, we first notice that $\mathbf{1}^\top$ is a right eigenvector of Φ_0 at eigenvalue 1, which follows from the facts that $P_{\bar{\pi}^*}\mathbf{1}^\top = \mathbf{1}^\top$, and $(P_1(\tilde{s}) - P_0(\tilde{s}))\mathbf{1}^\top = 0$. Brauer's theorem then implies that $\Phi = \Phi_0 - \mathbf{1}^\top \mu^* = \Phi_0 + xy^\top$ has the same spectrum as Φ_0 except that one occurrence of 1 is replaced by $1 + y^\top x = 1 - \mu^*\mathbf{1}^\top = 0$. $\square$

In Proposition 5.23 below, we prove the fact mentioned in Remark 5.4: the Uniform Global Attractor Property (UGAP) used in prior work [107, 108] implies Assumption 5.3. We first restate UGAP here for ease of reference.

Assumption 5.22 (Uniform Global Attractor Property). Define the mean transition map

$$\phi(x) \triangleq \mathbb{E}[X_1([N]) \mid X_0([N]) = x],$$

where the actions are determined by the LP-Priority policy. Let $x_{t+1} = \phi(x_t)$ for $t = 0, 1, \dots$ starting from any $x_0 \in \Delta(\mathbb{S})$. For each $\epsilon > 0$, there exists $T < \infty$ independent of both x_0 and N such that $\|x_t - \mu^*\| \leq \epsilon$ for all $t \geq T$ and all $x_0 \in \Delta(\mathbb{S})$.

Proposition 5.23 (UGAP implies local stability). *Assuming non-degeneracy (Assumption 5.2) and UGAP (Assumption 5.22), we have $\rho(\Phi) < 1$, so local stability (Assumption 5.3) holds.*

Proof. By non-degeneracy and Lemma 5.8, there exists a radius $\eta_0 > 0$ such that whenever $\|x - \mu^*\|_1 \leq \eta_0$, the condition (5.8) holds and the mean transition map is linear, given by: $\phi(x) - \mu^* = (x - \mu^*) \Phi$.

Suppose for contradiction that $\rho(\Phi) \geq 1$, and let $\lambda \in \mathbb{C}$ be an eigenvalue of Φ with $|\lambda| \geq 1$. We start by showing the following claim: There exist a right eigenvector y of Φ at λ and constants $b_{\min}, b_{\max} > 0$ such that for all $t \geq 0$,

$$b_{\min} |\lambda|^t \|\operatorname{Re} y\|_1 \leq \|(\operatorname{Re} y) \Phi^t\|_1 \leq b_{\max} |\lambda|^t \|\operatorname{Re} y\|_1, \quad (5.33)$$

where $\operatorname{Re} y$ denotes the entrywise real part of the vector y . We let $\operatorname{Im} y$ denote the entrywise imaginary part of y , and let $\bar{y}$ be the entrywise complex conjugate of y . We discuss two cases based on whether Φ has a real eigenvector at λ :

- *Case A: There exists a real right eigenvector y of Φ at λ .* Then $\operatorname{Im} y = 0$, $\operatorname{Re} y = y$, and $y \Phi^t = \lambda^t y$, so $\|(\operatorname{Re} y) \Phi^t\|_1 = |\lambda|^t \|\operatorname{Re} y\|_1$ and (5.33) holds with $b_{\min} = b_{\max} = 1$.
- *Case B: All right eigenvectors of Φ at λ are complex.* Let y be any right eigenvector of Φ at λ . Because $y = \operatorname{Re} y + i \operatorname{Im} y$, the two real vectors $\operatorname{Re} y$ and $\operatorname{Im} y$ must be linearly independent; otherwise, either $\operatorname{Re} y$ or $\operatorname{Im} y$ would be an eigenvector of Φ at λ , contradicting the hypothesis of Case B. Denoting θ_0 as the argument of the complex scalar λ , we have

$$(\operatorname{Re} y) \Phi^t = \frac{y + \bar{y}}{2} \Phi^t = \frac{1}{2} (\lambda^t y + \bar{\lambda}^t \bar{y}) = \operatorname{Re}(\lambda^t y).$$

Therefore, the L_1 norm of $(\operatorname{Re} y) \Phi^t$ can be written as

$$\|(\operatorname{Re} y) \Phi^t\|_1 = \|\operatorname{Re}(\lambda^t y)\|_1 = |\lambda|^t \|\operatorname{Re}(e^{i\theta_0 t} y)\|_1. \quad (5.34)$$

To prove (5.33), it remains to show that the function $g(\theta) \triangleq \|\operatorname{Re}(e^{i\theta}y)\|_1 / \|\operatorname{Re} y\|_1$ is upper and lower bounded by some constants $b_{\max}, b_{\min} > 0$, uniformly for all $\theta \in [0, 2\pi]$. This follows from the following three facts:

1. Because $\operatorname{Re} y$ and $\operatorname{Im} y$ are linearly independent, the denominator $\|\operatorname{Re} y\|_1$ is non-zero, so $g(\theta)$ is finite for all θ .
2. Because $\operatorname{Re} y$ and $\operatorname{Im} y$ are linearly independent, the numerator $\|\operatorname{Re}(e^{i\theta}y)\|_1 > 0$ for all θ , so $g(\theta) > 0$ for all θ .
3. The function $g(\theta)$ is continuous over the compact set $[0, 2\pi]$, so it can attain its maximum and minimum.

These three facts imply the existence of $b_{\max}, b_{\min} > 0$ such that for all $\theta \in [0, 2\pi]$,

$$b_{\min} \leq \|\operatorname{Re}(e^{i\theta}y)\|_1 / \|\operatorname{Re} y\|_1 \leq b_{\max},$$

which, when combined with (5.34), implies (5.33).

Now we conclude the proof of the proposition using (5.33). We normalize y so that $\|\operatorname{Re} y\|_1 = 1$. By UGAP with $\epsilon \triangleq \eta_0 b_{\min} / (2b_{\max} + 2) < \eta_0$, there exists a constant $T < \infty$ such that $\|x_t - \mu^*\|_1 \leq \epsilon$ for all $t \geq T$ and all $x_0 \in \Delta(\mathbb{S})$. We fix $x_0 = \mu^* + \delta \operatorname{Re} y$, where $\delta \triangleq \eta_0 / ((b_{\max} + 1)|\lambda|^{T+1})$.

We now establish by induction that for all $t = 0, 1, \dots, T + 1$,

$$x_t - \mu^* = \delta (\operatorname{Re} y) \Phi^t \tag{5.35}$$

$$\|x_t - \mu^*\|_1 \leq \eta_0. \tag{5.36}$$

- *Base case* ($t = 0$). By construction, $x_0 - \mu^* = \delta \operatorname{Re} y = \delta (\operatorname{Re} y) \Phi^0$, and $\|x_0 - \mu^*\|_1 = \delta \|\operatorname{Re} y\|_1 = \delta < \eta_0$.
- *Induction step*. Assume that we have proved the induction hypothesis for some $t \leq T$. We first prove (5.35) for $t + 1$: because (5.36) holds for t , by the definition of x_{t+1} , we have

$$x_{t+1} = \mu^* + (x_t - \mu^*)\Phi.$$

Substituting the induction hypothesis that $x_t - \mu^* = \delta (\operatorname{Re} y) \Phi^t$, we get $x_{t+1} - \mu^* = \delta (\operatorname{Re} y) \Phi^{t+1}$, which proves (5.35) for $t + 1$. Then, by the upper bound in (5.33) and the definition of δ , we have

$$\|x_{t+1} - \mu^*\|_1 = \delta \|(\operatorname{Re} y) \Phi^{t+1}\|_1 \leq \delta b_{\max} |\lambda|^{t+1} \leq \delta b_{\max} |\lambda|^{T+1} < \eta_0,$$

which proves (5.36) for $t + 1$.

By induction, (5.35) and (5.36) hold for all $t = 0, 1, \dots, T + 1$. Applying (5.35) at $T + 1$, we get

$$\|x_{T+1} - \mu^*\|_1 = \delta \left\| (\operatorname{Re} y) \Phi^{T+1} \right\|_1 \geq \delta b_{\min} |\lambda|^{T+1} = 2\epsilon > \epsilon.$$

This contradicts UGAP, which requires $\|x_{T+1} - \mu^*\|_1 \leq \epsilon$. Therefore, we must have $\rho(\Phi) < 1$. $\square$

5.B General lemmas on subroutine dynamics and weighted norms

This section collects general-purpose lemmas used throughout the proofs. They describe the MDP-level behavior of arms following a subroutine and the weighted norms used to measure convergence; they are not specific to the two-set policy, whose policy-specific lemmas appear in Appendix 5.C. Appendix 5.B.1 proves two lemmas on the transition dynamics under the subroutines $\bar{\pi}^*$ -Fixed-Ratio Control (Subroutine 1) and LP-Priority (Subroutine 2). Appendix 5.B.2 introduces the weight matrices W and U and establishes the pseudo-contraction properties of $P_{\bar{\pi}^*}$ and Φ under the corresponding weighted L_2 norms. Appendix 5.B.3 proves the existence of a feasibility-ensuring pair $(\eta, \epsilon_N^{\text{fe}})$.

5.B.1 Transition dynamics under subroutines

The lemma below characterizes the dynamics of a N -armed RB system if all arms follow $\bar{\pi}^*$ -Fixed-Ratio Control (Subroutine 1).

Lemma 5.24. *For any time step t ,*

$$\mathbb{E}[X_{t+1}([N]) \mid X_t, \text{all arms follow Subroutine 1}] = X_t([N])P_{\bar{\pi}^*} \quad a.s., \quad (5.37)$$

where expectation is taken entry-wise for the random vector $X_{t+1}([N])$; $P_{\bar{\pi}^*}$ is the transition matrix induced by $\bar{\pi}^*$ on the single-armed MDP, i.e., $P_{\bar{\pi}^*}(s, s') = \sum_{a \in \mathbb{A}} \bar{\pi}^*(a|s)P(s, a, s')$ for $s, s' \in \mathbb{S}$.

Proof. By definition, if all N arms follow $\bar{\pi}^*$ -Fixed-Ratio Control at time t , the expected number of arms in state s taking action a is equal to $X_t([N], s)\bar{\pi}^*(a|s)$. Among these arms, the fraction of arms transitioning to state s' in the next time step equals $P(s, a, s')$. Since the state transition of each arm is Markovian

given the state and action, the expected number of arms going from state s to state s' and taking action a equals $X_t([N], s)\bar{\pi}^*(a|s)P(s, a, s')$. Summing over s and a , we get

$$\begin{aligned} & \mathbb{E}[X_{t+1}([N], s') \mid X_t, \text{all arms follow Subroutine 1}] \\ &= \sum_{s \in \mathbb{S}, a \in \mathbb{A}} X_t([N], s)\bar{\pi}^*(a|s)P(s, a, s') \quad \forall s' \in \mathbb{S}, \end{aligned}$$

which completes the proof of the lemma. $\square$

If all arms in the RB system follow the LP-Priority (Subroutine 2), the expectation of the scaled state-count vector $X_t([N])$ has a linear dynamics, as stated in the lemma below. The lemma is basically a restatement of Lemma 2 in [107], although in [107], the LP-Priority policy was considered, and it was assumed that $y^*(s, 1) + y^*(s, 0) > 0$ for all $s \in \mathbb{S}$.

Lemma 5.8. *Under Assumption 5.2, suppose (5.8) holds for the N -armed system at time step t . We have*

$$\mathbb{E}[X_{t+1}([N]) - \mu^* \mid X_t, \text{all arms follow Subroutine 2}] = (X_t([N]) - \mu^*)\Phi, \quad (5.9)$$

where Φ is the $|\mathbb{S}|$ -by- $|\mathbb{S}|$ matrix defined in (5.7).

Proof. Throughout this proof, we use the shorthand x for $x([N])$, and omit “ $\mathbf{A}_t \sim \text{LP-Priority}$ ” in the conditioning.

We fixed an arbitrary x and let $X_t = x$. For any $s \neq \tilde{s}$ and $a \in \mathbb{A}$, the expected number of arms taking action a equals $x(s)\bar{\pi}^*(a|s)$; for $s = \tilde{s}$, the expected number of arms taking action 1 equals $\alpha - \sum_{s \neq \tilde{s}} x(s)\bar{\pi}^*(1|s)$, and the expected number of arms taking action 0 equals $1 - \alpha - \sum_{s \neq \tilde{s}} x(s)\bar{\pi}^*(0|s)$. Therefore, for any $s' \in \mathbb{S}$ such that $s' \neq \tilde{s}$,

$$\begin{aligned} & \mathbb{E}[X_{t+1}(s') \mid X_t = x] \\ &= \sum_{s \neq \tilde{s}, a \in \mathbb{A}} x(s)\bar{\pi}^*(a|s)P(s, a, s') \\ & \quad + \left(\alpha - \sum_{s \neq \tilde{s}} x(s)\bar{\pi}^*(1|s) \right) P(\tilde{s}, 1, s') + \left(1 - \alpha - \sum_{s \neq \tilde{s}} x(s)\bar{\pi}^*(0|s) \right) P(\tilde{s}, 0, s') \\ &= \sum_{s \neq \tilde{s}} x(s)P_{\bar{\pi}^*}(s, s') + \left(\alpha - \sum_{s \neq \tilde{s}} x(s)\bar{\pi}^*(1|s) \right) P_1(\tilde{s}, s') \\ & \quad + \left(1 - \alpha - \sum_{s \neq \tilde{s}} x(s)\bar{\pi}^*(0|s) \right) P_0(\tilde{s}, s') \end{aligned}$$

$$\begin{aligned}
&= \sum_{s \in \mathbb{S}} x(s) P_{\bar{\pi}^*}(s, s') - x(\tilde{s}) P_{\bar{\pi}^*}(\tilde{s}, s') \\
&\quad + \left(\alpha - \sum_{s \neq \tilde{s}} x(s) \bar{\pi}^*(1|s) \right) P_1(\tilde{s}, s') \\
&\quad + \left(1 - \alpha - \sum_{s \neq \tilde{s}} x(s) \bar{\pi}^*(0|s) \right) P_0(\tilde{s}, s').
\end{aligned} \tag{5.38}$$

Because $P_{\bar{\pi}^*}(\tilde{s}, s') = \bar{\pi}^*(1|\tilde{s}) P_1(\tilde{s}, s') + \bar{\pi}^*(0|\tilde{s}) P_0(\tilde{s}, s')$, we expand the second term in (5.38) and rearrange the terms to get

$$\begin{aligned}
&\mathbb{E}[X_{t+1}(s') | X_t = x] \\
&= \sum_{s \in \mathbb{S}} x(s) P_{\bar{\pi}^*}(s, s') + \left(\alpha - \sum_{s \in \mathbb{S}} x(s) \bar{\pi}^*(1|s) \right) P_1(\tilde{s}, s') \\
&\quad + \left(1 - \alpha - \sum_{s \in \mathbb{S}} x(s) \bar{\pi}^*(0|s) \right) P_0(\tilde{s}, s') \\
&= \sum_{s \in \mathbb{S}} x(s) P_{\bar{\pi}^*}(s, s') + \left(\alpha - \sum_{s \in \mathbb{S}} x(s) \bar{\pi}^*(1|s) \right) P_1(\tilde{s}, s') \\
&\quad - \left(\alpha - \sum_{s \in \mathbb{S}} x(s) \bar{\pi}^*(1|s) \right) P_0(\tilde{s}, s')
\end{aligned} \tag{5.39}$$

$$\begin{aligned}
&= \sum_{s \in \mathbb{S}} x(s) \left(P_{\bar{\pi}^*}(s, s') - \bar{\pi}^*(1|s) (P_1(\tilde{s}, s') - P_0(\tilde{s}, s')) \right) \\
&\quad + \alpha (P_1(\tilde{s}, s') - P_0(\tilde{s}, s')),
\end{aligned} \tag{5.40}$$

where (5.39) uses the facts that $\bar{\pi}^*(0|s) = 1 - \bar{\pi}^*(1|s)$ and $\sum_{s \in \mathbb{S}} x(s) = 1$.

Next, we calculate $\mathbb{E}[X_{t+1}(\tilde{s}) | X_t = x]$. Observe that $X_{t+1}(\tilde{s}) = 1 - \sum_{s' \neq \tilde{s}} X_{t+1}(s')$, so

$$\begin{aligned}
&\mathbb{E}[X_{t+1}(\tilde{s}) | X_t = x] \\
&= 1 - \sum_{s' \neq \tilde{s}} \mathbb{E}[X_{t+1}(s') | X_t = x] \\
&= \sum_{s \in \mathbb{S}} x(s) \left(1 - \sum_{s' \neq \tilde{s}} P_{\bar{\pi}^*}(s, s') + \sum_{s' \neq \tilde{s}} \bar{\pi}^*(1|s) (P_1(\tilde{s}, s') - P_0(\tilde{s}, s')) \right) \\
&\quad - \alpha \sum_{s' \neq \tilde{s}} (P_1(\tilde{s}, s') - P_0(\tilde{s}, s'))
\end{aligned}$$

$$\begin{aligned}
&= \sum_{s \in \mathbb{S}} x(s) \left(P_{\bar{\pi}^*}(s, \tilde{s}) - \bar{\pi}^*(1|s) (P_1(\tilde{s}, \tilde{s}) - P_0(\tilde{s}, \tilde{s})) \right) \\
&\quad + \alpha (P_1(\tilde{s}, \tilde{s}) - P_0(\tilde{s}, \tilde{s})).
\end{aligned} \tag{5.41}$$

Note that (5.41) has the same form as (5.40), with s' replaced by $\tilde{s}$.

We rewrite (5.40) and (5.41) in matrix form to get

$$\mathbb{E}[X_{t+1} | X_t = x] = x \left(P_{\bar{\pi}^*} - c_{\bar{\pi}^*}^\top (P_1(\tilde{s}) - P_0(\tilde{s})) \right) + \alpha (P_1(\tilde{s}) - P_0(\tilde{s})). \tag{5.42}$$

Recall that we have let $\Phi = P_{\bar{\pi}^*} - \mathbf{1}^\top \mu^* - (c_{\bar{\pi}^*} - \alpha \mathbf{1})^\top (P_1(\tilde{s}) - P_0(\tilde{s}))$, we can verify that the RHS of (5.42) is the same as $(x - \mu^*)\Phi + \mu^*$. Specifically,

$$\begin{aligned}
&(x - \mu^*)\Phi + \mu^* \\
&= (x - \mu^*) \left(P_{\bar{\pi}^*} - \mathbf{1}^\top \mu^* - (c_{\bar{\pi}^*} - \alpha \mathbf{1})^\top (P_1(\tilde{s}) - P_0(\tilde{s})) \right) + \mu^* \\
&= x \left(P_{\bar{\pi}^*} - c_{\bar{\pi}^*}^\top (P_1(\tilde{s}) - P_0(\tilde{s})) \right) \\
&\quad + x \left(-\mathbf{1}^\top \mu^* - \alpha \mathbf{1}^\top (P_1(\tilde{s}) - P_0(\tilde{s})) \right) \\
&\quad - \mu^* \left(P_{\bar{\pi}^*} - c_{\bar{\pi}^*}^\top (P_1(\tilde{s}) - P_0(\tilde{s})) \right) \\
&\quad - \mu^* \left(-\mathbf{1}^\top \mu^* - \alpha \mathbf{1}^\top (P_1(\tilde{s}) - P_0(\tilde{s})) \right) + \mu^* \\
&= x \left(P_{\bar{\pi}^*} - c_{\bar{\pi}^*}^\top (P_1(\tilde{s}) - P_0(\tilde{s})) \right) - \mu^* - \alpha (P_1(\tilde{s}) - P_0(\tilde{s})) \\
&\quad - \mu^* + \alpha (P_1(\tilde{s}) - P_0(\tilde{s})) + \mu^* + \alpha (P_1(\tilde{s}) - P_0(\tilde{s})) + \mu^* \\
&= x \left(P_{\bar{\pi}^*} - c_{\bar{\pi}^*}^\top (P_1(\tilde{s}) - P_0(\tilde{s})) \right) + \alpha (P_1(\tilde{s}) - P_0(\tilde{s})).
\end{aligned} \tag{5.43}$$

where to get (5.43) we use the facts that $x\mathbf{1}^\top = \mu^*\mathbf{1}^\top = 1$ and $\mu^*c_{\bar{\pi}^*}^\top = \alpha$. Substituting the RHS of (5.42) with $(x - \mu^*)\Phi + \mu^*$, we have proved that under the LP-Priority, $\mathbb{E}[X_{t+1} | X_t = x] = (x - \mu^*)\Phi + \mu^*$. This completes the proof of Lemma 5.8. $\square$

5.B.2 Weighted L_2 norms for quantifying the convergence of distributions

As discussed in Section 5.3.3 and Section 5.3.4, if a set of arms follow the $\bar{\pi}^*$ -Fixed-Ratio Control or the LP-Priority for a long time, the distribution

of their states converge to the optimal stationary distribution μ^* . In our analysis, it is convenient to define a proper metric to witness this convergence in each time step and quantify the convergence rate. For this purpose, we introduce two matrices, W and U , and study the W -weighted (U -weighted) L_2 norm on $\mathbb{S}$. Note that the same matrix W and W -weighted norm have also been used in Chapter 4 [153].

The definition of W and U are given as follows.

Definition 5.25. Assume Assumption 5.1, 5.2 and 5.3. Let W and U be $|\mathbb{S}|$ -by- $|\mathbb{S}|$ matrices given by

$$W = \sum_{k=0}^{\infty} (P_{\bar{\pi}^*} - \mathbf{1}^\top \mu^*)^k ((P_{\bar{\pi}^*} - \mathbf{1}^\top \mu^*)^\top)^k \quad (5.44)$$

$$U = \sum_{k=0}^{\infty} \Phi^k (\Phi^\top)^k, \quad (5.45)$$

Next, we show that the matrices W and U are well-defined and positive definite, and lower bound their eigenvalues by 1. Then if we let λ_W and λ_U denote maximal eigenvalues of W and U , the eigenvalues of W and U are in the range $[1, \lambda_W]$ and $[1, \lambda_U]$, respectively.

Lemma 5.26. Assume Assumption 5.1, 5.2 and 5.3, the matrices W and U given in Definition 5.25 are well-defined. Moreover, their eigenvalues are lower bounded by 1.

Proof. The well-definedness and eigenvalue lower bound for W has been proved in Chapter 4 [153]. The main fact used in the proof is that all eigenvalues of $P_{\bar{\pi}^*} - \mathbf{1}^\top \mu^*$ have moduli strictly less than 1 under Assumption 5.1. Since we have assumed the same things for Φ in Assumption 5.3, the proof for U is analogous with $P_{\bar{\pi}^*} - \mathbf{1}^\top \mu^*$ replaced by Φ . $\square$

With matrices W and U , we can define some weighted L_2 norms. Let $\|\cdot\|_W$ denote the W -weighted L_2 norm, given by $\|u\|_W = \sqrt{uWu^\top}$ for any row vector $u \in \mathbb{R}^{|\mathbb{S}|}$; let $\|\cdot\|_U$ denote the U -weighted L_2 norm, defined similarly.

Lemma 5.27 below shows that $P_{\bar{\pi}^*}$ and Φ are pseudo-contractions on the probability simplex $\Delta(\mathbb{S})$, under the W -weighted L_2 norm and the U -weighted L_2 norm, respectively.

Lemma 5.27 (Pseudo-contraction under the weighted L_2 norms). Assume Assumption 5.1, 5.2 and 5.3. For any distribution $v \in \Delta(\mathbb{S})$, we have

$$\|(v - \mu^*)P_{\bar{\pi}^*}\|_W \leq \rho_w \|v - \mu^*\|_W \quad (5.46)$$

$$\|(v - \mu^*)\Phi\|_U \leq \rho_u \|v - \mu^*\|_U. \quad (5.47)$$

where $\|\cdot\|_W$ is the W -weighted L_2 norm, i.e., $\|u\|_W = \sqrt{uWu^\top}$ for any row vector $u \in \mathbb{R}^{|\mathbb{S}|}$; $\|\cdot\|_U$ is the U -weighted L_2 norm defined similarly; $\rho_w = 1 - 1/(2\lambda_W)$; $\rho_u = 1 - 1/(2\lambda_U)$.

Proof. The inequality (5.46) has been proved in Chapter 4 [153]. The proof of (5.47) is analogous. $\square$

5.B.3 Proof of Lemma 5.10

In this section, we prove Lemma 5.10, which show that there exists $\eta = \Theta(1)$ and $\epsilon_N^{\text{fe}} = O(1/N)$ such that $(\eta, \epsilon_N^{\text{fe}})$ is feasibility ensuring, i.e., for any system state x and $D \subseteq [N]$ such that

$$\|x(D) - m(D)\mu^*\|_U \leq \eta m(D) - \epsilon_N^{\text{fe}}, \quad (5.10)$$

we have

$$\sum_{s \neq \tilde{s}} \bar{\pi}^*(1|s)x(D, s) \leq \alpha m(D) - \frac{|S^\emptyset| + 1}{N} \quad (5.48)$$

$$\sum_{s \neq \tilde{s}} \bar{\pi}^*(0|s)x(D, s) \leq (1 - \alpha)m(D) - \frac{|S^\emptyset| + 1}{N}. \quad (5.49)$$

Lemma 5.10 (Feasibility-ensuring pair). *For any system state x and $D \subseteq [N]$, if (5.10) holds with $\eta = |\mathbb{S}|^{-1/2} \min\{y^*(\tilde{s}, 0), y^*(\tilde{s}, 1)\}$ and $\epsilon_N^{\text{fe}} = |\mathbb{S}|^{-1/2}(|S^\emptyset| + 1)/N$, then (5.8) also holds.*

Proof. We first show (5.48). Observe that

$$\sum_{s \in \mathbb{S}} \bar{\pi}^*(1|s)\mu^*(s) = \sum_{s \in \mathbb{S}, a \in \mathbb{A}} y^*(s, a) = \alpha.$$

Therefore,

$$\begin{aligned} & \alpha m(D) - \sum_{s \neq \tilde{s}} \bar{\pi}^*(1|s)x(D, s) \\ &= y^*(\tilde{s}, 1)m(D) + \sum_{s \neq \tilde{s}} \bar{\pi}^*(1|s)(m(D)\mu^*(s) - x(D, s)) \\ &\geq y^*(\tilde{s}, 1)m(D) - \sum_{s \neq \tilde{s}} |m(D)\mu^*(s) - x(D, s)| \end{aligned}$$

$$\begin{aligned}
&\geq y^*(\tilde{s}, 1)m(D) - \|m(D)\mu^* - x(D)\|_1 \\
&\geq y^*(\tilde{s}, 1)m(D) - |\mathbb{S}|^{1/2} \|m(D)\mu^* - x(D)\|_U,
\end{aligned}$$

where to get the last expression, we use the argument that U 's eigenvalues are at least 1, so $\|v\|_1 \leq |\mathbb{S}|^{1/2} \|v\|_2 \leq |\mathbb{S}|^{1/2} \|v\|_U$ for any $|\mathbb{S}|$ -dim vector v . To bound $\|m(D)\mu^* - x(D, s)\|_U$, recall that D is chosen such that $\delta(x, D) \geq 0$, so by the definition of $\delta(x, D)$, we have

$$\|x(D) - m(D)\mu^*\|_U \leq \eta m(D) - \epsilon_N^{\text{fe}} \leq |\mathbb{S}|^{-1/2} y^*(\tilde{s}, 1)m(D) - \frac{|\mathbb{S}|^{-1/2}(|S^\emptyset| + 1)}{N}.$$

Therefore, combining the above calculations, we have

$$\alpha m(D) - \sum_{s \neq \tilde{s}} \bar{\pi}^*(1|s)x(D, s) \geq \frac{|S^\emptyset| + 1}{N},$$

which implies (5.48). We can prove (5.49) similarly. $\square$

5.C Proving the exponential asymptotic optimality of Two-Set Policy (Theorem 5.12)

This section is organized as follows. Appendix 5.C.1 establishes that the two-set policy is well-defined (Lemma 5.28). The remaining subsections prove the four main lemmas from the proof sketch in Section 5.5: Appendix 5.C.2 proves Lemma 5.17 on the properties of the distance-based Lyapunov function V_1 , together with three supporting lemmas on subset Lyapunov functions and the dynamics of D_t^{LP} (Lemmas 5.29 to 5.31) that are used in this proof; Appendix 5.C.3 proves Lemma 5.18 on the instantaneous reward when all arms can follow LP-Priority; Appendix 5.C.4 proves Lemma 5.19 on the drift of truncated V_1 ; and Appendix 5.C.5 proves Lemma 5.20 on the drift of the local-approximation correction Ψ . The logical relationship between the lemmas and the theorem is illustrated in the flowchart Figure 5.4. The general-purpose lemmas on subroutine dynamics and weighted norms developed in Appendix 5.B feed into the analysis chiefly through Lemma 5.29, and thence into the proof of Lemma 5.17; the only other direct use in this section is a single invocation of Lemma 5.8 in the proof of Lemma 5.20.

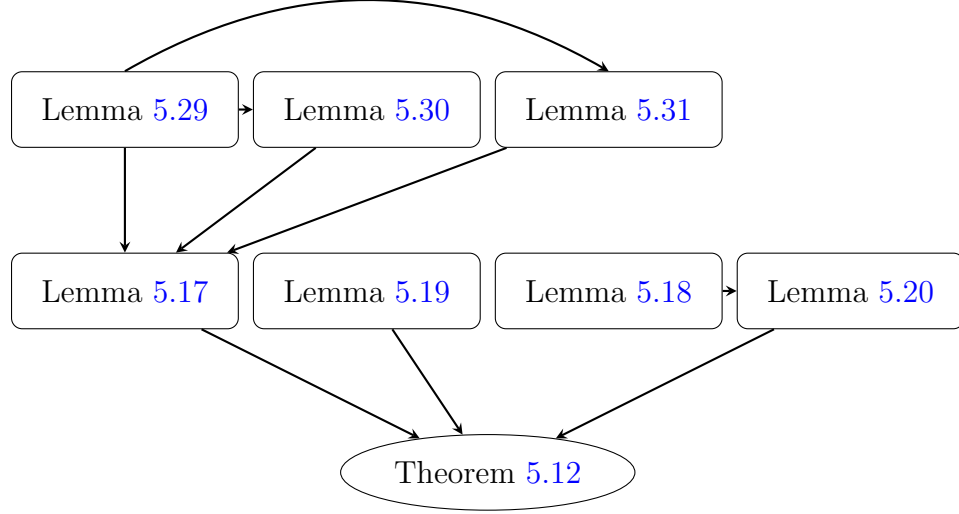


Figure 5.4: The flowchart for the proof of Theorem 5.12. The top row shows three supporting lemmas on two-set-policy dynamics, all stated and proved within Appendix 5.C.2 in service of Lemma 5.17; the middle row shows the four main lemmas from the proof sketch in Section 5.5. We omit Lemma 5.28 from the flowchart for readability since it is used in the proofs of all other lemmas and the theorem.

5.C.1 Subroutine conformity

This subsection establishes that the two-set policy (Algorithm 10) is well-defined: whenever $(\eta, \epsilon_N^{\text{fe}})$ is feasibility-ensuring, the arms in D_t^{LP} can run LP-Priority, the arms in $D_t^{\bar{\pi}^*}$ can run $\bar{\pi}^*$ -Fixed-Ratio Control, and the actions of the remaining arms can be chosen to meet the global budget constraint.

Lemma 5.28 (Subroutine conformity). *Assume Assumptions 5.1, 5.2 and 5.3 hold. Consider the two-set policy described in Algorithm 10 with a pair of feasibility-ensuring $(\eta, \epsilon_N^{\text{fe}})$. The steps in Lines 9-11 are well-defined. In other words, we can always let the arms in D_t^{LP} follow the LP-Priority, let the arms in $D_t^{\bar{\pi}^*}$ follow the $\bar{\pi}^*$ -Fixed-Ratio Control, and choose actions for the arms not in $D_t^{\text{LP}} \cup D_t^{\bar{\pi}^*}$ to satisfy the budget constraint $\sum_{i \in [N]} A_t(i) = \alpha N$.*

Proof of Lemma 5.28. We first argue that the arms in D_t^{LP} can follow the LP-Priority. By the definition of D_t^{LP} , if $D_t^{\text{LP}} \neq \emptyset$, then $\delta(X_t, D_t^{\text{LP}}) \geq 0$. Because $(\eta, \epsilon_N^{\text{fe}})$ is feasibility ensuring, (5.8) holds for the arms in D_t^{LP} , i.e.,

$$\sum_{s \neq \tilde{s}} \bar{\pi}^*(1|s) X_t(D_t^{\text{LP}}, s) \leq \alpha m(D_t^{\text{LP}}) - \frac{|S^\emptyset| + 1}{N}, \quad (5.50)$$

$$\sum_{s \neq \tilde{s}} \bar{\pi}^*(0|s) X_t(D_t^{\text{LP}}, s) \leq (1 - \alpha) m(D_t^{\text{LP}}) - \frac{|S^\emptyset| + 1}{N}. \quad (5.51)$$

which allows these arms to follow the LP-Priority.

It remains to show that $A_t(i)$ for $i \notin D_t^{\text{LP}} \cup D_t^{\bar{\pi}^*}$ can be chosen such that $\sum_{i \in [N]} A_t(i) = \alpha N$, if the arms in D_t^{LP} and $D_t^{\bar{\pi}^*}$ follow the LP-Priority and the $\bar{\pi}^*$ -Fixed-Ratio Control, respectively.

By the definition of the LP-Priority, $\sum_{i \in D_t^{\text{LP}}} A_t(i) \in \{\lfloor \alpha |D_t^{\text{LP}}| \rfloor, \lceil \alpha |D_t^{\text{LP}}| \rceil\}$. Because $A_t(i) \in \{0, 1\}$ for $i \in D_t^{\bar{\pi}^*}$, we have

$$\lfloor \alpha |D_t^{\text{LP}}| \rfloor \leq \sum_{i \in D_t^{\text{LP}} \cup D_t^{\bar{\pi}^*}} A_t(i) \leq \lceil \alpha |D_t^{\text{LP}}| \rceil + |D_t^{\bar{\pi}^*}|.$$

Because we can choose $\sum_{i \notin D_t^{\text{LP}} \cup D_t^{\bar{\pi}^*}} A_t(i)$ to be anything integer between 0 and $N - |D_t^{\text{LP}}| - |D_t^{\bar{\pi}^*}|$ to satisfy the budget constraint, it suffices to show that

$$\lfloor \alpha |D_t^{\text{LP}}| \rfloor + (N - |D_t^{\text{LP}}| - |D_t^{\bar{\pi}^*}|) \geq \alpha N \quad (5.52)$$

$$\lceil \alpha |D_t^{\text{LP}}| \rceil + |D_t^{\bar{\pi}^*}| \leq \alpha N. \quad (5.53)$$

To show (5.52), observe that $|D_t^{\bar{\pi}^*}| = \lfloor \beta(N - |D_t^{\text{LP}}|) \rfloor \leq (1 - \alpha)(N - |D_t^{\text{LP}}|)$, so

$$\alpha |D_t^{\text{LP}}| + (N - |D_t^{\text{LP}}| - |D_t^{\bar{\pi}^*}|) \geq \alpha N.$$

Rounding both sides down to the nearest integers, and utilizing the fact that $N - |D_t^{\text{LP}}| - |D_t^{\bar{\pi}^*}|$ and αN are integers, we get (5.52). Similarly, to show (5.53), observe that $|D_t^{\bar{\pi}^*}| \leq \alpha(N - |D_t^{\text{LP}}|)$, so

$$\alpha |D_t^{\text{LP}}| + |D_t^{\bar{\pi}^*}| \leq \alpha N.$$

Rounding both sides up to the nearest integers, we get (5.53). $\square$

5.C.2 Proof of Lemma 5.17

Lemma 5.17 (Properties of V_1). *Assume Assumptions 5.1 to 5.3 hold and let $\bar{\eta} = \eta - \epsilon_N^{fe}$ for a feasibility-ensuring pair (η, ϵ_N^{fe}) . There exist $\rho_1 \in (0, 1)$ and $K_1, K_2, C > 0$ independent of N , such that for all $t \geq 0$ and all $\sigma = (x, D^{\text{LP}}, D^{\bar{\pi}^*}) \in \mathbb{X}$,*

$$\mathbb{E} \left[\left(V_1(\Sigma_{t+1}) - \rho_1 V_1(\Sigma_t) \right)^+ \middle| \Sigma_t = \sigma \right] \leq \frac{K_1}{\sqrt{N}} \quad (5.19)$$

$$\mathbb{E} \left[\left(V_1(\Sigma_{t+1}) - \rho_1 V_1(\Sigma_t) - \frac{(1-\rho_1)\bar{\eta}}{2} \right)^+ \middle| \Sigma_t = \sigma \right] \leq K_2 \exp(-CN), \quad (5.20)$$

and

$$V_1(\sigma) \geq \|x([N]) - \mu^*\|_U. \quad (5.21)$$

The proof of Lemma 5.17 uses three supporting lemmas—on the subset Lyapunov functions h_W, h_U and on the dynamics of D_t^{LP} —that are specific to the two-set policy and are used only within this subsection. We state them first, then prove Lemma 5.17, and finally prove each supporting lemma.

Statement of supporting lemmas

Recall the subset Lyapunov functions

$$h_W(x, D) \triangleq \|x(D) - m(D)\mu^*\|_W \quad (5.16)$$

$$h_U(x, D) \triangleq \|x(D) - m(D)\mu^*\|_U, \quad (5.17)$$

defined for each system state x and subset $D \subseteq [N]$. The next lemma records their drift, concentration, strength, and Lipschitz properties.

Lemma 5.29 (Properties of h_W and h_U). *Assume Assumptions 5.1, 5.2 and 5.3 hold. For any time step t and $D \subseteq [N]$, let X'_{t+1} be the system state at time $t+1$ if the arms in D follow the $\bar{\pi}^*$ -Fixed-Ratio Control at time t , then we have*

$$\mathbb{E}[(h_W(X'_{t+1}, D) - \rho_w h_W(X_t, D))^+ \mid X_t] \leq \frac{K_{W,1}}{\sqrt{N}} \quad a.s. \quad (5.54)$$

$$\mathbb{P}[h_W(X'_{t+1}, D) > \rho_w h_W(X_t, D) + r \mid X_t] \leq K_{W,2} \exp(-C_W N r^2) \quad a.s. \quad \forall r \geq 0, \quad (5.55)$$

where $\rho_w = 1 - 1/(2\lambda_W)$, and $K_{W,1}, K_{W,2}, C_W$ are positive constants independent of N . In addition, for any $D, D' \subseteq [N]$,

$$h_W(x, D) \geq \frac{1}{|\mathbb{S}|^{1/2}} \|x(D) - m(D)\mu^*\|_1 \quad (5.56)$$

$$|h_W(x, D) - h_W(x, D')| \leq 2\lambda_W^{1/2}(m(D' \setminus D) + m(D \setminus D')), \quad (5.57)$$

The same inequalities hold for h_U with the $\bar{\pi}^*$ -Fixed-Ratio Control and W replaced by the LP-Priority and U . We omit the details here.

The next two lemmas characterize how the set D_t^{LP} evolves. Lemma 5.30 shows that the arms in D_t^{LP} are most likely to remain in D_{t+1}^{LP} , so that D_t^{LP} is almost non-shrinking in the set-inclusive sense. Lemma 5.31 exploits the maximality of D_t^{LP} to show that the fraction of arms not in D_t^{LP} is small whenever $X_t(D_t^{\bar{\pi}^*})$ is close to $m(D_t^{\bar{\pi}^*})\mu^*$ or $X_t([N])$ is close to μ^* . Together they imply that D_t^{LP} tends to gradually expand and cover all arms as the two subroutines drive the states of the arms in D_t^{LP} and $D_t^{\bar{\pi}^*}$ to the optimal distribution.

Lemma 5.30 (Almost non-shrinking). *Assume Assumptions 5.1, 5.2 and 5.3 hold. Consider the two-set policy (Algorithm 10) such that $(\eta, \epsilon_N^{\text{fe}})$ is feasibility-ensuring, $\eta = \Theta(1)$ and $\epsilon_N^{\text{fe}} = O(1/N)$. For any $t \geq 0$, we have*

$$\mathbb{E}[m(D_t^{\text{LP}} \setminus D_{t+1}^{\text{LP}}) \mid \Sigma_t] \leq K_{D,1} m(D_t^{\text{LP}}) \exp(-C_D N m(D_t^{\text{LP}})^2) \quad \text{a.s.} \quad (5.58)$$

$$\mathbb{P}[m(D_t^{\text{LP}} \setminus D_{t+1}^{\text{LP}}) > 0 \mid \Sigma_t] \leq K_{D,2} \exp(-C_D N m(D_t^{\text{LP}})^2) \quad \text{a.s.}, \quad (5.59)$$

where $K_{D,1}, K_{D,2}, C_D$ are positive constants independent of N .

Lemma 5.31 (Sufficient coverage). *Assume Assumptions 5.1, 5.2 and 5.3 hold. Consider the two-set policy (Algorithm 10) such that $(\eta, \epsilon_N^{\text{fe}})$ is feasibility-ensuring, $\eta = \Theta(1)$, $\epsilon_N^{\text{fe}} = O(1/N)$, and $\epsilon_N^{\text{rd}} = O(1/N)$. For any $t \geq 0$, we have*

$$1 - m(D_t^{\text{LP}}) \leq K_{C,1} h_W(X_t, D_t^{\bar{\pi}^*}) + \frac{K_{C,2}}{N} \quad \text{a.s.} \quad (5.60)$$

$$m(D_t^{\text{LP}}) = 1 \quad \text{if } h_U(X_t, [N]) \leq \eta - \epsilon_N^{\text{fe}}, \quad (5.61)$$

where $K_{C,1}, K_{C,2}$ are positive constants independent of N .

Proof of Lemma 5.17

With these three lemmas in hand, we now prove Lemma 5.17.

Proof. Recall that $V_1(\sigma)$ is defined as

$$V_1(\sigma) = h_U(x, D^{\text{LP}}) + h_W(x, D^{\bar{\pi}^*}) + L_1(1 - m(D^{\text{LP}})). \quad (5.62)$$

Proof of (5.21). By Lemma 5.29, $h_U(x, D)$ is Lipschitz continuous in D , so $h_U(x, D^{\text{LP}})$ changes by at most $2\lambda_U^{1/2}(1 - m(D^{\text{LP}}))$ when we replace D^{LP} with $[N]$. Consequently,

$$V_1(\sigma) \geq h_U(x, D^{\text{LP}}) + 2\lambda_U^{1/2}(1 - m(D^{\text{LP}})) \geq h_U(x, [N]) = \|x([N]) - \mu^*\|_U.$$

An preliminary bound for proving (5.19) and (5.20). Next, to prove (5.19) and (5.20), we claim the following upper bound on $V_1(\Sigma_{t+1}) - V_1(\Sigma_t)$: there exists a constant $\rho_1 \in (0, 1)$ such that

$$\begin{aligned} V_1(\Sigma_{t+1}) - \rho_1 V_1(\Sigma_t) \\ \leq (h_U(X_{t+1}, D_t^{\text{LP}}) - \rho_u h_U(X_t, D_t^{\text{LP}})) + (h_W(X_{t+1}, D_t^{\bar{\pi}^*}) - \rho_w h_W(X_t, D_t^{\bar{\pi}^*})) \end{aligned} \quad (5.63)$$

$$+ 4(\lambda_U^{1/2} + \lambda_W^{1/2})m(D_t^{\text{LP}} \setminus D_{t+1}^{\text{LP}}) + \frac{2\lambda_W^{1/2}}{N}. \quad (5.64)$$

To show this, we start with the decomposition:

$$V_1(\Sigma_{t+1}) - V_1(\Sigma_t) = (V_1(X_{t+1}, D_t^{\text{LP}}, D_t^{\bar{\pi}^*}) - V_1(X_t, D_t^{\text{LP}}, D_t^{\bar{\pi}^*})) \quad (5.65)$$

$$+ (V_1(X_{t+1}, D_{t+1}^{\text{LP}}, D_{t+1}^{\bar{\pi}^*}) - V_1(X_{t+1}, D_t^{\text{LP}}, D_t^{\bar{\pi}^*})). \quad (5.66)$$

We refer to the difference term in right-hand side of (5.66) the *set-update term*, and the difference term in (5.65) the *state-transition term*. We calculate these terms separately.

For the state-transition term in (5.65), it follows directly from the definition of V_1 that

$$\begin{aligned} & V_1(X_{t+1}, D_t^{\text{LP}}, D_t^{\bar{\pi}^*}) - V_1(X_t, D_t^{\text{LP}}, D_t^{\bar{\pi}^*}) \\ & \leq (h_U(X_{t+1}, D_t^{\text{LP}}) - h_U(X_t, D_t^{\text{LP}})) + (h_W(X_{t+1}, D_t^{\bar{\pi}^*}) - h_W(X_t, D_t^{\bar{\pi}^*})) \end{aligned} \quad (5.67)$$

For the set-update term in (5.66), we can derive a bound fully in terms of the sizes of sets, using the Lipschitz continuity of $h_W(x, D)$ and $h_U(x, D)$ with respect to D given in (5.57) of Lemma 5.29:

$$V_1(X_{t+1}, D_{t+1}^{\text{LP}}, D_{t+1}^{\bar{\pi}^*}) - V_1(X_{t+1}, D_t^{\text{LP}}, D_t^{\bar{\pi}^*})$$

$$\begin{aligned}
&= h_U(X_{t+1}, D_{t+1}^{\text{LP}}) - h_U(X_{t+1}, D_t^{\text{LP}}) + h_W(X_{t+1}, D_{t+1}^{\bar{\pi}^*}) - h_W(X_{t+1}, D_t^{\bar{\pi}^*}) \\
&\quad - L_1(m(D_{t+1}^{\text{LP}}) - m(D_t^{\text{LP}})) \\
&\leq 2\lambda_U^{1/2}(m(D_{t+1}^{\text{LP}} \setminus D_t^{\text{LP}}) + m(D_t^{\text{LP}} \setminus D_{t+1}^{\text{LP}})) \\
&\quad + 2\lambda_W^{1/2}(m(D_{t+1}^{\bar{\pi}^*} \setminus D_t^{\bar{\pi}^*}) + m(D_t^{\bar{\pi}^*} \setminus D_{t+1}^{\bar{\pi}^*})) \\
&\quad - L_1(m(D_{t+1}^{\text{LP}} \setminus D_t^{\text{LP}}) - m(D_t^{\text{LP}} \setminus D_{t+1}^{\text{LP}})). \tag{5.68}
\end{aligned}$$

Next, we show that

$$m(D_{t+1}^{\bar{\pi}^*} \setminus D_t^{\bar{\pi}^*}) + m(D_t^{\bar{\pi}^*} \setminus D_{t+1}^{\bar{\pi}^*}) \leq 2m(D_{t+1}^{\text{LP}} \setminus D_t^{\text{LP}}) - \beta(m(D_{t+1}^{\text{LP}}) - m(D_t^{\text{LP}})) + \frac{1}{N}. \tag{5.69}$$

We discuss based on whether $D_{t+1}^{\bar{\pi}^*} \supseteq D_t^{\bar{\pi}^*} \setminus D_{t+1}^{\text{LP}}$ or $D_{t+1}^{\bar{\pi}^*} \subseteq D_t^{\bar{\pi}^*} \setminus D_{t+1}^{\text{LP}}$.

- If $D_{t+1}^{\bar{\pi}^*} \supseteq D_t^{\bar{\pi}^*} \setminus D_{t+1}^{\text{LP}}$, because $D_t^{\bar{\pi}^*}$ is disjoint from D_t^{LP} , it is not hard to see that

$$D_t^{\bar{\pi}^*} \setminus D_{t+1}^{\bar{\pi}^*} \subseteq D_{t+1}^{\text{LP}} \setminus D_t^{\text{LP}}, \quad m(D_t^{\bar{\pi}^*} \setminus D_{t+1}^{\bar{\pi}^*}) \leq m(D_{t+1}^{\text{LP}} \setminus D_t^{\text{LP}}). \tag{5.70}$$

Moreover, by the definition of $D_t^{\bar{\pi}^*}$, $m(D_t^{\bar{\pi}^*}) \geq \beta(1 - m(D_t^{\text{LP}})) - 1/N$ and $m(D_{t+1}^{\bar{\pi}^*}) \leq \beta(1 - m(D_{t+1}^{\text{LP}}))$, so

$$\begin{aligned}
m(D_{t+1}^{\bar{\pi}^*} \setminus D_t^{\bar{\pi}^*}) &= m(D_t^{\bar{\pi}^*} \setminus D_{t+1}^{\bar{\pi}^*}) + m(D_{t+1}^{\bar{\pi}^*}) - m(D_t^{\bar{\pi}^*}) \\
&\leq m(D_{t+1}^{\text{LP}} \setminus D_t^{\text{LP}}) - \beta(m(D_{t+1}^{\text{LP}}) - m(D_t^{\text{LP}})) + \frac{1}{N}. \tag{5.71}
\end{aligned}$$

Combining (5.70) and (5.71), we get (5.69).

- If $D_{t+1}^{\bar{\pi}^*} \subseteq D_t^{\bar{\pi}^*} \setminus D_{t+1}^{\text{LP}}$, we have $m(D_{t+1}^{\bar{\pi}^*} \setminus D_t^{\bar{\pi}^*}) = 0$ and

$$\begin{aligned}
m(D_t^{\bar{\pi}^*} \setminus D_{t+1}^{\bar{\pi}^*}) &= m(D_t^{\bar{\pi}^*}) - m(D_{t+1}^{\bar{\pi}^*}) \\
&= \lfloor \beta(1 - m(D_t^{\text{LP}})) \rfloor - \lfloor \beta(1 - m(D_{t+1}^{\text{LP}})) \rfloor \\
&\leq \beta(m(D_{t+1}^{\text{LP}}) - m(D_t^{\text{LP}})) + \frac{1}{N},
\end{aligned}$$

which implies (5.69) because $m(D_{t+1}^{\text{LP}} \setminus D_t^{\text{LP}}) \geq m(D_{t+1}^{\text{LP}}) - m(D_t^{\text{LP}})$ and $\beta < 1$.

Plugging (5.69) into (5.68) and rearranging the terms, we get an upper bound for the set-update term:

$$\begin{aligned}
&V_1(X_{t+1}, D_{t+1}^{\text{LP}}, D_{t+1}^{\bar{\pi}^*}) - V_1(X_{t+1}, D_t^{\text{LP}}, D_t^{\bar{\pi}^*}) \\
&\leq 4(\lambda_U^{1/2} + \lambda_W^{1/2})m(D_t^{\text{LP}} \setminus D_{t+1}^{\text{LP}}) + \frac{2\lambda_W^{1/2}}{N}, \tag{5.72}
\end{aligned}$$

where the terms involving to $m(D_{t+1}^{\text{LP}} \setminus D_t^{\text{LP}})$ have been canceled out.

Substituting the above calculation into (5.65) and (5.66), we get

$$\begin{aligned} V_1(\Sigma_{t+1}) - V_1(\Sigma_t) & \\ & \leq h_U(X_{t+1}, D_t^{\text{LP}}) - h_U(X_t, D_t^{\text{LP}}) \\ & \quad + h_W(X_{t+1}, D_t^{\bar{\pi}^*}) - h_W(X_t, D_t^{\bar{\pi}^*}) \\ & \quad + 4(\lambda_U^{1/2} + \lambda_W^{1/2})m(D_t^{\text{LP}} \setminus D_{t+1}^{\text{LP}}) + \frac{2\lambda_W^{1/2}}{N}. \end{aligned} \quad (5.73)$$

By Lemma 5.31, we have

$$\begin{aligned} V_1(\Sigma_t) & \leq h_U(X_t, D_t^{\text{LP}}) + h_W(X_t, D_t^{\bar{\pi}^*}) + L_1 \left(K_{C,1} h_W(X_t, D_t^{\bar{\pi}^*}) + \frac{K_{C,2}}{N} \right) \\ & \leq K_{Vh} \cdot ((1 - \rho_u) h_U(X_t, D_t^{\text{LP}}) + (1 - \rho_w) h_W(X_t, D_t^{\bar{\pi}^*})) + \frac{L_1 K_{C,2}}{N}, \end{aligned}$$

for some constant $K_{Vh} > 0$. Taking $\rho_1 = 1/K_{Vh}$, and multiplying both sides by ρ_1 , we get

$$\rho_1 V_1(\Sigma_t) \leq (1 - \rho_u) h_U(X_t, D_t^{\text{LP}}) + (1 - \rho_w) h_W(X_t, D_t^{\bar{\pi}^*}) + \frac{\rho_1 L_1 K_{C,2}}{N}. \quad (5.74)$$

Adding (5.74) and the bound in (5.73), we have proved the preliminary bound claimed in (5.63)–(5.64).

Proof of (5.19). Next, we prove (5.19), restated here for the ease of reference: there exists a constant K_1 independent of N , such that for any $\sigma \in \mathbb{X}$, we have

$$\mathbb{E} \left[\left(V_1(\Sigma_{t+1}) - \rho_1 V_1(\Sigma_t) \right)^+ \middle| \Sigma_t = \sigma \right] \leq \frac{K_1}{\sqrt{N}}. \quad (5.19)$$

To prove (5.19), we apply (5.63)–(5.64) to get

$$\begin{aligned} & \mathbb{E} \left[\left(V_1(\Sigma_{t+1}) - \rho_1 V_1(\Sigma_t) \right)^+ \middle| \Sigma_t = \sigma \right] \\ & \leq (h_U(X_{t+1}, D_t^{\text{LP}}) - \rho_u h_U(X_t, D_t^{\text{LP}}))^+ \\ & \quad + (h_W(X_{t+1}, D_t^{\bar{\pi}^*}) - \rho_w h_W(X_t, D_t^{\bar{\pi}^*}))^+ \\ & \quad + 4(\lambda_U^{1/2} + \lambda_W^{1/2})m(D_t^{\text{LP}} \setminus D_{t+1}^{\text{LP}}) + \frac{2\lambda_W^{1/2}}{N} \end{aligned}$$

$$\leq \frac{K_{U,1}}{\sqrt{N}} + \frac{K_{W,1}}{\sqrt{N}} + \frac{4(\lambda_U^{1/2} + \lambda_W^{1/2})K_{D,1}}{\sqrt{N}} + \frac{2\lambda_W^{1/2}}{N},$$

where the last step is due to Lemma 5.29 and Lemma 5.30. This proves (5.19).

Proof of (5.20). Finally, we prove (5.20), restated below: there exists constants $K_2, C > 0$ independent of N , such that

$$\mathbb{E} \left[\left(V_1(\Sigma_{t+1}) - \rho_1 V_1(\Sigma_t) - \frac{(1-\rho_1)\bar{\eta}}{2} \right)^+ \middle| \Sigma_t \right] \leq K_2 \exp(-CN) \quad a.s. \quad (5.20)$$

Let $r_0 \triangleq (1 - \rho_1)\bar{\eta}/2$ and let M be a finite upper bound on V_1 over $\mathbb{X}$. Since $(V_1(\Sigma_{t+1}) - \rho_1 V_1(\Sigma_t) - r_0)^+ \leq M$ a.s., we have

$$\begin{aligned} & \mathbb{E} \left[\left(V_1(\Sigma_{t+1}) - \rho_1 V_1(\Sigma_t) - r_0 \right)^+ \middle| \Sigma_t \right] \\ & \leq M \cdot \mathbb{P} \left[V_1(\Sigma_{t+1}) > \rho_1 V_1(\Sigma_t) + r_0 \middle| \Sigma_t \right] \quad a.s., \end{aligned} \quad (5.75)$$

so it suffices to show the probability on the right decays as $\exp(-CN)$ for some $C > 0$ independent of N and σ .

By (5.63)–(5.64) and the union bound,

$$\begin{aligned} & \mathbb{P} \left[V_1(\Sigma_{t+1}) > \rho_1 V_1(\Sigma_t) + r_0 \middle| \Sigma_t \right] \\ & \leq \mathbb{P} \left[h_U(X_{t+1}, D_t^{\text{LP}}) > \rho_u h_U(X_t, D_t^{\text{LP}}) + r_0/4 \middle| \Sigma_t \right] \\ & \quad + \mathbb{P} \left[h_W(X_{t+1}, D_t^{\bar{\pi}^*}) > \rho_w h_W(X_t, D_t^{\bar{\pi}^*}) + r_0/4 \middle| \Sigma_t \right] \\ & \quad + \mathbb{P} \left[4(\lambda_U^{1/2} + \lambda_W^{1/2})m(D_t^{\text{LP}} \setminus D_{t+1}^{\text{LP}}) > r_0/4 \middle| \Sigma_t \right] \\ & \quad + \mathbb{1} \left\{ 2\lambda_W^{1/2}/N > r_0/4 \right\}. \end{aligned} \quad (5.76)$$

We bound each term on the right-hand side.

By (5.96) and (5.55) of Lemma 5.29,

$$\begin{aligned} & \mathbb{P} \left[h_U(X_{t+1}, D_t^{\text{LP}}) > \rho_u h_U(X_t, D_t^{\text{LP}}) + r_0/4 \middle| \Sigma_t \right] \\ & \leq K_{U,2} \exp \left(-\frac{C_U}{16} N r_0^2 \right) \quad a.s., \end{aligned} \quad (5.77)$$

$$\begin{aligned} & \mathbb{P} \left[h_W(X_{t+1}, D_t^{\bar{\pi}^*}) > \rho_w h_W(X_t, D_t^{\bar{\pi}^*}) + r_0/4 \middle| \Sigma_t \right] \\ & \leq K_{W,2} \exp \left(-\frac{C_W}{16} N r_0^2 \right) \quad a.s. \end{aligned} \quad (5.78)$$

For the third term on the right-hand side of (5.76), set $c \triangleq r_0/(16(\lambda_U^{1/2} + \lambda_W^{1/2}))$. If $m(D_t^{\text{LP}}) \geq c$, Lemma 5.30 gives

$$\begin{aligned} \mathbb{P}[4(\lambda_U^{1/2} + \lambda_W^{1/2})m(D_t^{\text{LP}} \setminus D_{t+1}^{\text{LP}}) > r_0/4 \mid \Sigma_t] &\leq \mathbb{P}[m(D_t^{\text{LP}} \setminus D_{t+1}^{\text{LP}}) > 0 \mid \Sigma_t] \\ &\leq K_{D,2} \exp(-C_D N m(D_t^{\text{LP}})^2) \\ &\leq K_{D,2} \exp(-C_D N c^2). \end{aligned} \quad (5.79)$$

If $m(D_t^{\text{LP}}) < c$, then $4(\lambda_U^{1/2} + \lambda_W^{1/2})m(D_t^{\text{LP}} \setminus D_{t+1}^{\text{LP}}) \leq 4(\lambda_U^{1/2} + \lambda_W^{1/2})m(D_t^{\text{LP}}) < r_0/4$, so the event is empty and the probability is zero.

Finally, the indicator $\mathbb{1}\{2\lambda_W^{1/2}/N > r_0/4\}$ vanishes for $N > 8\lambda_W^{1/2}/r_0$, a fixed threshold independent of σ .

Combining (5.76), (5.77), (5.78), and (5.79), there exist constants $\tilde{K}, \tilde{C} > 0$ independent of N and σ such that, for N above a fixed threshold,

$$\begin{aligned} \mathbb{P}[V_1(\Sigma_{t+1}) > \rho_1 V_1(\Sigma_t) + r_0 \mid \Sigma_t] \\ \leq \tilde{K} \exp(-\tilde{C}N) \quad a.s. \end{aligned} \quad (5.80)$$

Substituting into (5.75) and setting $K_2 = M\tilde{K}$ and $C = \tilde{C}$ establishes (5.20). $\square$

It remains to prove the three supporting lemmas stated at the beginning of this subsection.

Proofs of supporting lemmas

Lemma 5.29 (Properties of h_W and h_U). *Assume Assumptions 5.1, 5.2 and 5.3 hold. For any time step t and $D \subseteq [N]$, let X'_{t+1} be the system state at time $t+1$ if the arms in D follow the $\bar{\pi}^*$ -Fixed-Ratio Control at time t , then we have*

$$\mathbb{E}[(h_W(X'_{t+1}, D) - \rho_w h_W(X_t, D))^+ \mid X_t] \leq \frac{K_{W,1}}{\sqrt{N}} \quad a.s. \quad (5.54)$$

$$\begin{aligned} \mathbb{P}[h_W(X'_{t+1}, D) > \rho_w h_W(X_t, D) + r \mid X_t] &\leq K_{W,2} \exp(-C_W N r^2) \\ &a.s. \quad \forall r \geq 0, \end{aligned} \quad (5.55)$$

where $\rho_w = 1 - 1/(2\lambda_W)$, and $K_{W,1}, K_{W,2}, C_W$ are positive constants independent of N . In addition, for any $D, D' \subseteq [N]$,

$$h_W(x, D) \geq \frac{1}{|\mathbb{S}|^{1/2}} \|x(D) - m(D)\mu^*\|_1 \quad (5.56)$$

$$|h_W(x, D) - h_W(x, D')| \leq 2\lambda_W^{1/2}(m(D' \setminus D) + m(D \setminus D')), \quad (5.57)$$

The same inequalities hold for h_U with the $\bar{\pi}^*$ -Fixed-Ratio Control and W replaced by the LP-Priority and U . We omit the details here.

Proof of Lemma 5.29. Proving (5.54) and (5.55) By the pseudo-contraction properties of W -weighted norms given in Lemma 5.27,

$$\begin{aligned} \rho_w h_W(X_t, D) &= \rho_w \|X_t(D) - m(D)\mu^*\|_W \\ &\geq \|(X_t(D) - m(D)\mu^*)P_{\bar{\pi}^*}\|_W \\ &= \|X_t(D)P_{\bar{\pi}^*} - m(D)\mu^*\|_W. \end{aligned}$$

Consequently,

$$\begin{aligned} h_W(X'_{t+1}, D) - \rho_w h_W(X_t, D) &\leq \|X'_{t+1}(D) - m(D)\mu^*\|_W \\ &\quad - \|X_t(D)P_{\bar{\pi}^*} - m(D)\mu^*\|_W \\ &\leq \|X'_{t+1}(D) - X_t(D)P_{\bar{\pi}^*}\|_W \\ &\leq \lambda_W^{1/2} \|X'_{t+1}(D) - X_t(D)P_{\bar{\pi}^*}\|_2 \\ &\leq \lambda_W^{1/2} \|X'_{t+1}(D) - X_t(D)P_{\bar{\pi}^*}\|_1. \end{aligned}$$

Therefore, it suffices to show that

$$\mathbb{E}[\|X'_{t+1}(D) - X_t(D)P_{\bar{\pi}^*}\|_1 \mid X_t] = O(1/\sqrt{N}) \quad a.s. \quad (5.81)$$

$$\mathbb{P}[\|X'_{t+1}(D) - X_t(D)P_{\bar{\pi}^*}\|_1 > r \mid X_t] = O(\exp(-C\lambda_W N r^2)) \quad a.s. \quad \forall r \geq 0. \quad (5.82)$$

To show (5.81) and (5.82), we examine the distribution of $\|X'_{t+1}(D) - X_t(D)P_{\bar{\pi}^*}\|_1$ given X_t . Note that by Lemma 5.24, $\mathbb{E}[X'_{t+1}(D) \mid X_t] = X_t(D)P_{\bar{\pi}^*}$, so

$$\begin{aligned} \|X'_{t+1}(D) - X_t(D)P_{\bar{\pi}^*}\|_1 &= \|X'_{t+1}(D) - \mathbb{E}[X'_{t+1}(D) \mid X_t]\|_1 \\ &\leq \|X'_{t+1}(D) - \mathbb{E}[X'_{t+1}(D) \mid X_t, (A_t(i))_{i \in D}]\|_1 \end{aligned} \quad (5.83)$$

$$+ \|\mathbb{E}[X'_{t+1}(D) \mid X_t, (A_t(i))_{i \in D}]\|_1 \quad (5.84)$$

$$- \mathbb{E}[X'_{t+1}(D) \mid X_t]\|_1. \quad (5.85)$$

We first derive an expectation bound and a high-probability bound for (5.83). Note that for each $s \in \mathbb{S}$,

$$X'_{t+1}(D, s) = \frac{1}{N} \sum_{i \in D} \mathbb{1}\{S'_{t+1}(i) = s\},$$

where $S'_{t+1}(i)$ denote the state of arm i at time $t+1$ if the arms in D follow the $\bar{\pi}^*$ -Fixed-Ratio Control at time t . Given X_t and $(A_t(i))_{i \in D}$, $\mathbb{1}\{S'_{t+1}(i) = s\}$ for $i \in D$ are independent Bernoulli random variables, each with mean $P(S_t(i), A_t(i), s)$. By Cauchy-Schwarz,

$$\begin{aligned} & \mathbb{E} \left[\left| X'_{t+1}(D, s) - \mathbb{E}[X'_{t+1}(D, s) \mid X_t, (A_t(i))_{i \in D}] \right| \mid X_t, (A_t(i))_{i \in D} \right] \\ & \leq \text{Var} \left[X'_{t+1}(D, s) \mid X_t, (A_t(i))_{i \in D} \right]^{1/2} \\ & = \frac{1}{N} \left(\sum_{i \in D} \text{Var} \left[\mathbb{1}\{S'_{t+1}(i) = s\}, \mid X_t, (A_t(i))_{i \in D} \right] \right)^{1/2} \\ & \leq \frac{1}{\sqrt{N}}. \end{aligned}$$

By Hoeffding inequality, for all $r \geq 0$,

$$\begin{aligned} & \mathbb{P} \left[\left| X'_{t+1}(D, s) - \mathbb{E}[X'_{t+1}(D, s) \mid X_t, (A_t(i))_{i \in D}] \right| > r \mid X_t, (A_t(i))_{i \in D} \right] \\ & \leq 2 \exp(-2Nr^2). \end{aligned}$$

Therefore, summing over all $s \in \mathbb{S}$ and taking expectations over $(A_t(i))_{i \in D}$, we get

$$\begin{aligned} & \mathbb{E} \left[\left\| X'_{t+1}(D, s) - \mathbb{E}[X'_{t+1}(D, s) \mid X_t, (A_t(i))_{i \in D}] \right\|_1 \mid X_t \right] \\ & \leq \frac{|\mathbb{S}|}{\sqrt{N}} \end{aligned} \tag{5.86}$$

$$\begin{aligned} & \mathbb{P} \left[\left\| X'_{t+1}(D, s) - \mathbb{E}[X'_{t+1}(D, s) \mid X_t, (A_t(i))_{i \in D}] \right\|_1 > r \mid X_t \right] \\ & \leq 2|\mathbb{S}| \exp \left(- \frac{2Nr^2}{|\mathbb{S}|^2} \right), \end{aligned} \tag{5.87}$$

where (5.87) comes from applying a union bound.

Next, we bound the term in (5.84). For any $s \in \mathbb{S}$, we have

$$\begin{aligned}\mathbb{E} [X'_{t+1}(D, s) | X_t, (A_t(i))_{i \in D}] &= \sum_{s' \in \mathbb{S}, a \in \mathbb{A}} \Gamma_t(s', a) P(s', a, s) \\ \mathbb{E} [X'_{t+1}(D, s) | X_t] &= \sum_{s' \in \mathbb{S}, a \in \mathbb{A}} \bar{\pi}^*(a|s) X_t(D, s') P(s', a, s),\end{aligned}$$

where $N \cdot \Gamma_t(s, a)$ is the number of arms in D with state s taking action a at time t . By the definition of the $\bar{\pi}^*$ -Fixed-Ratio Control (Subroutine 1),

$$\Gamma_t(s', 1) \in \{ \lceil N \bar{\pi}^*(1|s') X_t(D, s') \rceil / N, \lfloor N \bar{\pi}^*(1|s') X_t(D, s') \rfloor / N \},$$

and $\Gamma_t(s, 0) = X_t(D, s) - \Gamma_t(s, 1)$. Consequently,

$$\begin{aligned}& \left| \mathbb{E} [X'_{t+1}(D, s) | X_t, (A_t(i))_{i \in D}] - \mathbb{E} [X'_{t+1}(D, s) | X_t] \right| \\ & \leq \sum_{s' \in \mathbb{S}, a \in \mathbb{A}} |\bar{\pi}^*(a|s') X_t(D, s') - \Gamma_t(s', a)| P(s', a, s) \\ & \leq \frac{1}{N} \sum_{s' \in \mathbb{S}, a \in \mathbb{A}} P(s', a, s) \quad a.s.\end{aligned} \tag{5.88}$$

Summing over all $s \in \mathbb{S}$ in (5.88), we get

$$\left\| \mathbb{E} [X'_{t+1}(D, s) | X_t, (A_t(i))_{i \in D}] - \mathbb{E} [X'_{t+1}(D, s) | X_t] \right\|_1 \leq \frac{2|\mathbb{S}|}{N} \quad a.s. \tag{5.89}$$

Combining the above calculations, we get

$$\begin{aligned}& \mathbb{E} [\|X'_{t+1}(D) - X_t(D) P_{\bar{\pi}^*}\|_1 | X_t] \\ & \leq \mathbb{E} [\|X'_{t+1}(D) - \mathbb{E} [X'_{t+1}(D, s) | X_t, (A_t(i))_{i \in D}]\|_1 | X_t] \\ & \quad + \mathbb{E} [\|\mathbb{E} [X'_{t+1}(D, s) | X_t, (A_t(i))_{i \in D}] - \mathbb{E} [X'_{t+1}(D, s) | X_t]\|_1 | X_t] \\ & \leq \frac{1}{\sqrt{N}} + \frac{2|\mathbb{S}|}{N} \\ & \leq \frac{3|\mathbb{S}|}{\sqrt{N}},\end{aligned} \tag{5.90}$$

which implies (5.81). We also have that for any $r \geq 2|\mathbb{S}|/N$,

$$\mathbb{P} [\|X'_{t+1}(D) - X_t(D) P_{\bar{\pi}^*}\|_1 > r | X_t]$$

$$\begin{aligned}
&\leq \mathbb{P} \left[\left\| X'_{t+1}(D) - \mathbb{E} [X'_{t+1}(D, s) \mid X_t, (A_t(i))_{i \in D}] \right\|_1 > r - \frac{2|\mathbb{S}|}{N} \mid X_t \right] \\
&\leq 2|\mathbb{S}| \exp \left(- \frac{2N(r - 2|\mathbb{S}|/N)^2}{|\mathbb{S}|^2} \right) \\
&\leq 2|\mathbb{S}| \exp \left(- \frac{Nr^2}{|\mathbb{S}|^2} + \frac{8}{N} \right) \tag{5.91}
\end{aligned}$$

where we use the fact that $(r - 2|\mathbb{S}|/N)^2 \geq r^2/2 - (2|\mathbb{S}|/N)^2$ in the (5.91). Note that the above upper bound also holds for $r < 2|\mathbb{S}|/N$ because the term in (5.91) is greater than 1 for $r < 2|\mathbb{S}|/N$. Because $\exp(8/N) \leq e^8$, we have proved (5.82):

$$\begin{aligned}
&\mathbb{P} \left[\left\| X'_{t+1}(D) - X_t(D)P_{\bar{\pi}^*} \right\|_1 > r \mid X_t \right] \\
&\leq 2|\mathbb{S}|e^8 \exp \left(- \frac{Nr^2}{|\mathbb{S}|^2} \right).
\end{aligned}$$

Combining with the bound

$$h_W(X'_{t+1}, D) - \rho_w h_W(X_t, D) \leq \lambda_W^{1/2} \left\| X'_{t+1}(D) - X_t(D)P_{\bar{\pi}^*} \right\|_1$$

established earlier and the expectation bound (5.81), we obtain

$$\mathbb{E}[(h_W(X'_{t+1}, D) - \rho_w h_W(X_t, D))^+ \mid X_t] \leq \frac{K_{W,1}}{\sqrt{N}} \quad a.s. \tag{5.54}$$

$$\begin{aligned}
\mathbb{P}[h_W(X'_{t+1}, D) > \rho_w h_W(X_t, D) + r \mid X_t] &\leq K_{W,2} \exp(-C_W N r^2) \\
&a.s. \quad \forall r \geq 0, \tag{5.55}
\end{aligned}$$

with the explicit choices $K_{W,1} = 3\lambda_W^{1/2}|\mathbb{S}|$, $K_{W,2} = 2|\mathbb{S}|e^8$, and $C_W = 1/(\lambda_W|\mathbb{S}|^2)$.

Proving the analogs of (5.54) and (5.55) for h_U . Now let X'_{t+1} denote the system at time $t+1$ given that the arms in D follow the LP-Priority at time t . The proof is almost verbatim: We first show

$$\begin{aligned}
&\mathbb{E} \left[\left\| X'_{t+1}(D, s) - \mathbb{E}[X'_{t+1}(D, s) \mid X_t, (A_t(i))_{i \in D}] \right\|_1 \mid X_t \right] \\
&\leq \frac{|\mathbb{S}|}{\sqrt{N}} \tag{5.92}
\end{aligned}$$

$$\begin{aligned}
&\mathbb{P} \left[\left\| X'_{t+1}(D, s) - \mathbb{E}[X'_{t+1}(D, s) \mid X_t, (A_t(i))_{i \in D}] \right\|_1 > r \mid X_t \right] \\
&\leq 2|\mathbb{S}| \exp \left(- \frac{2Nr^2}{|\mathbb{S}|^2} \right), \tag{5.93}
\end{aligned}$$

using identical arguments. Then we can show that

$$\left\| \mathbb{E} [X'_{t+1}(D, s) \mid X_t, (A_t(i))_{i \in D}] - \mathbb{E} [X'_{t+1}(D, s) \mid X_t] \right\|_1 \leq \frac{4|\mathbb{S}|}{N} \quad a.s., \quad (5.94)$$

using the fact that under LP-Priority (Subroutine 2), for all $s' \in \mathbb{S}$ and $a \in \mathbb{A}$,

$$\begin{aligned} |\bar{\pi}^*(a|s')X_t(D, s') - \Gamma_t(s', a)| &\leq \frac{1}{N} \quad a.s. \quad \forall s' \neq \tilde{s} \\ |\bar{\pi}^*(a|\tilde{s})X_t(D, \tilde{s}) - \Gamma_t(\tilde{s}, a)| &\leq \frac{|\mathbb{S}|}{N} \quad a.s. \end{aligned}$$

Finally, we combine (5.92), (5.93) and (5.94) to get that under h_U ,

$$\mathbb{E}[(h_U(X'_{t+1}, D) - \rho_u h_U(X_t, D))^+ \mid X_t] \leq \frac{K_{U,1}}{\sqrt{N}} \quad a.s. \quad (5.95)$$

$$\mathbb{P}[h_U(X'_{t+1}, D) > \rho_u h_U(X_t, D) + r \mid X_t] \leq K_{U,2} \exp(-C_U N r^2) \quad a.s. \quad \forall r \geq 0, \quad (5.96)$$

for explicit positive constants $K_{U,1}, K_{U,2}, C_U$ derived analogously to $K_{W,1}, K_{W,2}, C_W$.

The rest of the inequalities in this lemma. The other two inequalities, (5.56) and (5.57), and their analogs for h_U are relatively straightforward from the definitions of the weighted L_2 norms, so we skip them here. Similar calculations have also been done in Chapter 4 [153] for h_W , so we refer the readers there for the detailed calculations. $\square$

Next, we restate and prove Lemma 5.30.

Lemma 5.30 (Almost non-shrinking). *Assume Assumptions 5.1, 5.2 and 5.3 hold. Consider the two-set policy (Algorithm 10) such that (η, ϵ_N^{fe}) is feasibility-ensuring, $\eta = \Theta(1)$ and $\epsilon_N^{fe} = O(1/N)$. For any $t \geq 0$, we have*

$$\mathbb{E}[m(D_t^{\text{LP}} \setminus D_{t+1}^{\text{LP}}) \mid \Sigma_t] \leq K_{D,1} m(D_t^{\text{LP}}) \exp(-C_D N m(D_t^{\text{LP}})^2) \quad a.s. \quad (5.58)$$

$$\mathbb{P}[m(D_t^{\text{LP}} \setminus D_{t+1}^{\text{LP}}) > 0 \mid \Sigma_t] \leq K_{D,2} \exp(-C_D N m(D_t^{\text{LP}})^2) \quad a.s., \quad (5.59)$$

where $K_{D,1}, K_{D,2}, C_D$ are positive constants independent of N .

Proof of Lemma 5.30. Without loss of generality, we assume that $D_t^{\text{LP}} \neq \emptyset$, because otherwise the two inequalities stated in the lemma hold automatically. By the definition of D_{t+1}^{LP} in Algorithm 10, $D_{t+1}^{\text{LP}} \supsetneq D_t^{\text{LP}}$ only when $\delta(X_{t+1}, D_t^{\text{LP}}) < 0$; when this happens, $m(D_t^{\text{LP}} \setminus D_{t+1}^{\text{LP}}) \leq m(D_t^{\text{LP}})$. Therefore,

$$\mathbb{E}[(m(D_t^{\text{LP}} \setminus D_{t+1}^{\text{LP}}))^+ \mid \Sigma_t] \leq m(D_t^{\text{LP}}) \mathbb{P}[\delta(X_{t+1}, D_t^{\text{LP}}) < 0 \mid \Sigma_t] \quad (5.97)$$

$$\mathbb{P}[m(D_t^{\text{LP}} \setminus D_{t+1}^{\text{LP}}) > 0 \mid \Sigma_t] = \mathbb{P}[\delta(X_{t+1}, D_t^{\text{LP}}) < 0 \mid \Sigma_t]. \quad (5.98)$$

Therefore, it remains to bound $\mathbb{P}[\delta(X_{t+1}, D_t^{\text{LP}}) < 0 \mid \Sigma_t]$. Recall that $\delta(x, D) = \eta m(D) - \|x(D) - m(D)\mu^*\|_U - \epsilon_N^{\text{fe}} = \eta m(D) - h_U(x, D) - \epsilon_N^{\text{fe}}$, so

$$\mathbb{P}[\delta(X_{t+1}, D_t^{\text{LP}}) < 0 \mid \Sigma_t] = \mathbb{P}[h_U(X_{t+1}, D_t^{\text{LP}}) > \eta m(D_t^{\text{LP}}) - \epsilon_N^{\text{fe}} \mid \Sigma_t],$$

When $D_t^{\text{LP}} \neq \emptyset$, we have $\delta(X_t, D_t^{\text{LP}}) \geq 0$, implying $h_U(X_t, D_t^{\text{LP}}) \leq \eta m(D_t^{\text{LP}}) - \epsilon_N^{\text{fe}}$. Therefore,

$$\begin{aligned} & \mathbb{P}[h_U(X_{t+1}, D_t^{\text{LP}}) \geq \eta m(D_t^{\text{LP}})] \\ & \leq \mathbb{P}[h_U(X_{t+1}, D_t^{\text{LP}}) - \rho_u h_U(X_t, D_t^{\text{LP}}) \geq (1 - \rho_u) \eta m(D_t^{\text{LP}}) - (1 - \rho_u) \epsilon_N^{\text{fe}}] \\ & \leq K_{U,2} \exp(-C_U N (1 - \rho_u)^2 (\eta m(D_t^{\text{LP}}) - \epsilon_N^{\text{fe}})^2), \end{aligned} \quad (5.99)$$

where the last inequality comes from (5.96) of Lemma 5.29 with the constants $K_{U,2}, C_U$ defined therein. To bound the exponent in (5.99), observe that $1 - \rho_u = \Theta(1), \eta = \Theta(1)$ and $\epsilon_N^{\text{fe}} = O(1/N)$, so

$$\begin{aligned} & C_U N (1 - \rho_u)^2 (\eta m(D_t^{\text{LP}}) - \epsilon_N^{\text{fe}})^2 \\ & \geq C_U N (1 - \rho_u)^2 \left(\eta^2 m(D_t^{\text{LP}})^2 - 2\eta m(D_t^{\text{LP}}) \epsilon_N^{\text{fe}} \right) \\ & \geq C_U N (1 - \rho_u)^2 \eta^2 m(D_t^{\text{LP}})^2 \\ & \quad - 2C_U (1 - \rho_u)^2 \eta N \epsilon_N^{\text{fe}}. \end{aligned}$$

Since $N \epsilon_N^{\text{fe}} = O(1)$, the second term in the last expression is bounded by a constant K_0 , so

$$\mathbb{P}[\delta(X_{t+1}, D_t^{\text{LP}}) < 0 \mid \Sigma_t] \leq K_{U,2} e^{K_0} \exp(-C_D N m(D_t^{\text{LP}})^2),$$

with $C_D = C_U (1 - \rho_u)^2 \eta^2$. Combining this bound with the earlier bounds (5.97) and (5.98) for $\mathbb{E}[m(D_t^{\text{LP}} \setminus D_{t+1}^{\text{LP}})]$ and $\mathbb{P}[m(D_t^{\text{LP}} \setminus D_{t+1}^{\text{LP}}) > 0]$ yields

$$\begin{aligned} & \mathbb{E}[m(D_t^{\text{LP}} \setminus D_{t+1}^{\text{LP}}) \mid \Sigma_t] \leq K_{D,1} m(D_t^{\text{LP}}) \exp(-C_D N m(D_t^{\text{LP}})^2) \quad a.s. \\ & \mathbb{P}[m(D_t^{\text{LP}} \setminus D_{t+1}^{\text{LP}}) > 0 \mid \Sigma_t] \leq K_{D,2} \exp(-C_D N m(D_t^{\text{LP}})^2) \quad a.s., \end{aligned}$$

with $K_{D,1} = K_{D,2} = K_{U,2} e^{K_0}$. □

Finally, we restate and prove Lemma 5.31.

Lemma 5.31 (Sufficient coverage). *Assume Assumptions 5.1, 5.2 and 5.3 hold. Consider the two-set policy (Algorithm 10) such that $(\eta, \epsilon_N^{\text{fe}})$ is feasibility-ensuring, $\eta = \Theta(1)$, $\epsilon_N^{\text{fe}} = O(1/N)$, and $\epsilon_N^{\text{rd}} = O(1/N)$. For any $t \geq 0$, we have*

$$1 - m(D_t^{\text{LP}}) \leq K_{C,1} h_W(X_t, D_t^{\bar{\pi}^*}) + \frac{K_{C,2}}{N} \quad a.s. \quad (5.60)$$

$$m(D_t^{\text{LP}}) = 1 \quad \text{if } h_U(X_t, [N]) \leq \eta - \epsilon_N^{\text{fe}}, \quad (5.61)$$

where $K_{C,1}, K_{C,2}$ are positive constants independent of N .

Proof of Lemma 5.31. Observe that (5.61) follows directly from the definition of the policy, so we only need to prove (5.60). We claim that either $m(D_t^{\bar{\pi}^*}) \leq \epsilon_N^{\text{rd}}$, or

$$\lambda_U^{1/2} h_W(X_t, D_t^{\bar{\pi}^*}) > \eta m(D_t^{\bar{\pi}^*}) - \epsilon_N^{\text{rd}} - \epsilon_N^{\text{fe}}. \quad (5.100)$$

We first show (5.60) assuming this claim. Recall that $m(D_t^{\bar{\pi}^*}) = \lfloor \beta(N - |D_t^{\text{LP}}|) \rfloor / N$. If $m(D_t^{\bar{\pi}^*}) \leq \epsilon_N^{\text{rd}}$,

$$\beta(1 - m(D_t^{\text{LP}})) - \frac{1}{N} \leq m(D_t^{\bar{\pi}^*}) \leq \epsilon_N^{\text{rd}},$$

so $1 - m(D_t^{\text{LP}}) \leq 1/(\beta N) + \epsilon_N^{\text{rd}}/\beta$, which implies (5.60) by the non-negativity of $h_W(X_t, D_t^{\bar{\pi}^*})$. If $m(D_t^{\bar{\pi}^*}) > \epsilon_N^{\text{rd}}$, (5.100) holds. Because $m(D_t^{\bar{\pi}^*}) = \lfloor \beta(N - |D_t^{\text{LP}}|) \rfloor / N \geq (\beta(N - |D_t^{\text{LP}}|) - 1)/N$, we have

$$\begin{aligned} \lambda_U^{1/2} h_W(X_t, D_t^{\bar{\pi}^*}) &> \eta(\beta(N - |D_t^{\text{LP}}|) - 1) \frac{1}{N} - \epsilon_N^{\text{rd}} - \epsilon_N^{\text{fe}} \\ &= \eta\beta(1 - m(D_t^{\text{LP}})) - \frac{\eta}{N} - \epsilon_N^{\text{rd}} - \epsilon_N^{\text{fe}}. \end{aligned}$$

Rearranging gives

$$1 - m(D_t^{\text{LP}}) < \frac{\lambda_U^{1/2}}{\eta\beta} h_W(X_t, D_t^{\bar{\pi}^*}) + \frac{1}{\beta N} + \frac{\epsilon_N^{\text{rd}} + \epsilon_N^{\text{fe}}}{\eta\beta}.$$

Combining the two cases, we have

$$1 - m(D_t^{\text{LP}}) \leq \frac{\lambda_U^{1/2}}{\eta\beta} h_W(X_t, D_t^{\bar{\pi}^*}) + \frac{1}{\beta N} + \frac{\epsilon_N^{\text{rd}} + \epsilon_N^{\text{fe}}}{\min(\eta, 1)\beta}.$$

Since $\epsilon_N^{\text{rd}}, \epsilon_N^{\text{fe}} = O(1/N)$, the additive term is $O(1/N)$, and we can write the bound as $K_{C,1}h_W(X_t, D_t^{\bar{\pi}^*}) + K_{C,2}/N$ for positive constants $K_{C,1}, K_{C,2}$ independent of N , establishing (5.60).

Now we prove the claim by contradiction. Suppose, at a certain time t , we have $m(D_t^{\bar{\pi}^*}) > \epsilon_N^{\text{rd}}$ and $\lambda_U^{1/2}h_W(X_t, D_t^{\bar{\pi}^*}) \leq \eta m(D_t^{\bar{\pi}^*}) - \epsilon_N^{\text{rd}} - \epsilon_N^{\text{fe}}$. Because $\|v\|_U \leq \lambda_U^{1/2} \|v\|_2 \leq \lambda_U^{1/2} \|v\|_W$ for any $v \in \mathbb{R}^{|\mathcal{S}|}$,

$$\begin{aligned} & \|X_t(D_t^{\bar{\pi}^*}) - m(D_t^{\bar{\pi}^*})\mu^*\|_U \\ & \leq \lambda_U^{1/2} \|X_t(D_t^{\bar{\pi}^*}) - m(D_t^{\bar{\pi}^*})\mu^*\|_W \leq \eta m(D_t^{\bar{\pi}^*}) - \epsilon_N^{\text{rd}} - \epsilon_N^{\text{fe}}. \end{aligned}$$

Combined with the triangular inequality and the definition of D_t^{LP} , we have

$$\begin{aligned} & \|X_t(D_t^{\text{LP}} \cup D_t^{\bar{\pi}^*}) - m(D_t^{\text{LP}} \cup D_t^{\bar{\pi}^*})\mu^*\|_U \\ & \leq \|X_t(D_t^{\text{LP}}) - m(D_t^{\text{LP}})\mu^*\|_U \\ & \quad + \|X_t(D_t^{\bar{\pi}^*}) - m(D_t^{\bar{\pi}^*})\mu^*\|_U \\ & \leq \eta m(D_t^{\text{LP}}) + \eta m(D_t^{\bar{\pi}^*}) - \epsilon_N^{\text{rd}} - \epsilon_N^{\text{fe}}. \end{aligned}$$

Consequently, $D_t^{\text{LP}} \cup D_t^{\bar{\pi}^*}$ is a superset of D_t^{LP} such that $\delta(X_t, D_t^{\text{LP}} \cup D_t^{\bar{\pi}^*}) \geq \epsilon_N^{\text{rd}}$ and $m(D_t^{\text{LP}} \cup D_t^{\bar{\pi}^*}) = m(D_t^{\text{LP}}) + m(D_t^{\bar{\pi}^*}) > m(D_t^{\text{LP}}) + \epsilon_N^{\text{rd}}$, contradicting the ϵ_N^{rd} -maximality of D_t^{LP} . We have thus proved the claim that implies (5.60). $\square$

5.C.3 Proof of Lemma 5.18

Lemma 5.18 (Instantaneous reward under LP-Priority). *Under the two-set policy, for any $\sigma \triangleq (x, D^{\text{LP}}, D^{\bar{\pi}^*}) \in \mathbb{X}$, if $\|x([N]) - \mu^*\|_U \leq \bar{\eta}$, we have*

$$r^\pi(\sigma) = \hat{r}(x([N])),$$

where $\hat{r}: \Delta(\mathcal{S}) \rightarrow \mathbb{R}$ is the affine function

$$\hat{r}(v) \triangleq \sum_{s \neq \tilde{s}, a \in \mathcal{A}} (r(s, a) - r(\tilde{s}, a)) \bar{\pi}^*(a|s) v(s) + \alpha r(\tilde{s}, 1) + (1 - \alpha) r(\tilde{s}, 0), \quad (5.22)$$

which can be written in vector form as $\hat{r}(v) = v \cdot \mathbf{g} + r_0$ for some $\mathbf{g} \in \mathbb{R}^{|\mathcal{S}|}$ and $r_0 \in \mathbb{R}$.

Proof. By (5.10), when $\|x([N]) - \mu^*\|_U \leq \bar{\eta} \triangleq \eta - \epsilon_N^{\text{fe}}$, LP-Priority is feasible for all arms. Since the two-set policy chooses D^{LP} to be a maximal set that can follow LP-Priority, we have $D^{\text{LP}} = [N]$ and $D^{\bar{\pi}^*} = \emptyset$.

By definition,

$$r^\pi(\sigma) = r^\pi(x, [N], \emptyset) = \sum_{s \in \mathbb{S}, a \in \mathbb{A}} r(s, a) \mathbb{E}[Y_t(s, a) \mid X_t, D_t^{\text{LP}} = [N], D_t^{\bar{\pi}^*} = \emptyset], \quad (5.101)$$

where t is an arbitrary time step, and $Y_t(s, a)$ is the random variable representing the fraction of arms in state s taking action a at time t . By the definition of LP-Priority,

$$\mathbb{E}[Y_t(s, a) \mid X_t, D_t^{\text{LP}} = [N], D_t^{\bar{\pi}^*} = \emptyset] = \bar{\pi}^*(a|s) X_t([N], s) \quad s \neq \tilde{s} \quad (5.102)$$

$$\mathbb{E}[Y_t(\tilde{s}, 1) \mid X_t, D_t^{\text{LP}} = [N], D_t^{\bar{\pi}^*} = \emptyset] = \alpha - \sum_{s \neq \tilde{s}} \bar{\pi}^*(1|s) X_t([N], s) \quad (5.103)$$

$$\mathbb{E}[Y_t(\tilde{s}, 0) \mid X_t, D_t^{\text{LP}} = [N], D_t^{\bar{\pi}^*} = \emptyset] = 1 - \alpha - \sum_{s \neq \tilde{s}} \bar{\pi}^*(0|s) X_t([N], s). \quad (5.104)$$

Therefore,

$$\begin{aligned} r^\pi(X_t, [N], \emptyset) &= \sum_{s \neq \tilde{s}, a \in \mathbb{A}} (r(s, a) - r(\tilde{s}, a)) \bar{\pi}^*(a|s) X_t([N], s) \\ &\quad + \alpha r(\tilde{s}, 1) + (1 - \alpha) r(\tilde{s}, 0), \end{aligned} \quad (5.105)$$

For any $v \in \Delta(\mathbb{S})$, we define $\hat{r}(v)$ as

$$\hat{r}(v) \triangleq \sum_{s \neq \tilde{s}, a \in \mathbb{A}} (r(s, a) - r(\tilde{s}, a)) \bar{\pi}^*(a|s) v(s) + \alpha r(\tilde{s}, 1) + (1 - \alpha) r(\tilde{s}, 0).$$

Then $\hat{r}(\cdot)$ is affine v , and $r^\pi(\sigma) = \hat{r}(X_t([N]))$. $\square$

5.C.4 Proof of Lemma 5.19

Lemma 5.19 (Drift of truncated V_1). *For any $\sigma, \sigma' \in \mathbb{X}$,*

$$\begin{aligned} (V_1(\sigma') - \frac{\bar{\eta}}{2})^+ - (V_1(\sigma) - \frac{\bar{\eta}}{2})^+ &\leq -\frac{(1 - \rho_1)\bar{\eta}}{2} \mathbf{1}\{V_1(\sigma) > \bar{\eta}\} \\ &\quad + \left(V_1(\sigma') - \rho_1 V_1(\sigma) - \frac{(1 - \rho_1)\bar{\eta}}{2} \right)^+. \end{aligned} \quad (5.27)$$

Proof. We use the abbreviation that $a \triangleq V_1(\sigma)$, $b \triangleq V_1(\sigma')$, and $r_0 \triangleq (1 - \rho_1)\bar{\eta}/2$. The key algebraic identity

$$b - \frac{\bar{\eta}}{2} = \left(b - \rho_1 a - r_0 \right) + \rho_1 \left(a - \frac{\bar{\eta}}{2} \right), \quad (5.106)$$

together with subadditivity $(x + y)^+ \leq x^+ + y^+$, yields

$$\left(b - \frac{\bar{\eta}}{2}\right)^+ - \left(a - \frac{\bar{\eta}}{2}\right)^+ \leq \left(b - \rho_1 a - r_0\right)^+ - (1 - \rho_1) \left(a - \frac{\bar{\eta}}{2}\right)^+. \quad (5.107)$$

We derive the claim from (5.107) by cases on a .

Case 1: $a \leq \bar{\eta}$. Since $(1 - \rho_1)(a - \bar{\eta}/2)^+ \geq 0$, (5.107) gives

$$\left(b - \frac{\bar{\eta}}{2}\right)^+ - \left(a - \frac{\bar{\eta}}{2}\right)^+ \leq \left(b - \rho_1 a - r_0\right)^+,$$

which matches the RHS of (5.27) because $\mathbb{1}\{a > \bar{\eta}\} = 0$.

Case 2: $a > \bar{\eta}$. Then $(a - \bar{\eta}/2)^+ = a - \bar{\eta}/2 > \bar{\eta}/2$, so $(1 - \rho_1)(a - \bar{\eta}/2)^+ > (1 - \rho_1)\bar{\eta}/2 = r_0$. Hence (5.107) gives

$$\left(b - \frac{\bar{\eta}}{2}\right)^+ - \left(a - \frac{\bar{\eta}}{2}\right)^+ \leq \left(b - \rho_1 a - r_0\right)^+ - r_0,$$

which matches the RHS of (5.27) because $\mathbb{1}\{a > \bar{\eta}\} = 1$. $\square$

5.C.5 Proof of Lemma 5.20

Lemma 5.20 (Drift of Ψ). *For any $\sigma = (x, D^{\text{LP}}, D^{\bar{\pi}^*}) \in \mathbb{X}$,*

$$\Delta\Psi(\sigma) \leq -(R^{\text{rel}} - r^\pi(\sigma)) + (4r_{\max}\lambda_Q + 2r_{\max}) \mathbb{1}\{\|x([N]) - \mu^*\|_U > \bar{\eta}\}, \quad (5.28)$$

where $\Psi(\sigma) = (\mu^* - x([N])) Q \mathbf{g}$, $Q = (I - \Phi)^{-1}$, and $\bar{\eta} = \eta - \epsilon_N^e$.

Proof.

$$\begin{aligned} \Delta\Psi(\sigma) &= \mathbb{E}[(\mu^* - X_{t+1}([N])) Q \mathbf{g} \mid \Sigma_t = \sigma] - (\mu^* - x([N])) Q \mathbf{g} \\ &= \mathbb{E}[(\mu^* - X_{t+1}([N])) Q \mathbf{g} \mid \Sigma_t = \sigma] \\ &\quad - (\mu^* - x([N])) (\Phi - \Phi + I) Q \mathbf{g} \\ &= \mathbb{E}[(\mu^* - X_{t+1}([N])) Q \mathbf{g} \mid \Sigma_t = \sigma] - (\mu^* - x([N])) \Phi Q \mathbf{g} \\ &\quad - (\mu^* - x([N])) \mathbf{g} \\ &= (\mathbb{E}[(\mu^* - X_{t+1}([N])) Q \mathbf{g} \mid \Sigma_t = \sigma] - (\mu^* - x([N])) \Phi Q \mathbf{g}) \quad (5.108) \\ &\quad - (R^{\text{rel}} - \hat{r}(x([N]))). \end{aligned}$$

We claim that the expression in (5.108) is bounded as:

$$\begin{aligned} &\mathbb{E}[(\mu^* - X_{t+1}([N])) Q \mathbf{g} \mid \Sigma_t = \sigma] - (\mu^* - x([N])) \Phi Q \mathbf{g} \\ &\leq 4r_{\max}\lambda_Q \mathbb{1}\{D^{\text{LP}} \neq [N]\}. \end{aligned} \quad (5.109)$$

We verify (5.109) by cases on D^{LP} .

- Case $D^{\text{LP}} = [N]$. By Lemma 5.8, $\mathbb{E}[X_{t+1}([N]) - \mu^* | \Sigma_t = \sigma] = (x([N]) - \mu^*)\Phi$, so

$$\mathbb{E}[(\mu^* - X_{t+1}([N]))Q\mathbf{g} | \Sigma_t = \sigma] = (\mu^* - x([N]))\Phi Q\mathbf{g},$$

and (5.109) holds with both sides equal to zero.

- Case $D^{\text{LP}} \neq [N]$. By Jensen's inequality and the triangle inequality,

$$\begin{aligned} & \mathbb{E}[(\mu^* - X_{t+1}([N]))Q\mathbf{g} | \Sigma_t = \sigma] - (\mu^* - x([N]))\Phi Q\mathbf{g} \\ & \leq \mathbb{E}[|(X_{t+1}([N]) - \mu^*)Q\mathbf{g}| | \Sigma_t] + |(x([N]) - \mu^*)\Phi Q\mathbf{g}| \\ & \leq \mathbb{E}[\|(X_{t+1}([N]) - \mu^*)Q\|_1 | \Sigma_t] \|\mathbf{g}\|_\infty + \|(x([N]) - \mu^*)\Phi Q\|_1 \|\mathbf{g}\|_\infty \\ & \leq 2\lambda_Q \|\mathbf{g}\|_\infty \\ & \leq 4r_{\max}\lambda_Q, \end{aligned}$$

where the second last step uses $X_{t+1}([N]), x([N]) \in \Delta(\mathbb{S})$ and the definition of λ_Q , and the last step uses $\|\mathbf{g}\|_\infty \leq 2r_{\max}$.

Substituting (5.109) into (5.108), we get

$$\begin{aligned} \Delta\Psi(\sigma) & \leq -(R^{\text{rel}} - \hat{r}(x([N]))) + 4r_{\max}\lambda_Q \mathbb{1}\{D^{\text{LP}} \neq [N]\} \\ & \leq -(R^{\text{rel}} - \hat{r}(x([N]))) + 4r_{\max}\lambda_Q \mathbb{1}\{\|x([N]) - \mu^*\|_U > \bar{\eta}\}. \end{aligned}$$

where the last step is due to $\mathbb{1}\{D^{\text{LP}} \neq [N]\} \leq \mathbb{1}\{\|x([N]) - \mu^*\|_U > \bar{\eta}\}$. Because $\hat{r}(x([N])) = r^\pi(\sigma)$ when $\|x([N]) - \mu^*\|_U > \bar{\eta}$ and $\hat{r}(x([N])) \leq r^\pi(\sigma) + 2r_{\max}$ otherwise, we get

$$\Delta\Psi(\sigma) \leq -(R^{\text{rel}} - r^\pi(\sigma)) + (4r_{\max}\lambda_Q + 2r_{\max})\mathbb{1}\{\|x([N]) - \mu^*\|_U > \bar{\eta}\}. \quad \square$$

5.D Examples and proofs of violating assumptions

5.D.1 Examples of violating Assumptions 5.1 to 5.2

An aperiodic but non-unichain example.

We first give an example of RBs where $P_{\bar{\pi}^*}$ is aperiodic but not unichain. Consider the RB problem whose single-armed MDP has two states, A and B , with no transitions between these two states. The reward function is given

by $r(A, 1) = r(B, 0) = 1$ and $r(A, 0) = r(B, 1) = 0$. Let $\alpha = 1/2$ in the budget constraint. Since there is no transition between these two states, the Markov chain induced by the optimal single-armed policy, $\bar{\pi}^*$, $P_{\bar{\pi}^*}$, is not unichain. However, each of the recurrent classes $\{A\}$ and $\{B\}$ is aperiodic in this Markov chain.

It is easy to verify that the optimal solution to the LP relaxation (LP) is $y^*(A, 1) = y^*(B, 0) = 1/2$ and $y^*(s, a) = 0$ for the rest of $s \in \mathbb{S}, a \in \mathbb{A}$; the optimal value of (LP) is $R^{\text{rel}} = 1$. Now consider the N -armed problem where all arms are initialized in state A . In this case, the long-run average reward of any policy is $1/2$. Therefore, $R^{\text{rel}} - R(\pi, \mathbf{S}_0) \geq 1/2 = \Omega(1/\sqrt{N})$ for any π .

A unichain but periodic example.

Next, we give an example of RBs where $P_{\bar{\pi}^*}$ is unichain but periodic, and $R^{\text{rel}} - R(\pi, \mathbf{S}_0) = \Theta(1)$. This example was included in Chapter 4 [153], and we paraphrase it here for completeness.

Consider the RB problem whose single-armed MDP has two states, A and B . The state of the single-armed MDP deterministically transitions to the other state in each time step, regardless of the action applied. Let the reward function be $r(A, 0) = r(B, 1) = 1$ and $r(A, 1) = r(B, 0) = 0$; let $\alpha = 1/2$ in the budget constraint. Obviously, the Markov chain induced by any single-armed policy is unichain and periodic with a period 2.

One can easily see the optimal solution to the LP relaxation (LP) is by $y^*(A, 0) = y^*(B, 1) = 1/2$ and $y^*(A, 1) = y^*(B, 0) = 0$; the optimal value of (LP) is $R^{\text{rel}} = 1$. Now consider the N -armed problem where all arms are initialized in state A . Then at any time t , either all arms are in state A or all arms are in state B . In this case, under any policy π , the long-run average $R(\pi, \mathbf{S}_0)$ is $1/2$, so $R^{\text{rel}} - R(\pi, \mathbf{S}_0) = 1/2 = \Omega(1/\sqrt{N})$.

An example satisfying Assumption 5.1 but not Assumption 5.2.

Next, we give an example of RB satisfying Assumption 5.1 but violating Assumption 5.2. Suppose each arm in the N -armed problem has state $S_t(i) = \text{Ber}(1/2)$ i.i.d. across i and t . Let $\alpha = 1/2$ in the budget constraint. The reward function $r(s, 1) = s$ and $r(s, 0) = 0$ for $s = 0, 1$. Intuitively, the decision maker observes the state of each arm, and pulls $N/2$ of them. Each arm pulled by the decision maker generates an amount of reward equal to its

state. The single-armed MDP is clearly aperiodic unichain under any policy because its state is i.i.d.

The optimal solution of the LP relaxation (LP) is $y^*(1, 1) = y^*(0, 0) = 1/2$, and $y^*(s, a) = 0$ for other $s \in \mathbb{S}, a \in \mathbb{A}$; the optimal value $R^{\text{rel}} = 1/2$. Intuitively, the optimal single-armed policy pulls the arm if and only if it is in state 1.

In the N -armed problem, the optimal policy is obviously pulling $N/2$ arms with the largest states. Then the optimal reward $R^*(N, \mathbf{S}_0)$ is given by

$$R^*(N, \mathbf{S}_0) = \frac{1}{N} \mathbb{E} \left[\min \left(\sum_{i \in [N]} S_i(t), \frac{N}{2} \right) \right],$$

for an arbitrary t . Since $S_t(i) = \text{Ber}(1/2)$ and are i.i.d. across i , it can be shown that $R^*(N, \mathbf{S}_0) = 1/2 - \Theta(1/\sqrt{N})$. Therefore, $R^{\text{rel}} - R^*(N, \mathbf{S}_0) = \Theta(1/\sqrt{N})$.

5.D.2 Proofs of Lemma 5.32 and Theorem 5.15

In this section, we prove our second main result, Theorem 5.15, which is a lower bound on $R^{\text{rel}} - R(\pi, \mathbf{S}_0)$ for regular unstable RBs (Definition 5.14); meanwhile we will also prove Lemma 5.32 stated in Section 5.4.2 as one of the preliminary steps for proving Theorem 5.15. This section is organized linearly, with a sequence of lemma statements and proofs; Theorem 5.15 will be proved at the end of this section.

We first restate and prove Lemma 5.32.

Lemma 5.32. *Suppose that the LP relaxation (LP) has a unique optimal solution y^* , and $y^*(s, 1) + y^*(s, 0) > 0$, $\forall s \in \mathbb{S}$. Then*

$$R^{\text{rel}} - R(\pi, \mathbf{S}_0) \geq \epsilon_0 \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[d_{\text{IC}}(Y_t^\pi)], \quad (5.110)$$

where ϵ_0 is a positive constant; d_{IC} is a function of state-action distribution given by

$$d_{\text{IC}}(y) = \sum_{s \in S^+} y(s, 0) + \sum_{s \in S^-} y(s, 1),$$

where $S^+ = \{s \in \mathbb{S}: y^*(s, 1) > 0, y(s, 0) = 0\}$, $S^- = \{s \in \mathbb{S}: y^*(s, 1) = 0, y^*(s, 0) > 0\}$.

Proof. We first investigate the dual of (LP), whose properties help us prove the lemma:

$$\begin{aligned}
& \underset{\lambda, \mu, \{f(s)\}_{s \in \mathbb{S}}}{\text{minimize}} \mu && \text{(Dual-LP)} \\
& \text{subject to} \quad r(s, a) + (\alpha - \mathbf{1}\{a = 1\})\lambda - \mu \\
& \quad + \sum_{s'} P(s, a, s')f(s') \leq f(s) \quad \forall s \in \mathbb{S}, a \in \mathbb{A} && (5.111) \\
& \quad \lambda, \mu, f(s) \in \mathbb{R} \quad \forall s \in \mathbb{S}
\end{aligned}$$

Note that the primal problem (LP) is always feasible and bounded, so (Dual-LP) is also feasible and bounded. By strong duality of linear programs, the optimal value of the dual problem is $\mu^* = R^{\text{rel}}$. Moreover, there always exists an optimal solution of the dual problem (λ^*, f^*, μ^*) that satisfies strict complementarity with the unique primal optimal solution y^* [120]:

$$\begin{aligned}
& r(s, a) - \lambda^* \mathbf{1}\{a = 1\} + \lambda^* \alpha - \mu^* + \sum_{s'} P(s, a, s')f^*(s') - f^*(s) = 0 \\
& \text{iff} \quad y^*(s, a) \neq 0.
\end{aligned} \tag{5.112}$$

Let $r_{\lambda^*}(s, a) = r(s, a) - \lambda^* \mathbf{1}\{a = 1\} + \lambda^* \alpha$. By (5.112) and the definition of S^+ and S^- , there exists a constant $\epsilon_0 > 0$ such that

$$r_{\lambda^*}(s, 0) - R^{\text{rel}} + \sum_{s'} P(s, 0, s')f^*(s') \leq f^*(s) - \epsilon_0 \quad \forall s \in S^+ \tag{5.113}$$

$$r_{\lambda^*}(s, 1) - R^{\text{rel}} + \sum_{s'} P(s, 1, s')f^*(s') \leq f^*(s) - \epsilon_0 \quad \forall s \in S^-. \tag{5.114}$$

Define the function $Q^*: \mathbb{S} \times \mathbb{A} \rightarrow \mathbb{R}$ as

$$Q^*(s, a) = r_{\lambda^*}(s, a) - R^{\text{rel}} + \sum_{s'} P(s, a, s')f^*(s').$$

Then (5.111) implies that $f^*(s) \geq Q^*(s, a)$ for any $s \in \mathbb{S}, a \in \mathbb{A}$; (5.113) and (5.114) imply that $Q^*(s, 1) - Q^*(s, 0)$ is at least ϵ_0 for $s \in S^+$ and at most $-\epsilon_0$ for $s \in S^-$.

Now, we are ready to bound $R^{\text{rel}} - R(\pi, \mathbf{S}_0)$: for any policy π and initial state vector $\mathbf{S}_0$,

$$R^{\text{rel}} - R(\pi, \mathbf{S}_0)$$

$$\begin{aligned}
&= \mathbb{E} \left[R^{\text{rel}} - \frac{1}{N} \sum_{i \in [N]} r_{\lambda^*}(S_\infty(i), A_\infty(i)) \right] \\
&= \frac{1}{N} \sum_{i \in [N]} \mathbb{E} \left[R^{\text{rel}} - r_{\lambda^*}(S_\infty(i), A_\infty(i)) \right] \\
&= \frac{1}{N} \sum_{i \in [N]} \mathbb{E} \left[R^{\text{rel}} - r_{\lambda^*}(S_\infty(i), A_\infty(i)) \right. \\
&\quad \left. - \sum_{s' \in \mathbb{S}} P(S_\infty(i), A_\infty(i), s') f^*(s') + f^*(S_\infty(i)) \right] \tag{5.115}
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{N} \sum_{i \in [N]} \mathbb{E} \left[-Q^*(S_\infty(i), A_\infty(i)) + f^*(S_\infty(i)) \right] \\
&\geq \frac{1}{N} \sum_{i \in [N]} \epsilon_0 \mathbb{E} \left[\mathbf{1}\{S_\infty(i) \in S^+, A_\infty(i) = 0\} \right. \\
&\quad \left. + \mathbf{1}\{S_\infty(i) \in S^-, A_\infty(i) = 1\} \right] \tag{5.116}
\end{aligned}$$

$$= \epsilon_0 \mathbb{E} \left[\sum_{s \in S^+} Y_\infty(s, 0) + \sum_{s \in S^-} Y_\infty(s, 1) \right]. \tag{5.117}$$

where to get (5.115), we use the fact that $\mathbb{E}[\sum_{s' \in \mathbb{S}} P(S_\infty(i), A_\infty(i), s') f^*(s')] = \mathbb{E}[f^*(S_\infty(i))]$ for any function $f: \mathbb{S} \rightarrow \mathbb{R}$. $\square$

Next, we show Lemma 5.33, which implies that under any actions for the N -arms, one can find a subset of arms (can be empty) such that these almost follow the LP-Priority; the size of the subset is determined by $d_{\text{IC}}(Y_t)$. Essentially, this lemma shows that under any policy, if $d_{\text{IC}}(Y_t)$ is small, the transition dynamics will be close to the transition dynamics under the LP-Priority.

Lemma 5.33. *Consider an instance of regular unstable RBs (Definition 5.14). Under any policy π , at any time t , there exists a set $D_t \subseteq [N]$ depending on X_t and $\mathbf{A}_t$ with size $m(D_t) \geq 1 - 2d_{\text{IC}}(Y_t)/\beta - 1/(\beta N)$, where $\beta = \min(\alpha, 1 - \alpha)$, such that if we resample actions for the arms in D_t using the LP-Priority, we get almost the same actions as A_t up to at most 1 arm. For such D_t , we have*

- The expectation of $X_{t+1}(D_t)$ satisfies

$$\left\| \mathbb{E}[X_{t+1}(D_t) - m(D_t)\mu^* \mid X_t, \mathbf{A}_t] - (X_t(D_t) - m(D_t)\mu^*)\Phi \right\|_1 \leq \frac{2}{N}. \tag{5.118}$$

- Let $X'_{t+1}(D_t)$ be the state distribution at time $t+1$ when the actions for the arms in D_t are resampled using the LP-Priority, then there exists a coupling between X'_{t+1} and X_{t+1} such that

$$\|X'_{t+1}(D_t) - X_{t+1}(D_t)\|_1 \leq \frac{2}{N} \quad a.s. \quad (5.119)$$

Proof. We construct the subset D_t by removing a suitable subset of arms from $[N]$. We first remove each arm i with state $S_t(i) \in S^+$ and $A_t(i) = 0$ or $S_t(i) \in S^-$ and $A_t(i) = 1$. After this step, we get a subset D' with $N(1 - d_{\text{IC}}(Y_t))$ arms. The budget usage of D' equals $\alpha N - N \sum_{s \in S^-} Y_t(s, 1)$. Then D_t is obtained by removing arms that take action 1 or action 0 from D' such that the budget usage of D_t equals $\lfloor \alpha |D_t| \rfloor$. If $\alpha N - N \sum_{s \in S^-} Y_t(s, 1) > \alpha |D'|$, it is not hard to show that the number of arms removed from D' is bounded by

$$\begin{aligned} |D'| - |D_t| &\leq \frac{1}{1 - \alpha} \left(\left| \alpha N - N \sum_{s \in S^-} Y_t(s, 1) - \alpha |D'| \right| + 1 \right) \\ &\leq \frac{1}{1 - \alpha} (N d_{\text{IC}}(Y_t) + 1); \end{aligned}$$

If $\alpha N - N \sum_{s \in S^-} Y_t(s, 1) > \alpha |D'|$, it is not hard to show that the number of arms removed from D' is bounded by

$$\begin{aligned} |D'| - |D_t| &\leq \frac{1}{\alpha} \left(\left| \alpha N - N \sum_{s \in S^-} Y_t(s, 1) - \alpha |D'| \right| + 1 \right) \\ &\leq \frac{1}{\alpha} (N d_{\text{IC}}(Y_t) + 1). \end{aligned}$$

In both cases, we have

$$|D'| - |D_t| \leq \frac{1}{\beta} (N d_{\text{IC}}(Y_t) + 1), \quad (5.120)$$

where recall that $\beta = \min(\alpha, 1 - \alpha)$. Now we get a subset D_t such that $m(D_t) \geq 1 - d_{\text{IC}}(Y_t) - (d_{\text{IC}}(Y_t) + 1/N)/\beta \geq 1 - 2d_{\text{IC}}(Y_t)/\beta - 1/(\beta N)$. Moreover, for each $i \in D_t$, $A_t(i) = 1$ if $S_t(i) \in S^+$; $A_t(i) = 0$ if $S_t(i) \in S^-$; the total budget usage of D_t equals $\lfloor \alpha |D_t| \rfloor$.

Following similar calculations as the proof of Lemma 5.8, we can show that

$$\mathbb{E}[X_{t+1}(D_t) | X_t, \mathbf{A}_t] = X_t(D_t) \left(P_{\bar{\pi}^*} - c_{\bar{\pi}^*}^\top (P_1(\tilde{s}) - P_0(\tilde{s})) \right)$$

$$+ m(D_t) \frac{\lfloor \alpha N \rfloor}{N} (P_1(\tilde{s}) - P_0(\tilde{s})).$$

whereas by the definition of Φ ,

$$\begin{aligned} & (X_t(D_t) - m(D_t)\mu^*)\Phi + m(D_t)\mu^* \\ &= X_t(D_t) \left(P_{\pi^*} - c_{\pi^*}^\top (P_1(\tilde{s}) - P_0(\tilde{s})) \right) + m(D_t)\alpha(P_1(\tilde{s}) - P_0(\tilde{s})). \end{aligned}$$

Comparing the above two equations, we get (5.118).

To show (5.119), note that if we resample the actions using the LP-Priority for the arms in D_t , the actions of the arms in S^+ and S^- are unchanged; at most one arm with state $\tilde{s}$ gets a different action after resampling. Therefore, we can couple X'_{t+1} and X_{t+1} such that for each arm that gets the same actions after resampling, the next states of the arm are the same. In this way, we get $\|X'_{t+1}(D_t) - X_{t+1}(D_t)\|_1 \leq 2/N$. $\square$

Next, we conduct Lyapunov analysis with the help of Lemma 5.33. To construct Lyapunov functions, we first define two seminorms. Consider the Jordan decomposition of Φ , $\Phi = VJV^{-1}$. We split J into $J = J_+ + J_-$, where J_+ contains the Jordan blocks of J with eigenvalues' moduli greater than 1, and J_- contains the Jordan blocks of J with eigenvalues' moduli smaller than 1. Let I_+ be the diagonal matrix with 1 on the entries where J_+ is non-zero; let I_- be the identity matrix subtracting I_+ . We define the matrices U_+ and U_- via the infinite series:

$$U_+ = \sum_{k=1}^{\infty} (VJ_+^\dagger V^{-1})^k (VJ_+^\dagger V^{-1})^{\top k} \quad (5.121)$$

$$U_- = VI_-V^{-1}(VI_-V^{-1})^\top + \sum_{k=1}^{\infty} (VJ_-V^{-1})^k (VJ_-V^{-1})^{\top k} \quad (5.122)$$

where the superscript $\dagger$ denotes pseudo-inverse.

In the next lemma, we show that U_+ and U_- are well-defined, and prove several useful facts about the semi-norms weighted by U_+ and U_- .

Lemma 5.34. *For any regular unstable RBs (Definition 5.14), the matrices U_+ and U_- given in (5.121) and (5.122) are well-defined and positive semidefinite. Further, if we let $\|v\|_{U_+} = \sqrt{vU_+v^\top}$ and $\|v\|_{U_-} = \sqrt{vU_-v^\top}$ for any $v \in \mathbb{R}^{|\mathcal{S}|}$, then*

$$\|v\Phi\|_{U_+} \geq (1 + \epsilon_1) \|v\|_{U_+} \quad (5.123)$$

$$\|v\Phi\|_{U_+} \leq (1 + \epsilon_2) \|v\|_{U_+} \quad (5.124)$$

$$\|v\Phi\|_{U_-} \leq (1 - \epsilon_3) \|v\|_{U_-}, \quad (5.125)$$

where $\epsilon_1, \epsilon_2, \epsilon_3$ are positive constants. Moreover, there exists a positive constant K_{norm} such that for any $v \in \mathbb{R}^{|\mathcal{S}|}$,

$$\|v\|_{U_+} + \|v\|_{U_-} \geq K_{\text{norm}} \|v\|_1. \quad (5.126)$$

Proof. To show the well-definedness U_+ and U_- , note that the moduli of all eigenvalues of $VJ_+^\dagger V^{-1}$ and $VJ_- V^{-1}$ are strictly less than 1, so their k -th powers must converge exponentially as $k \rightarrow \infty$, implying the convergence of the series in (5.121) and (5.122). Moreover, since each term in the two series is positive semi-definite, both series converge to positive semi-definite matrices.

Next, we show (5.123). Observe that

$$\begin{aligned} \Phi U_+ \Phi^\top &= \sum_{k=1}^{\infty} (VJJ_+^{\dagger k} V^{-1}) (VJJ_+^{\dagger k} V^{-1})^\top \\ &= (VJJ_+^\dagger V^{-1}) (VJJ_+^\dagger V^{-1})^\top + \sum_{k=2}^{\infty} (VJJ_+^{\dagger k} V^{-1}) (VJJ_+^{\dagger k} V^{-1})^\top \\ &= (VI_+ V^{-1}) (VI_+ V^{-1})^\top + \sum_{k=2}^{\infty} (VJ_+^{\dagger(k-1)} V^{-1}) (VJ_+^{\dagger(k-1)} V^{-1})^\top \\ &= (VI_+ V^{-1}) (VI_+ V^{-1})^\top + U_+. \end{aligned}$$

We claim that there exists a positive constant ϵ'_1 such that for any $v \in \mathbb{R}^{|\mathcal{S}|}$,

$$v(VI_+ V^{-1}) (VI_+ V^{-1})^\top v^\top \geq \epsilon'_1 v U_+ v^\top, \quad (5.127)$$

Without loss of generality, we only need to show (5.127) for v not being the zero vector. By normalization, it suffices to show that the function $v \mapsto v U_+ v^\top$ is bounded on the set $\{v: \|v(VI_+ V^{-1})\|_2 = 1\}$. Let $\mathcal{V}_+$ be the linear subspace spanned by rows of V corresponding to the non-zero entries on I_+ 's diagonal, and let $\mathcal{V}_-$ be the linear subspace spanned by rows of V corresponding to the zero entries on I_+ 's diagonal. Observe that $\mathbb{R}^{|\mathcal{S}|}$ can be represented as the direct sum $\mathcal{V}_+ \oplus \mathcal{V}_-$, which implies that any $v \in \mathbb{R}^{|\mathcal{S}|}$ can be decomposed as $v = v^\parallel + v^\perp$, where $v^\parallel \in \mathcal{V}_+$ and $v^\perp \in \mathcal{V}_-$. Moreover, $\mathcal{V}_-$ is the kernel of $VI_+ V^{-1}$, and $VI_+ V^{-1}$ is the identity mapping when restricted

to $\mathcal{V}_+$. Therefore, if $\|v(VI_+V^{-1})\|_2 = 1$, then $\|v^\parallel\|_2 = 1$. Observe that $\mathcal{V}_-$ is also the kernel of U_- , so

$$vU_+v^\top = (v^\parallel + v^\perp)U_+(v^\parallel + v^\perp)^\top = v^\parallel U_+(v^\parallel)^\top.$$

Because $\{v \in \mathcal{V}_+ : \|v\|_2 = 1\}$ is a compact set, $vU_+v^\top$ is bounded when $\|v(VI_+V^{-1})\|_2 = 1$. After getting (5.127), we have that

$$\begin{aligned} \|v\Phi\|_{U_+} &= \sqrt{v\Phi U_+ \Phi^\top v^\top} \\ &\geq \sqrt{(1 + \epsilon'_1)vU_+v^\top} \\ &= \sqrt{1 + \epsilon'_1} \|v\|_{U_+}. \end{aligned}$$

Letting $\epsilon_1 = \sqrt{1 + \epsilon'_1} - 1$, we obtain (5.123).

Next, we prove (5.124). For any $v \in \mathbb{R}^{|\mathcal{S}|}$, we have $vU_+v^\top \geq \|vVJ_+^\dagger V^{-1}\|_2^2$. We claim that there exists a positive constant ϵ'_2 such that

$$\|vVI_+V^{-1}\|_2 \leq \epsilon'_2 \|vVJ_+^\dagger V^{-1}\|_2. \quad (5.128)$$

To show (5.128), without loss of generality, it suffices to show that it suffices to show that the function $v \mapsto \|vVI_+V^{-1}\|_2$ is bounded on the set $\{v : \|vVJ_+^\dagger V^{-1}\|_2 = 1\}$. Consider the decomposition $\mathbb{R}^{|\mathcal{S}|} = \mathcal{V}_+ \oplus \mathcal{V}_-$. Because the kernel of $VJ_+^\dagger V^{-1}$ is $\mathcal{V}_-$, and $VJ_+^\dagger V^{-1}$ is invertible when restricted to $\mathcal{V}_+$, the set $\{v : \|vVJ_+^\dagger V^{-1}\|_2 = 1\}$ is the sum of $\mathcal{V}_-$ and a compact subset of $\mathcal{V}_+$. Because $\mathcal{V}_-$ is also the kernel of vVI_+V^{-1} , $\|vVI_+V^{-1}\|_2$ is bounded. This proves (5.128). Because $\Phi U_+ \Phi^\top = (VI_+V^{-1})(VI_+V^{-1})^\top + U_+$, (5.128) implies that

$$\begin{aligned} \|v\Phi\|_{U_+} &\leq \sqrt{v\Phi U_+ \Phi v^\top} \\ &= \sqrt{v(VI_+V^{-1})(VI_+V^{-1})^\top v^\top + vU_+v^\top} \\ &\leq \sqrt{(\epsilon'_2)^2 + 1} \|v\|_{U_+}. \end{aligned}$$

Taking $\epsilon_2 = \sqrt{(\epsilon'_2)^2 + 1} - 1$ finishes the proof.

Next, we show (5.125). Observe that

$$\Phi U_- \Phi^\top = VJI_-V^{-1}(VJI_-V^{-1})^\top + \sum_{k=1}^{\infty} (VJJ_-^k V^{-1})(VJJ_-^k V^{-1})^\top$$

$$\begin{aligned}
&= \sum_{k=1}^{\infty} (V J_-^k V^{-1}) (V J_-^k V^{-1})^\top \\
&= U_- - V I_- V^{-1} (V I_- V^{-1})^\top.
\end{aligned}$$

We claim that there exists a positive constant ϵ'_3 such that for any $v \in \mathbb{R}^{|\mathcal{S}|}$,

$$v(V I_- V^{-1})(V I_- V^{-1})^\top v^\top \geq \epsilon'_3 v U_- v^\top. \quad (5.129)$$

The argument for (5.129) is similar to the argument for (5.127), so we omit it here. Therefore,

$$\begin{aligned}
\|v\Phi\|_{U_-} &= \sqrt{v\Phi U_- \Phi^\top v^\top} \\
&\leq \sqrt{(1 - \epsilon'_3) v U_- v^\top} \\
&= \sqrt{1 - \epsilon'_3} \|v\|_{U_-}.
\end{aligned}$$

Taking $\epsilon_3 = 1 - \sqrt{1 - \epsilon'_3}$, we obtain (5.125).

To show (5.126), note that

$$\begin{aligned}
\|v\|_{U_+} + \|v\|_{U_-} &= \|v U_+ v^\top\| + \|v U_- v^\top\| \\
&\geq \|v(U_+ + U_-)v^\top\| \\
&= \|v\|_{U_+ + U_-}.
\end{aligned}$$

Since all norms in a finite-dimensional Euclidean space are equivalent, it suffices to show $\|\cdot\|_{U_+ + U_-}$ is a norm, or equivalently, $U_+ + U_-$ is positive definite. For any $v \in \mathbb{R}^{|\mathcal{S}|}$, we have

$$\begin{aligned}
v(U_+ + U_-)v^\top &\geq v \left((V J_+^\dagger V^{-1})(V J_+^\dagger V^{-1})^\top + V I_- V^{-1} (V I_- V^{-1})^\top \right) v^\top \\
&= v V (J_+^\dagger V^{-1} V^{-\top} J_+^\dagger + I_- V^{-1} V^{-\top} I_-) V^\top v^\top.
\end{aligned}$$

By the definitions of J_+ and I_- , it is not hard to see that we always have $v(U_+ + U_-)v^\top \geq 0$, and v is the zero vector whenever $v(U_+ + U_-)v^\top = 0$. Therefore, $U_+ + U_-$ is positive definite, which implies (5.126). $\square$

Now we are ready to conduct the Lyapunov analysis. We define the Lyapunov functions h_- and h_+ : for any system state x and $D \subseteq [N]$, let

$$h_-(x, D) = \|x(D) - m(D)\mu^*\|_{U_-} \quad (5.130)$$

$$h_+(x, D) = \|x(D) - m(D)\mu^*\|_{U_+}. \quad (5.131)$$

Our next lemma gives a lower bound for the drift of $h_-(x, D)$ under *any* policy.

Lemma 5.35. *Consider an instance of regular unstable RBs (Definition 5.14). Let K_1 be any positive constant. Under any policy π , for any time step t , we have*

$$\begin{aligned} & \mathbb{E}[h_+(X_{t+1}, [N]) \mid X_t, \mathbf{A}_t] - h_+(X_t, [N]) \\ & \geq \frac{\epsilon_1 K_1 - (2 + \epsilon_1 + \epsilon_2) K_1 \mathbb{1}\{\mathcal{E}_t^c\}}{\sqrt{N}} - \frac{6\lambda_+^{1/2}}{\beta N} - \frac{8\lambda_+^{1/2} d_{\text{IC}}(Y_t)}{\beta} \quad a.s., \end{aligned} \quad (5.132)$$

where $\beta = \min(\alpha, 1 - \alpha)$; the event $\mathcal{E}_t$ is given by

$$\begin{aligned} \mathcal{E}_t = & \left\{ h_+(X_t, [N]) > \frac{K_1}{\sqrt{N}} + \frac{2\lambda_+^{1/2} d_{\text{IC}}(Y_t)}{\beta} + \frac{\lambda_+^{1/2}}{\beta N} \right. \\ & \left. \text{or } \|X_t([N]) - \mu^*\|_1 \leq \eta' - \frac{6d_{\text{IC}}(Y_t)}{\beta} - \frac{3}{\beta N} \right\}, \end{aligned}$$

and $K_1 = \lambda_+^{1/2} K_{\text{noise}} / (2 + \epsilon_1 + \epsilon_2)$.

Proof. Let D_t be the set given in Lemma 5.33. By the definition of h_+ and the fact that $m(D_t) \geq 1 - 2d_{\text{IC}}(Y_t) - 1/N$, we have

$$\begin{aligned} & \mathbb{E}[h_+(X_{t+1}, [N]) \mid X_t, \mathbf{A}_t] - h_+(X_t, [N]) \\ & \geq \mathbb{E}[h_+(X_{t+1}, D_t) \mid X_t, \mathbf{A}_t] - h_+(X_t, D_t) - 4\lambda_+^{1/2}(1 - m(D_t)) \quad (5.133) \\ & \geq \mathbb{E}[h_+(X_{t+1}, D_t) \mid X_t, \mathbf{A}_t] - h_+(X_t, D_t) - 8\lambda_+^{1/2} d_{\text{IC}}(Y_t)/\beta - 4\lambda_+^{1/2}/(\beta N). \end{aligned}$$

Note that (5.133) uses the definition of h_+ and the fact that $\|v\|_{U_+} \leq \lambda_+^{1/2} \|v\|_1$.

We first consider the case with $h_+(X_t, D_t) > K_1/\sqrt{N}$. We have

$$\begin{aligned} \mathbb{E}[h_+(X_{t+1}, D_t) \mid X_t, \mathbf{A}_t] & = \mathbb{E}[\|X_{t+1}(D_t) - m(D_t)\mu^*\|_{U_+} \mid X_t, \mathbf{A}_t] \\ & \geq \|\mathbb{E}[X_{t+1}(D_t) - m(D_t)\mu^* \mid X_t, \mathbf{A}_t]\|_{U_+} \quad (5.134) \end{aligned}$$

$$\geq \|(X_t(D_t) - m(D_t)\mu^*)\Phi\|_{U_+} - 2\lambda_+^{1/2}/N, \quad (5.135)$$

where (5.134) uses Jensen's inequality; (5.135) invokes Lemma 5.33 and the fact that $\|v\|_{U_+} \leq \lambda_+^{1/2} \|v\|_1$ for any $v \in \mathbb{R}^{|\mathcal{S}|}$. By Lemma 5.34, we have

$$\|v\Phi\|_{U_+} \geq (1 + \epsilon_1)\|v\|_{U_+}.$$

Letting $v = X_t(D_t) - m(D_t)\mu^*$, combined with (5.135), we get

$$\begin{aligned} \mathbb{E}[h_+(X_{t+1}, D_t) \mid X_t, \mathbf{A}_t] &\geq (1 + \epsilon_1)h_+(X_t, D_t) - \frac{2\lambda_+^{1/2}}{N} \\ &\geq h_+(X_t, D_t) + \frac{\epsilon_1 K_1}{\sqrt{N}} - \frac{2\lambda_+^{1/2}}{N}, \end{aligned}$$

which implies that

$$\mathbb{E}[h_+(X_{t+1}, [N]) \mid X_t, \mathbf{A}_t] - h_+(X_t, [N]) \geq \frac{\epsilon_1 K_1}{\sqrt{N}} - \frac{6\lambda_+^{1/2}}{\beta N} - \frac{8\lambda_+^{1/2} d_{\text{IC}}(Y_t)}{\beta}. \quad (5.136)$$

Next, we consider the case of $h_+(X_t, D_t) \leq K_1/\sqrt{N}$:

$$\begin{aligned} &\mathbb{E}[h_+(X_{t+1}, D_t) \mid X_t, \mathbf{A}_t] \\ &= \mathbb{E}[\|X_{t+1}(D_t) - m(D_t)\mu^*\|_{U_+} \mid X_t, \mathbf{A}_t] \\ &\geq \mathbb{E}[\|X_{t+1}(D_t) - m(D_t)\mu^* - (X_t(D_t) - m(D_t)\mu^*)\Phi\|_{U_+} \mid X_t, \mathbf{A}_t] \\ &\quad - \mathbb{E}[\|X_t(D_t) - m(D_t)\mu^*\|_{U_+} \mid X_t, \mathbf{A}_t] \\ &\geq \mathbb{E}[\|X_{t+1}(D_t) - m(D_t)\mu^* - (X_t(D_t) - m(D_t)\mu^*)\Phi\|_{U_+} \mid X_t, \mathbf{A}_t] \\ &\quad - (1 + \epsilon_2)h_+(X_t, D_t), \end{aligned} \quad (5.137)$$

where (5.137) uses the fact that $\|v\Phi\|_{U_+} \leq (1 + \epsilon_2)\|v\|_{U_+}$ for all $v \in \mathbb{R}^{|\mathcal{S}|}$, proved in Lemma 5.34. We plan to invoke Property (c) of regular unstable RBs to bound the first term in (5.137). Let $X'_{t+1}(D_t)$ be the state distribution at time $t + 1$ when the arms in D_t follow LP-Priority. By Lemma 5.33, there exists a coupling such that $\|X'_{t+1}(D_t) - X_{t+1}(D_t)\|_1 \leq 2/N$ almost surely, so

$$\begin{aligned} &\mathbb{E}[\|X_{t+1}(D_t) - m(D_t)\mu^* - (X_t(D_t) - m(D_t)\mu^*)\Phi\|_{U_+} \mid X_t, \mathbf{A}_t] \\ &\geq \mathbb{E}[\|X'_{t+1}(D_t) - m(D_t)\mu^* - (X_t(D_t) - m(D_t)\mu^*)\Phi\|_{U_+} \mid X_t, \mathbf{A}_t] - \frac{2\lambda_+^{1/2}}{N}. \end{aligned} \quad (5.138)$$

Let $\xi^\top$ be the right eigenvector of U_+ corresponding to the eigenvalue λ_+ , with norm $\|\xi\|_2 = 1$. Because U_+ is positive semidefinite, $U_+ \geq \lambda_+ \xi^\top \xi$, so

$$\|v\|_{U_+} \geq \sqrt{\lambda_+ v \xi^\top \xi v^\top} = \lambda_+^{1/2} |v \xi^\top|,$$

for any $v \in \mathbb{R}^{|\mathcal{S}|}$. Moreover, we argue that ξ satisfies the conditions in Property (c) of regular unstable RBs: the column space of U_+ is a subspace of the column space of Φ . Because $\mu^* \Phi = 0$, we have $\mu^* U_+ = 0$, thus $\mu^* \xi^\top = \lambda_+^{-1} \mu^* U_+ \xi^\top = 0$. Therefore, by Property (c) of regular unstable RBs,

$$\begin{aligned} & \mathbb{E}[\|X'_{t+1}(D_t) - m(D_t)\mu^* - (X_t(D_t) - m(D_t)\mu^*)\Phi\|_{U_+} \mid X_t, \mathbf{A}_t] \\ & \geq \lambda_+^{1/2} \mathbb{E}[\|(X'_{t+1}(D_t) - m(D_t)\mu^* - (X_t(D_t) - m(D_t)\mu^*)\Phi)\xi^\top\| \mid X_t, \mathbf{A}_t] \\ & \geq \frac{\lambda_+^{1/2} K_{\text{noise}}}{\sqrt{N}} \mathbb{1}\{\|X_t(D_t) - m(D_t)\mu^*\|_1 \leq m(D_t)\eta'\}. \end{aligned} \quad (5.139)$$

Combining (5.137), (5.138) and (5.139), we get

$$\begin{aligned} & \mathbb{E}[h_+(X_{t+1}, D_t) \mid X_t, \mathbf{A}_t] - h_+(X_t, D_t) \\ & \geq \frac{\lambda_+^{1/2} K_{\text{noise}}}{\sqrt{N}} \mathbb{1}\{\|X_t(D_t) - m(D_t)\mu^*\|_1 \leq m(D_t)\eta'\} \\ & \quad - (2 + \epsilon_2)h_+(X_t, D_t) - \frac{2\lambda_+^{1/2}}{N} \\ & \geq \frac{\lambda_+^{1/2} K_{\text{noise}}}{\sqrt{N}} \mathbb{1}\{\|X_t(D_t) - m(D_t)\mu^*\|_1 \leq m(D_t)\eta'\} \\ & \quad - \frac{(2 + \epsilon_2)K_1}{\sqrt{N}} - \frac{2\lambda_+^{1/2}}{N}. \end{aligned}$$

Therefore, when $h_+(X_t, D_t) \leq K_1/\sqrt{N}$,

$$\begin{aligned} & \mathbb{E}[h_+(X_{t+1}, [N]) \mid X_t, \mathbf{A}_t] - h_+(X_t, [N]) \\ & \geq \frac{\lambda_+^{1/2} K_{\text{noise}}}{\sqrt{N}} \mathbb{1}\{\|X_t(D_t) - m(D_t)\mu^*\|_1 \leq m(D_t)\eta'\} \\ & \quad - \frac{(2 + \epsilon_2)K_1}{\sqrt{N}} - \frac{6\lambda_+^{1/2}}{\beta N} - \frac{8\lambda_+^{1/2} d_{\text{IC}}(Y_t)}{\beta}. \end{aligned} \quad (5.140)$$

Summarizing the bounds (5.136) and (5.140) and substituting $K_1 = \lambda_+^{1/2} K_{\text{noise}}/(2 + \epsilon_1 + \epsilon_2)$ in the bounds, we get

- When $h_+(X_t, D_t) > K_1/\sqrt{N}$ or $\|X_t(D_t) - m(D_t)\mu^*\|_1 \leq m(D_t)\eta'$, we have

$$\mathbb{E}[h_+(X_{t+1}, [N]) \mid X_t, \mathbf{A}_t] - h_+(X_t, [N]) \geq \frac{\epsilon_1 K_1}{\sqrt{N}} - \frac{6\lambda_+^{1/2}}{\beta N} - \frac{8\lambda_+^{1/2} d_{\text{IC}}(Y_t)}{\beta}. \quad (5.141)$$

- When $h_+(X_t, D_t) \leq K_1/\sqrt{N}$ and $\|X_t(D_t) - m(D_t)\mu^*\|_1 > m(D_t)\eta'$, we have

$$\begin{aligned} & \mathbb{E}[h_+(X_{t+1}, [N]) \mid X_t, \mathbf{A}_t] - h_+(X_t, [N]) \\ & \geq -\frac{(2 + \epsilon_2)K_1}{\sqrt{N}} - \frac{6\lambda_+^{1/2}}{\beta N} - \frac{8\lambda_+^{1/2} d_{\text{IC}}(Y_t)}{\beta}. \end{aligned} \quad (5.142)$$

Note that the conditions separating the two cases above are in terms of $X_t(D_t)$. To complete the proof, we need to state the conditions in terms of $X_t([N])$ instead, observe that

$$\begin{aligned} h_+(X_t, D_t) & \geq h_+(X_t, [N]) - \lambda_+^{1/2}(1 - m(D_t)) \\ & \geq h_+(X_t, [N]) - 2\lambda_+^{1/2} d_{\text{IC}}(Y_t)/\beta - \lambda_+^{1/2}/(\beta N) \\ \|X_t(D_t) - m(D_t)\mu^*\|_1 & \leq \|X_t([N]) - \mu^*\|_1 + 2(1 - m(D_t)) \\ & \leq \|X_t([N]) - \mu^*\|_1 + 4d_{\text{IC}}(Y_t)/\beta + 2/(\beta N). \end{aligned}$$

Therefore, $h_+(X_t, [N]) > K_1/\sqrt{N} + 2\lambda_+^{1/2} d_{\text{IC}}(Y_t)/\beta + \lambda_+^{1/2}/(\beta N)$ implies $h_+(X_t, D_t) > K_1/\sqrt{N}$; $\|X_t([N]) - \mu^*\|_1 \leq \eta' - 6d_{\text{IC}}(Y_t)/\beta - 3/(\beta N)$ implies that

$$\|X_t(D_t) - m(D_t)\mu^*\|_1 \leq \eta' - \frac{2d_{\text{IC}}(Y_t)}{\beta} - \frac{1}{\beta N} \leq \eta' - (1 - m(D_t)) \leq m(D_t)\eta',$$

which implies the bounds in the lemma statement. $\square$

In the next lemma, we upper bound the drift of $h_-(x, D)$ under *any policy*.

Lemma 5.36. *Consider an instance of regular unstable RBs (Definition 5.14). Under any policy π , for any time step t , we have*

$$\begin{aligned} \mathbb{E}[h_-(X_{t+1}, [N]) \mid X_t, \mathbf{A}_t] & \leq (1 - \epsilon_3)h_-(X_t, [N]) + \frac{\lambda_-^{1/2}|\mathbb{S}|}{\sqrt{N}} \\ & \quad + \frac{10\lambda_-^{1/2}|\mathbb{S}|}{\beta N} + \frac{8\lambda_-^{1/2} d_{\text{IC}}(Y_t)}{\beta} \quad a.s., \end{aligned} \quad (5.143)$$

where λ_- is the largest eigenvalue of U_- ; $\beta = \min(\alpha, 1 - \alpha)$.

Proof. Let D_t be the set given in Lemma 5.33. Let $X'_{t+1}(D_t)$ be the state distribution at time $t + 1$ when the actions for the arms in D_t are resampled using LP-Priority. By Lemma 5.33, there exists a coupling such that $\|X'_{t+1}(D_t) - X_{t+1}(D_t)\|_1 \leq 2/N$ almost surely. By the definition of h_- and the fact that $m(D_t) \geq 1 - 2d_{\text{IC}}(Y_t)/\beta - 1/(\beta N)$, we have

$$\begin{aligned}
& \mathbb{E}[h_-(X_{t+1}, [N]) \mid X_t, \mathbf{A}_t] - (1 - \epsilon_3)h_-(X_t, [N]) \\
& \leq \mathbb{E}[h_-(X_{t+1}, D_t) \mid X_t, \mathbf{A}_t] - (1 - \epsilon_3)h_-(X_t, D_t) + 4\lambda_-^{1/2}(1 - m(D_t)) \\
& \leq \mathbb{E}[h_-(X_{t+1}, D_t) \mid X_t, \mathbf{A}_t] - (1 - \epsilon_3)h_-(X_t, D_t) \\
& \quad + 8\lambda_-^{1/2}d_{\text{IC}}(Y_t)/\beta + 4\lambda_-^{1/2}/(\beta N) \\
& \leq \mathbb{E}[h_-(X'_{t+1}, D_t) \mid X_t, \mathbf{A}_t] - (1 - \epsilon_3)h_-(X_t, D_t) \\
& \quad + 8\lambda_-^{1/2}d_{\text{IC}}(Y_t)/\beta + 6\lambda_-^{1/2}/(\beta N).
\end{aligned} \tag{5.144}$$

Because X'_{t+1} is obtained by letting arms in D_t follows the LP-Priority, using similar arguments as the proof of Lemma 5.29, we get

$$\mathbb{E}[h_-(X'_{t+1}, D_t) \mid X_t, \mathbf{A}_t] - (1 - \epsilon_3)h_-(X_t, D_t) \leq \frac{\lambda_-^{1/2}|\mathbb{S}|}{\sqrt{N}} + \frac{4\lambda_-^{1/2}|\mathbb{S}|}{N}. \tag{5.145}$$

Plugging (5.145) into (5.144), and using the facts that $\beta \leq 1$ and $|\mathbb{S}| \geq 1$, we finish the proof. $\square$

Now we are ready to prove Theorem 5.15.

Proof of Theorem 5.15. For any system state x and subset $D \subseteq [N]$, define

$$h_{\text{lb}}(x, D) = L_{\text{lb}}h_+(x, D) - h_-(x, D),$$

where $L_{\text{lb}} = 2\lambda_-^{1/2}|\mathbb{S}|/\epsilon_1$. We will show that with an appropriate L_{lb} and for N larger than a constant threshold, we have

$$\mathbb{E}[h_{\text{lb}}(X_{t+1}, [N])] - \mathbb{E}[h_{\text{lb}}(X_t, [N])] \geq \Omega\left(\frac{1}{\sqrt{N}}\right) - \Theta(1) \cdot \mathbb{E}[d_{\text{IC}}(Y_t)] \quad \forall t \geq 0. \tag{5.146}$$

If we can prove (5.146), then taking the long-run average on two sides of the inequality implies

$$0 = \mathbb{E}[h_{\text{lb}}(X_\infty, [N])] - \mathbb{E}[h_{\text{lb}}(X_\infty, [N])] \geq \Omega\left(\frac{1}{\sqrt{N}}\right) - \Theta(1) \cdot \mathbb{E}[d_{\text{IC}}(Y_\infty)],$$

so $\mathbb{E}[d_{\text{IC}}(Y_\infty)] = \Omega(1/\sqrt{N})$. Combined with Lemma 5.32, we get $R^{\text{rel}} - R(\pi, \mathbf{S}_0) = \Omega(1/\sqrt{N})$, which proves Theorem 5.15.

Now we prove (5.146) using the formulas of $h_+(X_t, [N])$ and $h_-(X_t, [N])$'s drifts given in Lemma 5.35 and Lemma 5.36.

$$\begin{aligned}
& \mathbb{E}[h_{\text{lb}}(X_{t+1}, [N]) \mid X_t, \mathbf{A}_t] - h_{\text{lb}}(X_t, [N]) \\
&= L_{\text{lb}} \left(\mathbb{E}[h_+(X_{t+1}, [N]) \mid X_t, \mathbf{A}_t] - h_+(X_t, [N]) \right) \\
&\quad - \left(\mathbb{E}[h_-(X_{t+1}, [N]) \mid X_t, \mathbf{A}_t] - h_-(X_t, [N]) \right) \\
&\geq L_{\text{lb}} \left(\frac{\epsilon_1 K_1 - (2 + \epsilon_1 + \epsilon_2) K_1 \mathbb{1}\{\mathcal{E}_t^c\}}{\sqrt{N}} - \frac{6\lambda_+^{1/2}}{\beta N} \right. \\
&\quad \left. - 8\lambda_+^{1/2} \beta^{-1} d_{\text{IC}}(Y_t) \right) \\
&\quad + \epsilon_3 h_-(X_t, [N]) - \frac{\lambda_-^{1/2} |\mathbb{S}|}{\sqrt{N}} \\
&\quad - \frac{10\lambda_-^{1/2} |\mathbb{S}|}{\beta N} - 8\lambda_-^{1/2} \beta^{-1} d_{\text{IC}}(Y_t) \\
&= \left(L_{\text{lb}} \epsilon_1 K_1 - \lambda_-^{1/2} |\mathbb{S}| \right) \frac{1}{\sqrt{N}} \tag{5.147}
\end{aligned}$$

$$\begin{aligned}
& + \left(\epsilon_3 h_-(X_t, [N]) - \frac{L_{\text{lb}}(2 + \epsilon_1 + \epsilon_2) K_1}{\sqrt{N}} \mathbb{1}\{\mathcal{E}_t^c\} \right) \\
& - (8L_{\text{lb}} \lambda_+^{1/2} + 8\lambda_-^{1/2}) \beta^{-1} d_{\text{IC}}(Y_t) - O\left(\frac{1}{N}\right). \tag{5.148}
\end{aligned}$$

The coefficient of the $1/\sqrt{N}$ in (5.147) equals $\lambda_-^{1/2} |\mathbb{S}|$, which is positive.

Comparing (5.147)-(5.148) with our goal, (5.146), it remains to lower bound the second term in (5.147). First, observe that given the event $\mathcal{E}_t$, the second term of (5.147) is non-negative. So we only need to consider the event $\mathcal{E}_t^c$. By definition, $\mathcal{E}_t^c$ implies that

$$\begin{aligned}
h_+(X_t, [N]) &\leq \frac{K_1}{\sqrt{N}} + \frac{2\lambda_+^{1/2} d_{\text{IC}}(Y_t)}{\beta} + \frac{\lambda_+^{1/2}}{\beta N}, \\
\|X_t([N]) - \mu^*\|_1 &> \eta' - \frac{6d_{\text{IC}}(Y_t)}{\beta} - \frac{3}{\beta N}.
\end{aligned}$$

By (5.126) in Lemma 5.34,

$$h_+(X_t, [N]) + h_-(X_t, [N]) \geq K_{\text{norm}} \|X_t([N]) - \mu^*\|_1,$$

so given $\mathcal{E}_t^c$,

$$\begin{aligned} h_-(X_t, [N]) &\geq K_{\text{norm}} \|X_t([N]) - \mu^*\|_1 - h_+(X_t, [N]) \\ &\geq K_{\text{norm}} \left(\eta' - \frac{6d_{\text{IC}}(Y_t)}{\beta} - \frac{3}{\beta N} \right) - \frac{K_1}{\sqrt{N}} - \frac{2\lambda_+^{1/2}d_{\text{IC}}(Y_t)}{\beta} - \frac{\lambda_+^{1/2}}{\beta N} \\ &= K_{\text{norm}}\eta' - \frac{K_1}{\sqrt{N}} - (6K_{\text{norm}} + 2\lambda_+^{1/2})\beta^{-1}d_{\text{IC}}(Y_t) - O\left(\frac{1}{N}\right). \end{aligned}$$

Therefore, given $\mathcal{E}_t^c$, combining the above lower bound, if $N \geq K_1^2(\epsilon_3 + L_{\text{lb}}(2 + \epsilon_1 + \epsilon_2))^2 / (K_{\text{norm}}\eta'\epsilon_3)^2$,

$$\begin{aligned} &\epsilon_3 h_-(X_t, [N]) - \frac{L_{\text{lb}}(2 + \epsilon_1 + \epsilon_2)K_1}{\sqrt{N}} \mathbb{1}_{\{\mathcal{E}_t^c\}} \\ &\geq \epsilon_3 K_{\text{norm}}\eta' - \frac{K_1(\epsilon_3 + L_{\text{lb}}(2 + \epsilon_1 + \epsilon_2))}{\sqrt{N}} \\ &\quad - \epsilon_3(6K_{\text{norm}} + 2\lambda_+^{1/2})\beta^{-1}d_{\text{IC}}(Y_t) - O\left(\frac{1}{N}\right), \\ &\geq -\epsilon_3(6K_{\text{norm}} + 2\lambda_+^{1/2})\beta^{-1}d_{\text{IC}}(Y_t) - O\left(\frac{1}{N}\right). \end{aligned} \tag{5.149}$$

Substituting (5.149) back to the second term of (5.147), we get

$$\begin{aligned} &\mathbb{E} [h_{\text{lb}}(X_{t+1}, [N]) \mid X_t, \mathbf{A}_t] - h_{\text{lb}}(X_t, [N]) \\ &\geq \frac{\lambda_-^{1/2}|\mathbb{S}|}{\sqrt{N}} - (6\epsilon_3 K_{\text{norm}} + 2\epsilon_3 \lambda_+^{1/2} + 8L_{\text{lb}}\lambda_+^{1/2} + 8\lambda_-^{1/2})\beta^{-1}d_{\text{IC}}(Y_t) - O\left(\frac{1}{N}\right), \end{aligned}$$

which implies (5.146) and finishes the proof. $\square$

5.D.3 Remarks on Item (c) of regular unstable RBs (Definition 5.14)

In this section, we prove an intuitive sufficient condition for Property (c) of regular unstable RBs and also provide a weaker version of this property.

We first state and prove the intuitive sufficient condition for Property (c): **Lemma 5.37.** *Consider an RB instance with the single-armed MDP $(\mathbb{S}, \mathbb{A}, P, r)$, where $P(s, a, s') > 0$ for all $s \in \mathbb{S}, a \in \mathbb{A}, s' \in \mathbb{S}$. Then this RB instance satisfies Property (c) in Definition 5.14.*

Proof. Throughout this proof, we fix X_t and A_t and omit them in the conditioning when writing expectations or probabilities.

Recalling the statement of Property (c), it suffices to show that for any $\xi \in \mathbb{R}^{|\mathbb{S}|}$ with $\|\xi\|_2 = 1$ and $\mu^* \xi^\top = 0$, there exists a constant $K_{\text{noise}} > 0$ such that for any N larger than a constant and any time step t , if the local optimal policy is applicable, we have

$$\mathbb{E}[\|(X_{t+1}([N]) - \mu^* - (X_t([N]) - \mu^*)\Phi)\xi^\top\| \mid X_t, \mathbf{A}_t] \geq \frac{K_{\text{noise}}}{\sqrt{N}}, \quad (5.150)$$

where $\mathbf{A}_t$ follows the LP-Priority. Note that in this case, the parameter η' in Property (c) of Definition 5.14 could be any positive number such that $\|X_t([N]) - \mu^*\|_1 \leq \eta'$ guarantees the condition (5.8) for applying the LP-Priority.

We first show that for any $\xi \in \mathbb{R}^{|\mathbb{S}|}$ such that $\|\xi\|_2 = 1$ and $\mu^* \xi^\top = 0$,

$$\xi_{\max} - \xi_{\min} \geq 1/\sqrt{|\mathbb{S}|}. \quad (5.151)$$

where $\xi_{\max} = \max_{s \in \mathbb{S}} \xi(s)$ and $\xi_{\min} = \min_{s \in \mathbb{S}} \xi(s)$. Observe that we must have either $\xi_{\max} \geq 1/\sqrt{|\mathbb{S}|}$ or $\xi_{\min} \leq -1/\sqrt{|\mathbb{S}|}$, because otherwise $\|\xi\|_2 < 1$. We first show (5.151) for the case with $\xi_{\max} \geq 1/\sqrt{|\mathbb{S}|}$. In this case, we must have $\xi_{\min} \leq 1/\sqrt{|\mathbb{S}|}$, because otherwise $\|\xi\|_2 > 1$. Let $\mathbf{1}_{\text{scaled}} \in \mathbb{R}^{|\mathbb{S}|}$ be the whose entries are all $1/\sqrt{|\mathbb{S}|}$. Then

$$\|\xi - \mathbf{1}_{\text{scaled}}\|_\infty \leq \max(\xi_{\max} - 1/\sqrt{|\mathbb{S}|}, 1/\sqrt{|\mathbb{S}|} - \xi_{\min}) \leq \xi_{\max} - \xi_{\min}.$$

On the other hand, because $\|\mu^*\|_1 = 1$ and $\mu^* \mathbf{1}_{\text{scaled}}^\top = 1/\sqrt{|\mathbb{S}|}$,

$$\|\xi - \mathbf{1}_{\text{scaled}}\|_\infty \geq \mu^* \mathbf{1}_{\text{scaled}}^\top - \mu^* \xi^\top = 1/\sqrt{|\mathbb{S}|}.$$

Combining the last two inequalities yields (5.151). For the case with $\xi_{\min} \leq -1/\sqrt{|\mathbb{S}|}$, (5.151) can be proved in an analogous way.

To show (5.150), we invoke the Paley-Zygmund inequality to get

$$\begin{aligned} & \mathbb{P} \left[(X_{t+1}([N])\xi^\top - \mathbb{E}[X_{t+1}([N])\xi^\top])^2 > \frac{1}{2} \text{Var}[X_{t+1}([N])\xi^\top] \right] \\ & \geq \frac{1}{4} \frac{\text{Var}[X_{t+1}([N])\xi^\top]^2}{\mathbb{E}[(X_{t+1}([N])\xi^\top - \mathbb{E}[X_{t+1}([N])\xi^\top])^4]}. \end{aligned} \quad (5.152)$$

Next, we lower bound the second moment of $X_{t+1}([N])\xi^\top - \mathbb{E}[X_{t+1}([N])\xi^\top]$ in (5.152). For any timestep t , given $X_t, \mathbf{A}_t$, we have

$$X_{t+1}([N])\xi^\top = \sum_{s \in \mathbb{S}} X_{t+1}([N], s)\xi(s) \quad (5.153)$$

$$= \frac{1}{N} \sum_{s \in \mathbb{S}} \sum_{i \in [N]} \mathbf{1}\{\mathbf{S}_{t+1}(i) = s\} \xi(s) \quad (5.154)$$

$$= \frac{1}{N} \sum_{i \in [N]} \xi(\mathbf{S}_{t+1}(i)). \quad (5.155)$$

Note that conditioned on $X_t, \mathbf{A}_t$, $\xi(\mathbf{S}_{t+1}(i))$ are independent across i . Therefore, the variance of $X_{t+1}([N])\xi^\top$ is given by

$$\text{Var}[X_{t+1}([N])\xi^\top] = \frac{1}{N^2} \sum_{i \in [N]} \text{Var}[\xi(\mathbf{S}_{t+1}(i))].$$

Because $P(s, a, s') > 0$ for all $s, s' \in \mathbb{S}, a \in \mathbb{A}$, the support of $\mathbf{S}_{t+1}(i)$ is the whole state space $\mathbb{S}$. Moreover, because $\xi_{\max} - \xi_{\min} > 1/\sqrt{|\mathbb{S}|}$, we have

$$\begin{aligned} \text{Var}[\xi(\mathbf{S}_{t+1}(i))] &\geq \min_{s, a, s'} P(s, a, s') \left((\xi_{\max} - \mathbb{E}[\xi(\mathbf{S}_{t+1}(i))])^2 \right. \\ &\quad \left. + (\xi_{\min} - \mathbb{E}[\xi(\mathbf{S}_{t+1}(i))])^2 \right) \\ &\geq \frac{1}{2} \min_{s, a, s'} P(s, a, s') (\xi_{\max} - \xi_{\min})^2 \\ &\geq \frac{\min_{s, a, s'} P(s, a, s')}{2|\mathbb{S}|}, \end{aligned}$$

where $\min_{s, a, s'}$ is a shorthand for $\min_{s, s' \in \mathbb{S}, a \in \mathbb{A}}$. Therefore,

$$\text{Var}[X_{t+1}([N])\xi^\top] \geq \frac{1}{N^2} \sum_{i \in [N]} \text{Var}[\xi(\mathbf{S}_{t+1}(i))] \geq \frac{\min_{s, a, s'} P(s, a, s')}{2|\mathbb{S}|N}. \quad (5.156)$$

Next, we upper bound the fourth moment of $X_{t+1}([N])\xi^\top - \mathbb{E}[X_{t+1}([N])\xi^\top]$. Using the fact that $X_{t+1}([N])\xi^\top = \sum_{i \in [N]} \xi(\mathbf{S}_{t+1}(i))/N$, we get

$$\mathbb{E} \left[\left(X_{t+1}([N])\xi^\top - \mathbb{E}[X_{t+1}([N])\xi^\top] \right)^4 \right]$$

$$\begin{aligned}
&= \frac{1}{N^4} \left(\sum_{i \in [N]} \mathbb{E} \left[\left(\xi(\mathbf{S}_{t+1}(i)) - \mathbb{E}[\xi(\mathbf{S}_{t+1}(i))] \right)^4 \right] \right. \\
&\quad \left. + 6 \sum_{1 \leq i < j \leq N} \text{Var}[\xi(\mathbf{S}_{t+1}(i))] \text{Var}[\xi(\mathbf{S}_{t+1}(j))] \right) \\
&\leq \frac{1}{N^4} \left(N(\xi_{\max} - \xi_{\min})^4 + 6 \frac{N(N-1)}{2} (\xi_{\max} - \xi_{\min})^4 \right) \\
&\leq \frac{3(\xi_{\max} - \xi_{\min})^4}{N^2}. \tag{5.157}
\end{aligned}$$

Substituting the calculations in (5.156) and (5.157) back to (5.152), we get

$$\begin{aligned}
&\mathbb{P} \left[\left(X_{t+1}([N])\xi^\top - \mathbb{E}[X_{t+1}([N])\xi^\top] \right)^2 > \frac{\min_{s,a,s'} P(s, a, s')}{4|\mathbb{S}|N} \right] \\
&\geq \left(\frac{\min_{s,a,s'} P(s, a, s')}{2|\mathbb{S}|N} \right)^2 \frac{N^2}{12(\xi_{\max} - \xi_{\min})^4} \\
&= \frac{\min_{s,a,s'} P(s, a, s')^2}{47|\mathbb{S}|^2(\xi_{\max} - \xi_{\min})^4}. \tag{5.158}
\end{aligned}$$

Because under the LP-Priority, we have

$$\left\| \mathbb{E}[X_{t+1}([N]) \mid X_t, \mathbf{A}_t] - \mathbb{E}[X_{t+1}([N])\xi^\top \mid X_t] \right\| \leq 2/N$$

and $\mathbb{E}[X_{t+1}([N]) \mid X_t] = X_t([N])\Phi$, we can rewrite the bound (5.158) to get

$$\begin{aligned}
&\mathbb{P} \left[\left| \left(X_{t+1}([N])\xi^\top - X_t([N])\Phi \right) \xi^\top \right| \right. \\
&\quad \left. > \frac{\min_{s,a,s'} P(s, a, s')^{1/2}}{2|\mathbb{S}|^{1/2}\sqrt{N}} - \frac{2}{N} \right] \\
&\geq \frac{\min_{s,a,s'} P(s, a, s')^2}{47|\mathbb{S}|^2(\xi_{\max} - \xi_{\min})^4}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E} \left[\left| \left(X_{t+1}([N])\xi^\top - X_t([N])\Phi \right) \xi^\top \right| \right] &\geq \left(\frac{\min_{s,a,s'} P(s, a, s')^{1/2}}{2|\mathbb{S}|^{1/2}\sqrt{N}} - \frac{2}{N} \right) \\
&\quad \cdot \frac{\min_{s,a,s'} P(s, a, s')^2}{47|\mathbb{S}|^2(\xi_{\max} - \xi_{\min})^4},
\end{aligned}$$

which implies (5.150) with K_{noise} being the coefficient of $1/\sqrt{N}$ in the last expression. $\square$

Next, we state the weaker version of Property (c) of regular unstable RBs.

Property 5.38 (Weaker version of Property (c) in Definition 5.14). There exists positive constants η' and K_{noise} with $\eta' \leq \min\{y^*(\tilde{s}, 0), y^*(\tilde{s}, 1)\} - 1/N$, such that for any large-enough N and any time step t , we have

$$\mathbb{E}[\|X_{t+1}([N]) - \mu^* - (X_t([N]) - \mu^*)\Phi\|_{U_+} \mid X_t, \mathbf{A}_t] \geq \frac{\lambda_+^{1/2} K_{\text{noise}}}{\sqrt{N}}$$

$$\text{if } \|X_t([N]) - \mu^*\|_1 \leq \eta',$$

where $\mathbf{A}_t$ follow the LP-Priority (Subroutine 2), which is always feasible if $\|X_t([N]) - \mu^*\|_1 \leq \min\{y^*(\tilde{s}, 0), y^*(\tilde{s}, 1)\} - 1/N$.

Intuitively, Property 5.38 assumes that when $X_t([N])$ is close to μ^* , the state-count vector in the next step, $X_{t+1}([N])$, has enough randomness; or precisely, $X_{t+1}([N]) - \mu^* - \mathbb{E}[X_{t+1}([N]) - \mu^* \mid X_t, \mathbf{A}_t]$ has enough variances in the directions orthogonal to the kernel of U_+ .

If Property (c) of regular unstable RBs is replaced by Property 5.38, the whole proof of Theorem 5.15 goes through almost verbatim, except one change: In the proof of Lemma 5.35, in the argument below (5.138), we apply Property 5.38 rather than Property (c).

5.D.4 Proof of Corollary 5.16

Corollary 5.16. *There exists an RB satisfying Assumptions 5.1 and 5.2 and violating Assumption 5.3 such that $R^{\text{rel}} - R(\pi, \mathcal{S}_0) \geq \Omega\left(\frac{1}{\sqrt{N}}\right)$.*

Proof. We will verify that the example in Figure 1 of [149] is regular unstable.

In this example, the single-armed MDP $(\mathbb{S}, \mathbb{A}, P, r)$ is defined as below: $\mathbb{S} = \{1, 2, 3\}$; $\mathbb{A} = \{0, 1\}$; the transition kernel P is given by

$$P(\cdot, 0, \cdot) = \begin{bmatrix} 0.02232142 & 0.10229283 & 0.87538575 \\ 0.03426605 & 0.17175704 & 0.79397691 \\ 0.52324756 & 0.45523298 & 0.02151947 \end{bmatrix},$$

$$P(\cdot, 1, \cdot) = \begin{bmatrix} 0.14874601 & 0.30435809 & 0.54689589 \\ 0.56845754 & 0.41117331 & 0.02036915 \\ 0.25265570 & 0.27310439 & 0.4742399 \end{bmatrix};$$

The reward function r is defined as

$$r(\cdot, 0) = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix},$$

$$r(\cdot, 1) = [0.37401552 \quad 0.11740814 \quad 0.07866135].$$

Let $\alpha = 0.4$ in the budget constraint. The optimal solution of the LP relaxation (LP) is:

$$y^*(\cdot, \cdot) = \begin{bmatrix} 0 & 0.29943 \\ 0.23768 & 0.10057 \\ 0.36232 & 0 \end{bmatrix},$$

where the first column corresponds to $y^*(s, 0)$ and the second column corresponds to $y^*(s, 1)$.

Since $P(s, a, s') > 0$ for all s, a, s' , $P_{\bar{\pi}^*}$ is an aperiodic unichain, so Assumption 5.1 holds. Also, from y^* , we can see that the state 2 is the unique neutral state, implying Assumption 5.2. Moreover, one can numerically verify that y^* is the unique solution to (LP), so Property (a) of Definition 5.14 is satisfied; the moduli of the eigenvalues of Φ are $\{1.133, 0, 0.059\}$, so Property (b) of Definition 5.14 is satisfied.

Finally, observe that in this example, $P(s, a, s') > 0$ for any $s, s' \in \mathbb{S}$ and $a \in \mathbb{A}$. Therefore, we invoke the sufficient condition Lemma 5.37 proved in Appendix 5.D.3 to conclude that this example satisfies Property (c) of Definition 5.14. $\square$

5.E Experiment details

5.E.1 Details of Two-Set Policy

Determine η and ϵ_N^{fe}

In the analysis, $\eta = |\mathbb{S}|^{-1/2} \min\{y^*(\tilde{s}, 1), y^*(\tilde{s}, 0)\}$, $\epsilon_N^{\text{fe}} = (|S^\emptyset| + 1)/N$. In the simulations, we will try to choose a pair of feasibility-ensuring $(\eta, \epsilon_N^{\text{fe}})$ with η being as large as possible. Recall that $(\eta, \epsilon_N^{\text{fe}})$ is feasibility-ensuring if for any system state x and $D \subseteq [N]$ satisfying

$$\|x(D) - m(D)\mu^*\|_U \leq \eta m(D) - \epsilon_N^{\text{fe}}, \quad (5.159)$$

we have

$$\sum_{s \neq \tilde{s}} \bar{\pi}^*(1|s)x(D, s) \leq \alpha m(D) - \frac{|S^\emptyset| + 1}{N} \quad \text{and}, \quad (5.160)$$

$$\sum_{s \neq \tilde{s}} \bar{\pi}^*(0|s)x(D, s) \leq (1 - \alpha)m(D) - \frac{|S^\emptyset| + 1}{N}. \quad (5.161)$$

Equivalently, $(\eta, \epsilon_N^{\text{fe}})$ is feasibility-ensuring if whenever

$$\sum_{s \neq \tilde{s}} \bar{\pi}^*(1|s)x(D, s) \geq \alpha m(D) - \frac{|S^\emptyset| + 1}{N} \quad \text{or} \ , \quad (5.162)$$

$$\sum_{s \neq \tilde{s}} \bar{\pi}^*(0|s)x(D, s) \geq (1 - \alpha)m(D) - \frac{|S^\emptyset| + 1}{N}, \quad (5.163)$$

we have $\|x(D) - m(D)\mu^*\|_U \geq \eta m(D) - \epsilon_N^{\text{fe}}$. Thus, ignoring the $O(1/N)$ term, we consider solving the following two Second-Order Conic Programs (SOCPs), where the decision variable $\mathbf{z}$ represent $X_t(D)/m(D)$ for some non-empty D :

$$\underset{\mathbf{z} \in \mathbb{R}^{|S|}}{\text{minimize}} \quad \|\mathbf{z} - \mu^*\|_U \quad (5.164a)$$

$$\text{subject to} \quad \sum_{s \neq \tilde{s}} \bar{\pi}^*(1|s)\mathbf{z}(s) \geq \alpha \quad (5.164b)$$

$$\mathbf{z}\mathbf{1}^\top = 1 \quad (5.164c)$$

$$\mathbf{z} \geq 0 \quad (5.164d)$$

$$\underset{\mathbf{z} \in \mathbb{R}^{|S|}}{\text{minimize}} \quad \|\mathbf{z} - \mu^*\|_U \quad (5.165a)$$

$$\text{subject to} \quad \sum_{s \neq \tilde{s}} \bar{\pi}^*(0|s)\mathbf{z}(s) \geq 1 - \alpha \quad (5.165b)$$

$$\mathbf{z}\mathbf{1}^\top = 1 \quad (5.165c)$$

$$\mathbf{z} \geq 0. \quad (5.165d)$$

The following proposition shows that one can obtain a pair of feasibility-ensuring $(\eta, \epsilon_N^{\text{fe}})$ from (5.164a) and (5.165a) if the two SOCPs are both feasible.

Proposition 5.39. *Suppose the SOCPs in (5.164a)–(5.164d) and (5.165a)–(5.165d) are feasible; let their optimal values be η_1 and η_2 , respectively. Let $\eta = \min(\eta_1, \eta_2)$ and let $\epsilon_N^{\text{fe}} = 2\sqrt{2}\lambda_U^{1/2}(|S^\emptyset| + 1)/N$. Then for any system state x and $D \in [N]$ satisfying (5.159), we have (5.160) and (5.161).*

Proof. In this proof, we write $\epsilon_0 = (|S^\emptyset| + 1)/N$ for simplicity. To get a contradiction, suppose (5.160) does not hold, i.e.,

$$\sum_{s \neq \tilde{s}} \bar{\pi}^*(1|s)x(D, s) > \alpha m(D) - \epsilon_0.$$

Let $p_{\max} = \max_{s \neq \tilde{s}} \bar{\pi}^*(1|s)$. Because (5.164a)–(5.164d) is feasible, we must have $p_{\max} \geq \alpha$. We define the sets S^a and S^b to be

$$\begin{aligned} S^a &= \{s \in \mathbb{S} : s \neq \tilde{s}, \bar{\pi}^*(1|s) = p_{\max}\} \\ S^b &= \{s \in \mathbb{S} : s \neq \tilde{s}, \bar{\pi}^*(1|s) < p_{\max}\}. \end{aligned}$$

Now we construct a feasible $\mathbf{z}$ to the SOCP in (5.164a)–(5.164d) as

$$z(s) = \begin{cases} (x(D, s) + c)/m(D) & \text{if } s \in S^a \\ (x(D, s) - d)^+/m(D) & \text{if } s \in S^b \cup \tilde{s}, \end{cases} \quad (5.166)$$

for some $d > 0$ to be determined later, and c is such that $\sum_{s \in \mathbb{S}} z(s) = 1$. Note that such c always exists for any d . Observe that when d is sufficiently large, $\sum_{s \neq \tilde{s}} \bar{\pi}^*(1|s)z(s) = p_{\max} \geq \alpha$. By the intermediate value theorem, we can let d be such that $\sum_{s \neq \tilde{s}} \bar{\pi}^*(1|s)z(s) = \alpha$. Then $\mathbf{z}$ is feasible to the SOCP in (5.164a)–(5.164d). Therefore, by the definition of η , we must have

$$\|\mathbf{z} - \mu^*\|_U \geq \eta. \quad (5.167)$$

To get a contradiction, we only need to show that

$$\|x(D) - m(D)\mathbf{z}\|_U \leq \epsilon_N^{\text{fe}} = 2\sqrt{2}\lambda_U^{1/2}\epsilon_0, \quad (5.168)$$

which would imply that

$$\begin{aligned} \|x(D) - m(D)\mu^*\|_U &\geq \|m(D)\mathbf{z} - m(D)\mu^*\|_U - \|x(D) - m(D)\mathbf{z}\|_U \\ &\geq \eta m(D) - \epsilon_N^{\text{fe}}. \end{aligned}$$

To show (5.168), observe that

$$\begin{aligned} &\|x(D) - m(D)\mathbf{z}\|_2 \\ &= \sqrt{\sum_{s \in S^a} (x(D, s) - m(D)z(s))^2 + \sum_{s \in S^b \cup \{\tilde{s}\}} (x(D, s) - m(D)z(s))^2} \end{aligned}$$

$$\begin{aligned}
&= \sqrt{\left(\sum_{s \in S^a} (x(D, s) - m(D)z(s))\right)^2 + \left(\sum_{s \in S^b \cup \{\tilde{s}\}} (x(D, s) - m(D)z(s))\right)^2} \\
&= \sqrt{2} \sum_{s \in S^b \cup \{\tilde{s}\}} (x(D, s) - m(D)z(s)).
\end{aligned}$$

To bound $\sum_{s \in S^a} (x(D, s) - m(D)z(s))$, observe that

$$\begin{aligned}
\epsilon_0 &\geq \alpha m(D) - \sum_{s \neq \tilde{s}} \bar{\pi}^*(1|s)x(D, s) \\
&= \sum_{s \neq \tilde{s}} \bar{\pi}^*(1|s)m(D)z(s) - \sum_{s \neq \tilde{s}} \bar{\pi}^*(1|s)x(D, s) \\
&= \sum_{s \in S^a} \bar{\pi}^*(1|s)(m(D)z(s) - x(D, s)) + \sum_{s \in S^b} \bar{\pi}^*(1|s)(m(D)z(s) - x(D, s)) \\
&= -p_{\max} \sum_{s \in S^b \cup \{\tilde{s}\}} (m(D)z(s) - x(D, s)) + \sum_{s \in S^b} \bar{\pi}^*(1|s)(m(D)z(s) - x(D, s)) \\
&= \sum_{s \in S^b} (p_{\max} - \bar{\pi}^*(1|s))(x(D, s) - m(D)z(s)) + p_{\max}(x(D, \tilde{s}) - m(D)z(\tilde{s})) \\
&\geq \frac{1}{2} \sum_{s \in S^b \cup \{\tilde{s}\}} (x(D, s) - m(D)z(s)).
\end{aligned}$$

where the last inequality uses the fact that $\bar{\pi}^*(1|s) \in \{0, 1/2, 1\}$ if $s \neq \tilde{s}$, so $p_{\max} - \bar{\pi}^*(1|s) \geq 1/2$ for $s \in S^b$ and $p_{\max} \geq 1/2$. Therefore,

$$\|x(D) - m(D)z\|_2 = \sqrt{2} \sum_{s \in S^b \cup \{\tilde{s}\}} (x(D, s) - m(D)z(s)) \leq 2\sqrt{2}\epsilon_0.$$

Because $\|v\|_U \leq \lambda_U^{1/2} \|v\|_2$ for any $v \in \mathbb{R}^{|\mathbb{S}|}$, we have proved (5.168), which leads to a contradiction. Therefore, (5.160) must hold. Similarly, (5.161) also holds. $\square$

Next, we discuss choosing η and ϵ_N^{fe} when either of the SOCPs is not feasible. We first temporarily let $\eta = \min(\eta_1, \eta_2)$ and $\epsilon_N^{\text{fe}} = 2\sqrt{2}\lambda_U^{1/2}(|S^\emptyset| + 1)/N$, where η_1 or η_2 is $+\infty$ if the corresponding SOCP is infeasible. We modify the values of η and ϵ_N^{fe} if the following two cases happen, and do nothing otherwise.

- If (5.164a)–(5.164d) is not feasible, but $(\alpha - \max_{s \neq \tilde{s}} \bar{\pi}^*(1|s))\epsilon_N^{\text{fe}}/\eta \geq (|S^\emptyset| + 1)/N$, then we reset

$$(\eta, \epsilon_N^{\text{fe}}) = (|\mathbb{S}|^{-1/2} \min\{y^*(\tilde{s}, 1), y^*(\tilde{s}, 0)\}, |\mathbb{S}|^{-1/2}(|S^\emptyset| + 1)/N),$$

which is guaranteed to be feasibility ensuring by Lemma 5.10.

- If (5.165a)–(5.165d) is not feasible, but $(1 - \alpha - \max_{s \neq \tilde{s}} \bar{\pi}^*(0|s))\epsilon_N^{\text{fe}}/\eta \geq (|S^\emptyset| + 1)/N$, then we again reset

$$(\eta, \epsilon_N^{\text{fe}}) = (|\mathbb{S}|^{-1/2} \min\{y^*(\tilde{s}, 1), y^*(\tilde{s}, 0)\}, |\mathbb{S}|^{-1/2}(|S^\emptyset| + 1)/N).$$

To see why $(\eta, \epsilon_N^{\text{fe}})$ obtained in this way is feasibility-ensuring, we focus on the situation where neither of the above two cases happen. Without loss of generality, we assume $m(D) \geq \epsilon_N^{\text{fe}}/\eta$, because otherwise (5.159) is always false. Suppose (5.164a)–(5.164d) is infeasible, but $(\alpha - \max_{s \neq \tilde{s}} \bar{\pi}^*(1|s))\epsilon_N^{\text{fe}}/\eta \geq (|S^\emptyset| + 1)/N$. Then for $m(D) \geq \epsilon_N^{\text{fe}}/\eta$,

$$\begin{aligned} \alpha m(D) - \sum_{s \neq \tilde{s}} \bar{\pi}^*(1|s)x(D, s) &\geq (\alpha - \max_{s \neq \tilde{s}} \bar{\pi}^*(1|s))m(D) \\ &\geq (\alpha - \max_{s \neq \tilde{s}} \bar{\pi}^*(1|s)) \frac{\epsilon_N^{\text{fe}}}{\eta} \\ &> (|S^\emptyset| + 1)/N, \end{aligned}$$

where the second inequality utilizes the fact that $\alpha > \max_{s \neq \tilde{s}} \bar{\pi}^*(1|s)$ when (5.164a)–(5.164d) is infeasible. Therefore, (5.160) holds automatically. Then if (5.165a)–(5.165d) is also infeasible, we can apply the same argument above; if (5.165a)–(5.165d) is feasible, with $(\eta, \epsilon_N^{\text{fe}}) = (\eta_2, 2\sqrt{2}\lambda_U^{1/2}(|S^\emptyset| + 1)/N)$, by the arguments in Proposition 5.39, (5.159) implies (5.161). Therefore, $(\eta, \epsilon_N^{\text{fe}})$ is feasibility-ensuring.

Updating D_t^{LP}

With η solved, now we implement the step of solving D_t^{LP} . When $\delta(X_t, [N]) \geq 0$, we simply let $D_t^{\text{LP}} = [N]$. Next, we specify the procedure for determining D_t^{LP} in two cases, based on whether $\delta(X_t, D_{t-1}^{\text{LP}}) \geq 0$.

Case 1: When $\delta(X_t, D_{t-1}^{\text{LP}}) \geq 0$. In this case, we solve the following SOCP parameterized by $\epsilon_N^{\text{rd}} = (1 + 2\lambda_U^{1/2})|\mathbb{S}|/N$:

$$\underset{\mathbf{z}, m}{\text{maximize}} \quad \sum_{s \in \mathbb{S}} z(s) \tag{5.169a}$$

$$\text{subject to} \quad \|\mathbf{z} - m\mu^*\|_U \leq \eta m - \frac{|S^\emptyset| + 1}{N} - \epsilon_N^{\text{rd}} \tag{5.169b}$$

$$\sum_{s \in \mathbb{S}} z(s) = m \quad (5.169c)$$

$$X_t(D_{t-1}^{\text{LP}}, s) \leq z(s) \leq X_t([N], s) \quad \forall s \in \mathbb{S}. \quad (5.169d)$$

Based on whether the SOCP is feasible, we define D_t^{LP} and argue that it satisfies the requirements of Algorithm 10 in different ways, as specified below.

If the SOCP is infeasible, we choose $D_t^{\text{LP}} = D_{t-1}^{\text{LP}}$. To see why D_t^{LP} is ϵ_N^{rd} -maximal feasible set, observe that $\delta(X_t, D_t^{\text{LP}}) = \delta(X_t, D_{t-1}^{\text{LP}}) \geq 0$; moreover, the infeasibility of the SOCP implies that there is no $D' \supseteq D_t^{\text{LP}}$ such that $\delta(X_t, D') \geq \epsilon_N^{\text{rd}}$.

If the SOCP is feasible, let the optimal solution solved from this SOCP be $(\mathbf{z}^*, m^*)$. We let D_t consist of $\lfloor Nz^*(s) \rfloor$ arms with state s for each state $s \in \mathbb{S}$. Then we have $X_t(D_{t-1}^{\text{LP}}, s) \leq X_t(D_t^{\text{LP}}, s) \leq X_t([N], s)$, so it is equivalent to letting $D_t^{\text{LP}} \supseteq D_{t-1}^{\text{LP}}$. Next, we show that D_t^{LP} is a ϵ_N^{rd} -maximal feasible set. We first show that $\delta(X_t, D_t^{\text{LP}}) \geq 0$. Observe that $\|X_t(D_t^{\text{LP}}) - \mathbf{z}^*\|_1 \leq |\mathbb{S}|/N$, so

$$\begin{aligned} & \left| \|X_t(D_t^{\text{LP}}) - m(D_t^{\text{LP}})\mu^*\|_U - \|\mathbf{z}^* - m^*\mu^*\|_U \right| \\ & \leq \|(X_t(D_t^{\text{LP}}) - \mathbf{z}^*) - (m(D_t^{\text{LP}}) - m^*)\mu^*\|_U \\ & \leq \|(X_t(D_t^{\text{LP}}) - \mathbf{z}^*)\|_U + |m(D_t^{\text{LP}}) - m^*| \|\mu^*\|_U \\ & \leq \lambda_U^{1/2} \|(X_t(D_t^{\text{LP}}) - \mathbf{z}^*)\|_2 + \lambda_U^{1/2} |m(D_t^{\text{LP}}) - m^*| \|\mu^*\|_2 \\ & \leq \lambda_U^{1/2} \|(X_t(D_t^{\text{LP}}) - \mathbf{z}^*)\|_1 + \lambda_U^{1/2} |m(D_t^{\text{LP}}) - m^*| \|\mu^*\|_1 \\ & \leq \frac{2\lambda_U^{1/2}|\mathbb{S}|}{N}. \end{aligned}$$

Therefore,

$$\begin{aligned} \delta(X_t, D_t^{\text{LP}}) &= \eta m(D_t^{\text{LP}}) - \|X_t(D_t^{\text{LP}}) - m(D_t^{\text{LP}})\mu^*\|_U - \frac{|S^\emptyset| + 1}{N} \\ &\geq m^* - \frac{|\mathbb{S}|}{N} - \|\mathbf{z}^* - m^*\mu^*\|_U - \frac{2\lambda_U^{1/2}|\mathbb{S}|}{N} - \frac{|S^\emptyset| + 1}{N} \\ &\geq \epsilon_N^{\text{rd}} - \frac{(1 + 2\lambda_U^{1/2})|\mathbb{S}|}{N} \\ &= 0. \end{aligned}$$

To show the maximality of D_t^{LP} , observe that for any $D' \supseteq D_t^{\text{LP}}$ satisfying $\delta(X_t, D') \geq \epsilon_N^{\text{rd}}$, $(X_t(D'), m(D'))$ satisfies all the constraints of this SOCP, so $m(D') \leq m^* \leq m(D_t^{\text{LP}}) + |\mathbb{S}|/N \leq m(D_t^{\text{LP}}) + \epsilon_N^{\text{rd}}$.

Caes 2: When $\delta(X_t, D_{t-1}^{\text{LP}}) < 0$. In this case, we solve two SOCPs. The first SOCP is given by

$$\underset{\mathbf{z}, m}{\text{maximize}} \quad \sum_{s \in \mathbb{S}} z(s) \quad (5.170a)$$

$$\text{subject to} \quad \|\mathbf{z} - m\mu^*\|_U \leq \eta m - \frac{|S^\emptyset| + 1}{N} - \epsilon_N^{\text{rd}} \quad (5.170b)$$

$$\sum_{s \in \mathbb{S}} z(s) = m \quad (5.170c)$$

$$0 \leq z(s) \leq X_t(D_{t-1}^{\text{LP}}, s) \quad \forall s \in \mathbb{S}. \quad (5.170d)$$

We define an intermediate set $D_{\text{temp}}^{\text{LP}}$ based on the outcome of solving this SOCP. If this SOCP is infeasible, we let $D_{\text{temp}}^{\text{LP}} = \emptyset$. If the SOCP is feasible, let the optimal solution be $(\mathbf{z}^*, m^*)$. We let the set $D_{\text{temp}}^{\text{LP}}$ consist of $\lfloor Nz^*(s) \rfloor$ arms with state s , for each $s \in \mathbb{S}$. One can verify that the set can be chosen as a subset of D_{t-1}^{LP} because $z^*(s) \leq X_t(D_{t-1}^{\text{LP}}, s)$. Also, the constraint (5.170b) ensures that $D_{\text{temp}}^{\text{LP}}$ satisfies $\delta(X_t, D_{\text{temp}}^{\text{LP}}) \geq 0$, following a similar argument as before.

After obtaining $D_{\text{temp}}^{\text{LP}}$, we solve the second SOCP given by

$$\underset{\mathbf{z}, m}{\text{maximize}} \quad \sum_{s \in \mathbb{S}} z(s) \quad (5.171a)$$

$$\text{subject to} \quad \|\mathbf{z} - m\mu^*\|_U \leq \eta m - \frac{|S^\emptyset| + 1}{N} - \epsilon_N^{\text{rd}} \quad (5.171b)$$

$$\sum_{s \in \mathbb{S}} z(s) = m \quad (5.171c)$$

$$X_t(D_{\text{temp}}^{\text{LP}}, s) \leq z(s) \leq X_t([N], s) \quad \forall s \in \mathbb{S}. \quad (5.171d)$$

If this SOCP is infeasible, we let $D_t^{\text{LP}} = D_{\text{temp}}^{\text{LP}}$. We argue that D_t^{LP} is an ϵ_N^{rd} -maximal feasible set: first, because $\delta(X_t, D_{\text{temp}}^{\text{LP}}) \geq 0$ or $D_{\text{temp}}^{\text{LP}} = \emptyset$, the same holds for D_t^{LP} ; then observe that for any D' such that $D' \supseteq D_t^{\text{LP}} = D_{\text{temp}}^{\text{LP}}$ and $\delta(X_t, D') \geq \epsilon_N^{\text{rd}}$, $(X_t(D'), m(D'))$ satisfies the constraint of the SOCP, so such D' does not exist by the infeasibility, which implies that $D_{\text{temp}}^{\text{LP}}$ is ϵ_N^{rd} -maximal feasible.

If this SOCP is feasible, we let its optimal solution be $(\mathbf{z}^*, m^*)$. We let the set D_t^{LP} consist of $\lfloor Nz^*(s) \rfloor$ arms with state s , for each $s \in \mathbb{S}$. The constraint (5.171b) ensures that D_t^{LP} satisfies $\delta(X_t, D_t^{\text{LP}}) \geq 0$, following a similar argument as before. Moreover, for any set $D' \supseteq D_t^{\text{LP}} \supseteq D_{\text{temp}}^{\text{LP}}$ and

such that $\delta(X_t, D_t) \geq \epsilon_N^{\text{rd}}$, $(X_t(D'), m(D'))$ satisfies the constraints of this SOCP, so $m(D') \leq m^* \leq m(D_t^{\text{LP}}) + |\mathbb{S}|/N \leq m(D_t^{\text{LP}}) + \epsilon_N^{\text{rd}}$. Therefore, D_t^{LP} is ϵ_N^{rd} -maximal feasible.

Deciding actions for the arms in $(D_t^{\text{LP}} \cup D_t^{\bar{\pi}^*})^c$

In our implementation of the two-set policy, we choose the actions for the arms in $(D_t^{\text{LP}} \cup D_t^{\bar{\pi}^*})^c$ following similar ideas as the ID policy in Chapter 4 [153]. Specifically, we generate a random ideal action $\hat{A}_t(i)$ and temporarily set $A_t(i) = \hat{A}_t(i)$ for each $i \in (D_t^{\text{LP}} \cup D_t^{\bar{\pi}^*})^c$. If

$$\sum_{i \in [N]} A_t(i) > \alpha N,$$

we change $A_t(i)$ to 0 for $i \in (D_t^{\text{LP}} \cup D_t^{\bar{\pi}^*})^c$ starting from large i 's to small i 's, until $\sum_{i \in [N]} A_t(i) = \alpha N$. Similarly, if

$$\sum_{i \in [N]} A_t(i) < \alpha N,$$

we change $A_t(i)$ to 1 for $i \in (D_t^{\text{LP}} \cup D_t^{\bar{\pi}^*})^c$ starting from large i 's to small i 's, until $\sum_{i \in [N]} A_t(i) = \alpha N$. As guaranteed by Lemma 5.28, we can always achieve $\sum_{i \in [N]} A_t(i) = \alpha N$ through this process.

5.E.2 Details of RB instances simulated in Section 5.6

Definitions of the instances in Figure 5.1

Here, we provide some details on the RB instances simulated in Figure 5.1a and Figure 5.1b.

The RB instance corresponding to Figure 5.1a is defined in Chapter 4 [153]. We repeat its definition below for completeness. This RB instance is defined by a single-armed MDP with state space $\mathbb{S} = \{0, 1, \dots, 7\}$. Each state in $\mathbb{S}$ is labelled by a *preferred action*, which is chosen to be action 1 for states $\{0, 1, 2, 3\}$ and action 0 for states $\{4, 5, 6, 7\}$. If an arm is in state s and takes the preferred action, it moves to state $(s+1) \bmod 8$ with probability $p_{s,\text{R}}$, and stays in state s otherwise; if it does not take the preferred action, it moves to

state $(s - 1)^+$ with probability $p_{s,L}$; here the probabilities $p_{s,R} = 0.1$ for all $s \in \mathbb{S}$, and $p_{s,L}$ is defined as

$$p_{s,L} = \begin{cases} 1 & \text{if } s = 0, 1 \\ 0.5 - 0.1s/8 & \text{otherwise.} \end{cases}$$

The reward function is defined as $r(7, 0) = 0.1$, $r(0, 1) = 1/300$, and $r(s, a) = 0$ for all the other $s \in \mathbb{S}, a \in \mathbb{A}$. The budget parameter $\alpha = 0.45$. We initialize the states to be uniformly distributed over $\{4, 5, 6, 7\}$.

Here we briefly comment on the rationale of constructing this example. The main purpose of this transition structure is to let the LP index policy use a suboptimal priority order. Specifically, the LP index policy in this example prioritizes state 3 over state 5, and state 5 over state 4. Thus, if at a certain time step, all arms are in states $\{3, 4, 5\}$ with a suitable distribution, all arms in state 3 and 5 will be activated. Consequently, the arms in 3 or 5 will return to state 4, so all arms are stuck in these three states without getting any reward.

Apart from the transition structure, the value of the reward function $r(0, 1)$ is also specially chosen to make the problem non-indexable. The budget α is chosen to be less than 0.5, so that state 0 is the neutral state.

The RB instance in Figure 5.1b is constructed with similar ideas. The single-armed MDP of this instance has the state space $\mathbb{S} = \{0, 1, 2, \dots, 12\}$; the preferred actions are 1 for states $\{0, 1, \dots, 5\}$, and are 0 for the rest of the states. The transition rules after taking the preferred (non-preferred) action are the same as the last example, with $p_{\cdot,L}$ adjusted to

$$p_{s,L} = \begin{cases} 1 & \text{if } s = 0, 1 \\ 0.5 - 0.1s/12 & \text{otherwise.} \end{cases}$$

We let the reward function be defined as $r(7, 0) = 0.1$ and $r(0, 1) = 1/300$, and $r(s, a) = 0$ for all the other $s \in \mathbb{S}, a \in \mathbb{A}$. We let the budget parameter $\alpha = 0.4$. We initialize the states to be uniformly distributed over $\{6, 7, 8, 9, 10, 11\}$. In this example, the LP index policy prioritizes state 5 over state 7, and state 7 over state 6, so intuitively, the arms could be stuck on these three states.

One can potentially generalize the above two examples, and construct more RB instances that satisfy Assumptions 5.1–5.3, but violate UGAP under the Whittle index policy and the LP index policy.

Details of generating uniformly random examples

Finally, we comment on the details of generating uniformly random examples for the experiments in Figures 5.2 and 5.3: For each $s \in \mathbb{S}$ and $a \in \mathbb{A}$, we let $P(s, a, \cdot)$ be a random vector sampled from the uniform distribution on the simplex $\Delta(\mathbb{S})$; for each $a \in \mathbb{S}$, let $r(\cdot, a)$ be a random vector sampled from the uniform distribution on $\Delta(\mathbb{S})$. We let all the random vectors described above to be independent.

Chapter 6

Projection-Based Lyapunov Method for Fully Heterogeneous Weakly-Coupled MDPs

6.1 Introduction

Heterogeneity poses a fundamental challenge for many real-world decision-making problems, where each problem consists of a large number of interacting components. However, despite its practical significance, heterogeneity remains largely understudied in the literature. In this chapter, we study *heterogeneous* settings of a prominent class of such problems, known as weakly-coupled Markov decision processes (WCMDPs) [142]. A WCMDP consists of N *arms* (or *subproblems*), where each arm itself is a Markov decision process (MDP). In a heterogeneous setting, the MDPs could be distinct. At each time step, the decision-maker selects an action for each arm, which affects the arm's transition probabilities and reward, and then the arms make state transitions independently. However, these actions are subject to a set of global *budget constraints*, where each constraint limits one type of total cost across all arms at each time step. The objective is to find a policy that maximizes the long-run *average reward* over an infinite time horizon. We focus on the *planning* setting, where all the model parameters (reward function, cost functions, budget, and transition kernel) are known.

WCMDPs have been used to model a wide range of applications, including online advertising [40, 304], job scheduling [295], healthcare [36], surveillance

[268], and machine maintenance [115]. A faithful modeling of these applications calls for *heterogeneity*. For instance, in [36], arms are beneficiaries of a health program and they could react to interventions differently; in [268], arms are targets of surveillance who have different locations and probabilities to be exposed; in [115], arms are machines that could require distinct repair schedules.

Although heterogeneity is crucial in the modeling of these applications, most existing work on average-reward WCMDPs establishes asymptotic optimality only for the homogeneous setting where all arms share the same set of model parameters [107, 108, 121, 144, 149, 151, 153, 266, 274, 294]. Only a few exceptions [149, 266, 293] address heterogeneity, but in highly specialized settings. Among these, a common approach to handle heterogeneity is to consider the *typed heterogeneous* setting, where the N arms are divided into a constant number of types as N scales up, with each type having distinct model parameters. While heterogeneous WCMDPs have been studied under the finite-horizon total-reward and discounted-reward criteria, these results do not extend to the average-reward setting we consider. We review related work in more detail at the end of this section and also in Appendix 6.A.

The key distinction between the homogeneous (or typed heterogeneous) setting and the fully heterogeneous setting is whether the arms can be divided into a *constant* number of homogeneous groups. In the former, the system dynamics depends only on the fraction of arms in each state in each homogeneous group. Thus, the effective dimension of the state space is *polynomial* in N . In contrast, in the fully heterogeneous setting, the state space grows *exponentially* in N , making the problem truly high-dimensional.

Our contribution. In this chapter, we study *fully heterogeneous* WCMDPs. We propose a policy we call the *ID policy with reassignment*, which generalizes the ID policy of Chapter 4 [153]. In our policy, we first perform an ID reassignment algorithm to reorder the arms, which ensures proper arm prioritization during policy execution. We then run a variant of the ID policy adapted to handle heterogeneity, which consists of two phases. The first phase is a pre-processing phase, where we compute an *optimal single-armed policy* for each arm (denoted as $\bar{\pi}_i^*$ for the i -th arm) that prescribes the *ideal action* the arm would take at each state. The second phase is the real-time phase. At each time step, the policy iterates over the arms according to their reassigned IDs, and it lets as many arms as possible follow their respective

ideal actions while satisfying the budget constraints. Unlike the original ID policy, which has only one optimal single-armed policy due to homogeneous arms, our policy computes N optimal single-armed policies, one for each arm. Our proposed policy is efficiently computable, with computational complexity polynomial in N .

We prove that the proposed ID policy with reassignment achieves an $O(1/\sqrt{N})$ optimality gap under mild assumptions as the number of arms N becomes large. Here, the optimality gap refers to the difference between the long-run average reward per arm under our policy and that under the optimal policy. *This is the first result establishing asymptotic optimality for fully heterogeneous average-reward WCMDPs.*

We remark that the original ID policy was designed for a special case of WCMDPs known as restless bandits, and it is for the homogeneous setting. While the generalization in our proposed policy is natural, identifying the appropriate generalization and establishing its optimality gap in the fully heterogeneous setting are technically challenging and require new theoretical approaches.

Technical novelty. The main technical innovation of the chapter is the introduction of a novel Lyapunov function for fully heterogeneous WCMDPs. Specifically, to prove the asymptotic optimality of a policy, a key step is to show that the system state is globally attracted to an *optimal region* where most arms can follow the ideal actions generated by their respective optimal single-armed policies $\bar{\pi}_i^*$'s. In the homogeneous setting, we can prove such convergence using state aggregation techniques that rely on the symmetry of arms. However, in the heterogeneous setting, states of different arms cannot be aggregated since arms are no longer symmetric. Our technique is to *project* arm states onto a set of carefully selected feature vectors, and define the Lyapunov function based on these projections. These feature vectors encode the minimal amount of information needed to evaluate the relevant functions of the system state (e.g., instantaneous reward or cost) and predict their future expectations. This projection-based Lyapunov function provides a principled way to measure deviations of the system state from the optimal region in a fully heterogeneous setting. A more detailed discussion of this approach can be found in Section 6.4.

Beyond WCMDPs, our techniques have the potential to be applied to more general heterogeneous large stochastic systems. Heterogeneity has

been a topic of strong interest in these systems, but it is known to be a challenging problem with limited theoretical results. Only recently have there been notable breakthroughs. [8, 9] extended the popular mean-field analysis to a class of heterogeneous large stochastic systems for the first time, but the results are only for transient distributions. Another line of work [302, 303] studied heterogeneous load-balancing systems. They first analyzed the transient distributions and then used interchange-of-limits arguments to extend the results to steady state. Our method provides a more direct framework for steady-state analysis and has the potential to generalize to a broader range of heterogeneous stochastic systems.

Related work. WCMDPs have been extensively studied with a rich body of literature. Here we briefly overview the most relevant work and refer the reader to Appendix 6.A for a detailed survey.

We first focus on the *average-reward* criterion. As mentioned earlier, most existing work considers the *homogeneous setting*. Early work on WCMDPs primarily focuses on the *restless bandit (RB)* problem, where each arm's MDP has a binary action space (active and passive actions) and there is only one budget constraint that limits the total number of active actions across all arms at each time step. The seminal work by Whittle [283] introduced the RB problem and the celebrated Whittle index policy, which was later shown to achieve an $o(1)$ optimality gap as $N \rightarrow \infty$ under a set of conditions [274]. Subsequent work on RBs has focused on designing policies that achieve asymptotic optimality under more relaxed conditions [149, 153, 266, 294], as well as improving the optimality gap to $O(1/\sqrt{N})$ [149, 153] or $O(\exp(-cN))$ [107, 108, 151]. Among these papers, [149, 266] address heterogeneous RBs. However, [266] focuses on the *typed heterogeneous* setting, where the N arms are divided into a constant number of types as $N \rightarrow \infty$. The paper [149] discussed in Chapter 3 includes an extension to the fully heterogeneous setting. However, their result relies on the Synchronization Assumption. The policies in both papers cannot be straightforwardly extended to general WCMDPs, which have multiple actions, multiple budget constraints, and state-dependent cost functions.

Beyond RBs, work on general average-reward WCMDPs is scarce. The closest to ours are [144, 266, 293], which studied WCMDPs with a single budget constraint and established $o(1)$ optimality gaps, but again in the homogeneous setting [144] or the *typed heterogeneous* setting [266, 293]. More

recently, [121] proved the first $o(1)$ optimality gap result for WCMDPs with general budget constraints, but in the *homogeneous* setting.

Under the *finite-horizon total-reward* or *discounted-reward* criteria, there has been more work on heterogeneous settings, including both the typed heterogeneous setting [87, 111] and, more recently, the fully heterogeneous setting [56–58, 297]. However, the optimality gap in these papers generally grows *super-linearly* with the (effective) time horizon, except under restrictive conditions. Consequently, it is difficult to extend these results to the average-reward setting and still achieve asymptotic optimality.

Although our work focuses on the planning setting where all model parameters are known, there has been growing interest in developing reinforcement learning algorithms for the learning setting with unknown parameters [24, 35, 88, 167, 168, 216, 217, 237, 292, 293]. Many of these approaches rely on well-designed planning policies as a foundation to achieve learning efficiency. In this context, our results can serve as an important building block for developing model-based learning algorithms for fully heterogeneous WCMDPs.

General notation. Let $\mathbb{R}$, $\mathbb{N}$, and $\mathbb{N}_+$ denote the sets of real numbers, non-negative integers, and positive integers, respectively. Let $[N] \triangleq \{1, 2, \dots, N\}$ for any $N \in \mathbb{N}_+$ and $[n_1 : n_2] \triangleq \{n_1, n_1 + 1, \dots, n_2\}$ for $n_1, n_2 \in \mathbb{N}_+$ with $n_1 \leq n_2$. Let $[0, 1]_N = \{i/N : i \in \mathbb{N}, 0 \leq i/N \leq 1\}$, the set of integer multiples of $1/N$ in $[0, 1]$. For a matrix $A \in \mathbb{R}^{d \times d}$, we denote its operator norm as $\|A\|_p = \sup_{x \neq \mathbf{0}} \|Ax\|_p / \|x\|_p$. We use boldface letters to denote matrices, and regular letters to denote vectors and scalars. We write $\mathbb{R}^{\mathbb{S}}$ for the set of real-valued vectors indexed by elements of $\mathbb{S}$, or equivalently, the set of real-valued functions on $\mathbb{S}$; for each $v \in \mathbb{R}^{\mathbb{S}}$, let $v(s)$ to denote its element corresponding to $s \in \mathbb{S}$.

6.2 Problem setup

We consider a weakly-coupled Markov decision process (WCMDP) that consists of N arms. Each arm has an ID $i \in [N]$ and is associated with a smaller MDP denoted as $\mathcal{M}_i = (\mathbb{S}, \mathbb{A}, \mathbb{P}_i, r_i, (c_{k,i})_{k \in [K]})$. Here $\mathbb{S}$ and $\mathbb{A}$ are the state space and the action space, respectively, both assumed to be finite; $\mathbb{P}_i$ describes the transition probabilities with $\mathbb{P}_i(s' | s, a)$ being the transition probability from state s to state s' when action a is taken. The state tran-

sitions of different arms are independent given the actions. When arm i is in state s and we take action a , a reward $r_i(s, a)$ is generated, as well as K types of costs $c_{k,i}(s, a)$, $k \in [K]$. We assume that the costs are nonnegative, i.e., $c_{k,i}(s, a) \geq 0$ for all $i \in \mathbb{N}_+$, $k \in [K]$, $s \in \mathbb{S}$, and $a \in \mathbb{A}$. Note that we allow the arms to be *fully heterogeneous*, i.e., the $\mathcal{M}_i$'s can be *all distinct*.

When taking an action for each arm in this N -armed system, we are subject to budget constraints. Specifically, suppose each arm i is in state s_i . Then the actions, a_i 's, should satisfy the constraints:

$$\sum_{i \in [N]} c_{k,i}(s_i, a_i) \leq \alpha_k N, \quad \forall k \in [K], \quad (6.1)$$

where each $\alpha_k > 0$ is a constant independent of N , and $\alpha_k N$ is referred to as the *budget* for type- k cost. We assume that there exists an action $0 \in \mathbb{A}$ that does not incur any type of cost for any arm at any state, i.e., $c_{k,i}(s, 0) = 0$ for all $k \in [K]$, $i \in [N]$, $s \in \mathbb{S}$. This assumption guarantees that there always exist valid actions (e.g., taking action 0 for every arm) regardless of the states of the arms.

Policy and system state. A policy π for the N -armed problem specifies the action for each of the N arms, in a possibly history-dependent way. Under policy π , let $S_{i,t}^\pi$ denote the state of the i th arm at time t , and we refer to $\mathbf{S}_t^\pi \triangleq (S_{i,t}^\pi)_{i \in [N]}$ as the *system state*. Similarly, let $A_{i,t}^\pi$ denote the action applied to arm i at time t , and we refer to $\mathbf{A}_t^\pi \triangleq (A_{i,t}^\pi)_{i \in [N]}$ as the *system action*. In this chapter, we also use an alternative representation of the system state, denoted as $\mathbf{X}_t^\pi$ and defined as follows. Let $X_{i,t}^\pi = (X_{i,t}^\pi(s))_{s \in \mathbb{S}} \in \mathbb{R}^{|\mathbb{S}|}$ be a row vector where the entry corresponding to state s is given by $X_{i,t}^\pi(s) = \mathbf{1}\{S_{i,t}^\pi = s\}$; i.e., $X_{i,t}^\pi$ is a one-hot row vector whose s 's entry is 1 if $S_{i,t}^\pi = s$ and is 0 otherwise. Then let $\mathbf{X}_t^\pi$ be an $N \times |\mathbb{S}|$ matrix whose i th row is $X_{i,t}^\pi$. It is easy to see that $\mathbf{X}_t^\pi$ contains the same information as $\mathbf{S}_t^\pi$, and we refer to both of them as the system state. In this chapter, we often encounter vectors like $X_{i,t}^\pi = (X_{i,t}^\pi(s))_{s \in \mathbb{S}}$, whose entries correspond to different states in $\mathbb{S}$. For such vectors, say u and v , we use the inner product to write a sum for convenience $\langle u, v \rangle \triangleq \sum_{s \in \mathbb{S}} u(s)v(s)$. We sometimes omit the superscript π when it is clear from context.

Maximizing average reward. Our objective is to maximize the long-run time-average reward subject to the budget constraints. To be more

precise, we follow the treatment for maximizing average reward in [230]. For any policy π and an initial state $\mathbf{S}_0$ of the N -armed system, consider the *limsup* average reward $R^+(\pi, \mathbf{S}_0)$ and the *liminf* average $R^-(\pi, \mathbf{S}_0)$, defined as $R^+(\pi, \mathbf{S}_0) = \limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i \in [N]} \mathbb{E} [r_i(S_{i,t}^\pi, A_{i,t}^\pi)]$ and $R^-(\pi, \mathbf{S}_0) = \liminf_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i \in [N]} \mathbb{E} [r_i(S_{i,t}^\pi, A_{i,t}^\pi)]$. If $R^+(\pi, \mathbf{S}_0) = R^-(\pi, \mathbf{S}_0)$, then the average reward of policy π under the initial condition $\mathbf{S}_0$ exists and is defined as

$$R(\pi, \mathbf{S}_0) = R^+(\pi, \mathbf{S}_0) = R^-(\pi, \mathbf{S}_0) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i \in [N]} \mathbb{E} [r_i(S_{i,t}^\pi, A_{i,t}^\pi)]. \quad (6.2)$$

Note that these reward notions divide the total reward from all arms by the number of arms, N , measuring the reward *per arm*. The WCMDP problem is to solve the following optimization problem:

$$\underset{\text{policy } \pi}{\text{maximize}} \quad R^-(\pi, \mathbf{S}_0) \quad (6.3a)$$

$$\text{subject to} \quad \sum_{i \in [N]} c_{k,i}(S_{i,t}^\pi, A_{i,t}^\pi) \leq \alpha_k N, \quad \forall k \in [K], \forall t \geq 0. \quad (6.3b)$$

Let the optimal value of this problem be denoted as $R^*(N, \mathbf{S}_0)$. Note that since the WCMDP is an MDP with finite state and action space, if we replace the $R^-(\pi, \mathbf{S}_0)$ in the objective (6.3a) with $R^+(\pi, \mathbf{S}_0)$, the optimal value stays the same [230, Proposition 9.1.6].

Asymptotic optimality. Recall that exactly solving the WCMDP problem is PSPACE-hard [220]. In this chapter, our goal is to design a policy π that is *efficiently computable* and *asymptotically optimal* as $N \rightarrow \infty$, with the following notion for asymptotic optimality. For any policy π , we define its *optimality gap* as $R^*(N, \mathbf{S}_0) - R^-(\pi, \mathbf{S}_0)$. We say the policy π is *asymptotically optimal* if

$$R^*(N, \mathbf{S}_0) - R^-(\pi, \mathbf{S}_0) = o(1) \quad \text{as } N \rightarrow \infty. \quad (6.4)$$

When we take this asymptotic regime as $N \rightarrow \infty$, we keep the number of constraints, K , as well as the budget coefficients, $\alpha_1, \alpha_2, \dots, \alpha_K$, fixed. We assume that the reward functions and cost functions are uniformly bounded, i.e., $\sup_{i \in \mathbb{N}_+} \max_{s \in \mathbb{S}, a \in \mathbb{A}} |r_i(s, a)| \triangleq r_{\max} < \infty$ and $\sup_{i \in \mathbb{N}_+} \max_{k \in [K], s \in \mathbb{S}, a \in \mathbb{A}} c_{k,i}(s, a) \triangleq c_{\max} < \infty$. This notion for asymptotic

optimality is consistent with that in the existing literature (e.g., [266, Definition 4.11]). We are interested in not only achieving asymptotic optimality but also characterizing the *order* of the optimality gap.

In the remainder of this chapter, we focus on stationary Markov policies, which are sufficient for achieving the optimal value because the WCMDP problem is an MDP with finite state and action spaces [230, Theorem 9.1.8]. Under any stationary Markovian policy, the long-run reward $R(\pi, \mathbf{S}_0) = R^+(\pi, \mathbf{S}_0) = R^-(\pi, \mathbf{S}_0)$ is always well-defined [230, Theorem 8.1.1].

LP relaxation and an upper bound on optimality gap. We consider the linear program (LP) below, which will play a critical role in performance analysis and policy design:

$$R_N^{\text{rel}} \triangleq \begin{aligned} & \underset{(y_i(s,a))_{i \in [N], s \in \mathbb{S}, a \in \mathbb{A}}}{\text{maximize}} && \frac{1}{N} \sum_{i \in [N]} \sum_{s \in \mathbb{S}, a \in \mathbb{A}} y_i(s, a) r_i(s, a) \end{aligned} \quad (6.5a)$$

$$\text{subject to} \quad \frac{1}{N} \sum_{i \in [N]} \sum_{s \in \mathbb{S}, a \in \mathbb{A}} y_i(s, a) c_{k,i}(s, a) \leq \alpha_k, \quad \forall k \in [K], \quad (6.5b)$$

$$\begin{aligned} & \sum_{s' \in \mathbb{S}, a' \in \mathbb{A}} \mathbb{P}_i(s | s', a') y_i(s', a') = \sum_{a \in \mathbb{A}} y_i(s, a), \\ & \forall s \in \mathbb{S}, \forall i \in [N], \end{aligned} \quad (6.5c)$$

$$\begin{aligned} & \sum_{s' \in \mathbb{S}, a' \in \mathbb{A}} y_i(s', a') = 1, \quad y_i(s, a) \geq 0, \\ & \forall s \in \mathbb{S}, \forall a \in \mathbb{A}, \forall i \in [N]. \end{aligned} \quad (6.5d)$$

Lemma 6.1 below establishes a connection between this LP and the WCMDP.

Lemma 6.1. *The optimal value of any N -armed WCMDP problem is upper bounded by the optimal value of the corresponding linear program in (6.5), i.e.,*

$$R^*(N, \mathbf{S}_0) \leq R_N^{\text{rel}}, \quad \forall N, \forall \mathbf{S}_0.$$

An immediate implication of Lemma 6.1 is that for any policy π , its optimality gap is upper bounded as

$$R^*(N, \mathbf{S}_0) - R^-(\pi, \mathbf{S}_0) \leq R_N^{\text{rel}} - R^-(\pi, \mathbf{S}_0). \quad (6.6)$$

Therefore, to derive an upper bound for the optimality gap, it suffices to control $R_N^{\text{rel}} - R^-(\pi, \mathbf{S}_0)$, which we will show is $O(1/\sqrt{N})$ in Theorem 6.3.

To see the intuition of Lemma 6.1, we interpret the optimization variable $y_i(s, a)$ as the long-run fraction of time arm i spends in state s and takes action a . We refer to $y_i(s, a)$ as arm i 's *state-action frequency* for the state-action pair (s, a) . Then the constraints in (6.5b) of the LP can be viewed as relaxations of the budget constraints in (6.3b) for the WCMDP. The constraints in (6.5c)–(6.5d) guarantee that $y_i(s, a)$'s are proper stationary time fractions. Therefore, the LP is a relaxation of the WCMDP and thus achieves a higher optimal value. The formal proof of Lemma 6.1 is provided in Appendix 6.C.

Our LP (6.5) serves a similar role to the LP used in previous work on restless bandits and WCMDPs (see, e.g., [108, 121, 144, 149, 266, 274, 293]), but with different forms and dependencies on N . Both our LP and the LP in previous work relax the hard budget constraints to time-average constraints. However, in the homogeneous arm setting [108, 121, 144, 274], the LP has only one set of state-action frequencies $y(s, a)$, and the LP is independent of N . As a result, both the optimal value of the LP and the complexity of solving it are independent of N . Prior work for the typed-heterogeneous setting [149, 266, 293] divides the arms into a constant number K of types and defines a set of state-action frequency $y_k(s, a)$ for each type k . The optimal value of the resulting LP and the complexity of solving it depend on the number of types K . Chapter 3 [149] includes a generalization to the fully heterogeneous setting, resulting in an LP similar to ours, although restricted to restless bandits.

In our LP, we define a separate set of state-action frequencies $y_i(s, a)$ for each arm $i \in [N]$, making the LP explicitly depend on N . Therefore, the optimal value R_N^{rel} depends on N , and the complexity of solving the LP grows with N . Nevertheless, because the number of variables and constraints scales linearly with N , our LP can still be solved in polynomial time.

6.3 ID policy with reassignment

In this section, we introduce the ID policy with reassignment, generalized from the ID policy designed for homogeneous restless bandits in Chapter 4 [153]. Our policy first performs an ID reassignment procedure, and then proceeds to run a variant of the ID policy adapted to handle heterogeneity. We begin by

introducing a building block of our policy, referred to as optimal single-armed policies, followed by the ID reassignment algorithm and the execution of the adapted ID policy.

Optimal single-armed policies. Once we obtain a solution to the LP in (6.5), we can construct a policy for each arm i , which we refer to as an *optimal single-armed policy* for arm i . In particular, let $(y_i^*(s, a))_{i \in [N], s \in \mathbb{S}, a \in \mathbb{A}}$ be an arbitrary optimal solution to the LP in (6.5). Then for arm i , the optimal single-armed policy, $\bar{\pi}_i^*$, is defined as

$$\bar{\pi}_i^*(a | s) = \begin{cases} \frac{y_i^*(s, a)}{\sum_{a \in \mathbb{A}} y_i^*(s, a)}, & \text{if } \sum_{a \in \mathbb{A}} y_i^*(s, a) > 0, \\ \frac{1}{|\mathbb{A}|}, & \text{if } \sum_{a \in \mathbb{A}} y_i^*(s, a) = 0, \end{cases} \quad (6.7)$$

where $\bar{\pi}_i^*(a | s)$ is the probability of taking action a given that the arm's current state is s . Note that due to heterogeneity, this optimal single-armed policy $\bar{\pi}_i^*$ can be different for different arms.

The rationale behind these policies is as follows. If each arm i individually follows its optimal single-armed policy $\bar{\pi}_i^*$, then the average reward per arm (total reward divided by N) achieves the upper bound R_N^{rel} given by the LP. However, this strategy only guarantees that the budget constraints are satisfied in a *time-average* sense, rather than conforming to the *hard* constraints in the original N -armed WCMDP. Thus, having each arm follow its optimal single-armed policy is not a valid policy for the original N -armed problem. Nevertheless, these optimal single-armed policies $\bar{\pi}_i^*$'s serve as a guide for how the arms should ideally behave to maximize rewards. The ID policy uses the $\bar{\pi}_i^*$'s as a reference. It is then designed to ensure that even under the hard budget constraints, most arms follow their optimal single-armed policies most of the time, yielding a diminishing gap to R_N^{rel} in reward.

ID reassignment. We first define a few quantities that will be used in the ID reassignment algorithm. For each arm $i \in [N]$ and each cost type $k \in [K]$, the expected cost under the optimal single-armed policy is defined as $C_{k,i}^* = \sum_{s \in \mathbb{S}, a \in \mathbb{A}} y_i^*(s, a) c_{k,i}(s, a)$. Based on $C_{k,i}^*$'s, we divide the budget constraints into *active* constraints and *inactive* constraints as follows. For each cost type $k \in [K]$, we say the type- k budget constraint is *active* if

$$\sum_{i \in [N]} C_{k,i}^* \geq \frac{\alpha_k}{2} N, \quad (6.8)$$

Algorithm 11 ID reassignment

```

1: Input: optimal state-action frequencies  $(y_i^*(s, a))_{i \in [N], s \in \mathbb{S}, a \in \mathbb{A}}$ , budgets
    $(\alpha_k)_{k \in [K]}$ 
2: Output: new arm ID, recorded in  $\text{newID}(i)$ , for each arm with old ID
    $i \in [N]$ 
3: Compute  $(C_{k,i}^*)_{i \in [N], k \in [K]}$  and the set of active constraints  $\mathcal{A}$  using (6.8)
4: if  $\mathcal{A} = \emptyset$  then
5:    $\text{newID}(i) \leftarrow i$  for all  $i \in [N]$   $\triangleright$  No need for ID reassignment
6: else
7:   Initialize  $\mathcal{F} \leftarrow \emptyset$   $\triangleright$  Set of arms that have been assigned new IDs
8:   Initialize  $\mathcal{D}_k \leftarrow \{i \in [N] : C_{k,i}^* \geq \delta\}$  for all  $k \in \mathcal{A}$ 
9:    $\delta \leftarrow \alpha_{\min}/4 \triangleq \min_{k \in [K]} \alpha_k/4$ ;  $d \leftarrow \left\lceil \frac{(c_{\max} - \delta)K}{\alpha_{\min}/2 - \delta} \right\rceil$ 
10:  for  $\ell = 0, 1, \dots, \lfloor N/d \rfloor - 1$  do
11:     $\mathcal{I}(\ell) \leftarrow [\ell d + 1 : (\ell + 1)d]$ ;  $j \leftarrow \ell d + 1$ 
12:    for  $k \in \mathcal{A}$  do
13:      if  $\sum_{i \in \mathcal{F}} C_{k,i}^* \mathbf{1}\{\text{newID}(i) \in \mathcal{I}(\ell)\} < \delta$  then
14:        Pick any  $i$  from  $\mathcal{D}_k$ ; set  $\text{newID}(i) \leftarrow j$  and add  $i$  to  $\mathcal{F}$ 
15:        Remove  $i$  from  $\mathcal{D}_{k'}$  for all  $k'$ 
16:         $j \leftarrow j + 1$ 
17:  For all  $i \in [N] \setminus \mathcal{F}$ , assign values to their  $\text{newID}(i)$ 's randomly from
    $[N] \setminus \{\text{newID}(i') : i' \in \mathcal{F}\}$ 

```

and *inactive* otherwise. Let $\mathcal{A} \subseteq [K]$ denote the set of cost types corresponding to active constraints. Note that replacing $\alpha_k N/2$ with any constant and strict fraction of $\alpha_k N$ will not change the results.

Based on the costs $C_{k,i}^*$'s and the active constraints, the ID reassignment algorithm rearranges arms so that the cost incurred by each contiguous segment of arms is approximately proportional to the length of the segment. We give a brief explanation of how the reassigned IDs affect the execution of the policy at the end of this section. The algorithm is formally described in Algorithm 11, with more details and properties provided in Appendix 6.D. In the remainder of this chapter, we use the *reassigned IDs* to refer to arms, i.e., arm i refers to the arm whose new ID assigned by the ID reassignment algorithm is i .

Algorithm 12 ID policy with reassignment

-
- 1: **Input:** N -armed WCMDP instance $(\mathcal{M}_i)_{i \in [N]}$
 - 2: **Preprocessing:**
 - 3: Solve the LP in (6.5) and obtain the optimal state-action frequencies $(y_i^*(s, a))_{i \in [N], s \in \mathbb{S}, a \in \mathbb{A}}$
 - 4: Calculate the optimal single-armed policies $(\bar{\pi}_i^*)_{i \in [N]}$ using (6.7)
 - 5: Perform ID reassignment using Algorithm 11
 - 6: **Real-time:**
 - 7: **for** $t = 0, 1, 2, \dots$ **do**
 - 8: Sample ideal actions $\hat{A}_{i,t} \sim \bar{\pi}_i^*(\cdot \mid S_{i,t})$ for all $i \in [N]$
 - 9: $I \leftarrow 1$
 - 10: **while** $\sum_{i \in [I]} c_{k,i}(S_{i,t}, \hat{A}_{i,t}) \leq \alpha_k N, \forall k \in [K]$ **do**
 - 11: For arm I , take action $A_{I,t} = \hat{A}_{I,t}$; $I \leftarrow I + 1$
 - 12: For each arm $i \in \{I, I + 1, \dots, N\}$, take action $A_{i,t} = 0$
-

Constructing ID policy. We are now ready to describe our generalized ID policy, formally described in Algorithm 12. The policy begins with a one-time preprocessing phase: we solve the associated LP, construct the optimal single-armed policies, and reassign arm IDs using the ID reassignment algorithm (Algorithm 11). After the preprocessing, the policy proceeds at each time step t as follows. For each arm i (where i is the reassigned ID), we first sample an action $\hat{A}_{i,t}$, referred to as an *ideal action*, from the optimal single-armed policy $\bar{\pi}_i^*(\cdot \mid S_{i,t})$. We then attempt to execute these ideal actions, i.e., set the real actions equal to the ideal actions, in ascending order of arm IDs, starting from $i = 1$, then $i = 2$, and so on. We continue the attempt until we have used up at least one type of cost budget, at which point we let the remaining arms take action 0 (the no-cost action).

This ID policy is a natural generalization of the ID policy designed for homogeneous restless bandits in Chapter 4 [153]. At a high level, using arm IDs to decide the priority order for executing ideal actions guarantees that a subset of arms (those with smaller IDs) can persistently follow their ideal actions. This persistency gives these arms time to converge to their optimal state-action frequencies, which in turn allows their instantaneous costs to converge to steady-state values. This convergence creates slack in the budget constraints, thereby allowing more arms to follow their ideal actions. In contrast, if we do not use IDs but instead randomly select a subset of

arms to follow their ideal actions, the convergence may be disrupted. Indeed, Chapter 3 [149] provides an example where this randomized strategy fails to achieve asymptotic optimality in restless bandits.

When there is no ambiguity, we refer to the ID policy with reassignment simply as the ID policy.

How the reassigned IDs affect the policy execution. In the heterogeneous setting, arms differ in their cost consumptions under their optimal single-armed policies. The ID reassignment algorithm is designed to prevent “plateaus” in cumulative cost as we progress from smaller to larger IDs. If such a plateau exists, the subset of arms allowed to follow their ideal actions (determined by the budget constraints) can become sensitive to the randomness in action sampling, potentially leading to performance instability. The ID reassignment algorithm ensures a regularity property of the cost consumptions, stated in Lemma 6.7 in Appendix 6.D, which eliminates such plateaus.

6.4 Main results and technical overview

Before we present the main results, we first state our main assumption. This assumption is for the optimal single-armed policies $\bar{\pi}_i^*$'s. Note that each $\bar{\pi}_i^*$ is a stationary Markov policy. Therefore, under this policy, the state of arm i forms a Markov chain. Let the transition probability matrix of this Markov chain be denoted as $P_i = (P_i(s, s'))_{s \in \mathbb{S}, s' \in \mathbb{S}}$, where the row index is the current state s and the column index is the next state s' . Then $P_i(s, s')$ can be written as

$$P_i(s, s') = \sum_{a \in \mathbb{A}} \mathbb{P}_i(s' \mid s, a) \bar{\pi}_i^*(a \mid s). \quad (6.9)$$

One can verify that the stationary distribution of this Markov chain is $\mu_i^* = (\mu_i^*(s))_{s \in \mathbb{S}}$ with $\mu_i^*(s) = \sum_{a \in \mathbb{A}} y_i^*(s, a)$, which we refer to as the *optimal state distribution* for arm i . Let τ_i be the *mixing time* of this Markov chain, defined as

$$\tau_i = \max_{s \in \mathbb{S}} \min \{t \in \mathbb{N}: \|P_i^t(s, \cdot) - \mu_i^*(\cdot)\|_1 \leq 1/e\}, \quad (6.10)$$

where P_i^t is the t -step transition probability matrix. The mixing time τ_i is finite if the Markov chain P_i is unichain (one recurrent class, possibly with transient states) and aperiodic.

Assumption 6.2. For each arm $i \in \mathbb{N}_+$, the induced Markov chain under the optimal single-armed policy $\bar{\pi}_i^*$ is an aperiodic unichain. Furthermore, the mixing times of these Markov chains have a uniform upper bound; i.e., there exists a positive τ such that for all $i \in \mathbb{N}_+$,

$$\tau_i \leq \tau. \quad (6.11)$$

We remark that in the homogeneous or typed heterogeneous settings, once we make the aperiodic unichain assumption in Assumption 6.2, the uniform upper bound on mixing times automatically exists.

Next, we state our main theorem, Theorem 6.3, whose proof is provided in Appendix 6.E.

Theorem 6.3. *Consider an N -armed WCMDP problem satisfying Assumption 6.2, with initial system state $\mathbf{S}_0$. Let policy π be the ID policy with reassignment (Algorithm 12). Then the optimality gap of π is bounded as*

$$R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0) \leq \frac{C_{\text{ID}}}{\sqrt{N}},$$

where C_{ID} is a positive constant independent of N .

We re-emphasize that our proposed ID policy with reassignment is the first efficiently computable policy that achieves an $O(1/\sqrt{N})$ optimality gap for fully heterogeneous average-reward WCMDPs. In contrast, the best-known optimality gap for efficiently computable policies for average-reward WCMDPs is $o(1)$, achieved only under restrictive budget constraints and typed-heterogeneity.

We comment that the primary goal of this chapter is to characterize the optimality gap in terms of its order in N , which is in line with the main focus of the large body of prior work on restless bandits and WCMDPs. While our analysis also gives an explicit expression for the constant C_{ID} , which shows that $C_{\text{ID}} = O(K^5 \max\{r_{\max}, c_{\max}\}^7 \tau^4 / \alpha_{\min}^6)$, we have not attempted to optimize its dependence on other problem parameters, either through refined analysis or alternative policy design.

Remark 6.4 (Generalization of result). Our result can be generalized to the setting where a $g(N)$ fraction of arms have unbounded mixing times, and the mixing times of the remaining arms scale with N with an upper bound τ . In this case, we can modify the ID policy by reassigning this $g(N)$ fraction of arms the largest IDs, effectively ignoring these arms. Applying Theorem 6.3 to the remaining arms then implies $R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0) \leq$

$O(K^5 \max\{r_{\max}, c_{\max}\}^7 \tau^4 / (\alpha_{\min}^6 \sqrt{N}) + r_{\max} g(N))$. Consequently, this modified policy is asymptotically optimal when $g(N) = o(1)$ and $\tau = o(N^{1/8})$.

Adapting the proof of Theorem 6.3 also gives the following finite-time bound (see Appendix 6.I).

Proposition 6.5 (Finite-time bound). *Under the same conditions as Theorem 6.3, we have for any $T \geq 1$,*

$$R^*(N, \mathbf{S}_0) - \frac{1}{TN} \sum_{t=0}^{T-1} \sum_{i \in [N]} \mathbb{E} [r_i(S_{i,t}^\pi, A_{i,t}^\pi)] \leq \frac{C_{\text{ID}}}{\sqrt{N}} + \frac{C_{\text{finite}}}{T}, \quad (6.12)$$

where C_{finite} is another positive constant independent of N .

Technical overview

Our technical approach uses the Lyapunov drift method, which has found widespread applications in queueing systems, Markov decision processes, reinforcement learning, and so on. While the basic framework of the drift method is standard, the *key challenge* lies in constructing the right Lyapunov function with the desired properties, where the difficulty is exacerbated by the full heterogeneity of the problem under study. Our construction of such a Lyapunov function is highly novel, yet still natural. We reiterate that fully heterogeneous, high-dimensional stochastic systems are poorly understood in the existing literature. Our approach opens up the possibility of analyzing the steady-state behavior of such systems through the Lyapunov drift method.

In the remainder of this section, we consider the ID policy, also referred to as the policy π . Let $\mathbf{X}_t$ denote the system state under it, with the superscript π omitted for brevity. To make this overview more intuitive, here let us assume that $\mathbf{X}_t$ converges to its steady state $\mathbf{X}_\infty$ in a proper sense such that taking expectations in steady state is the same as taking time averages. However, note that our formal results do not need this assumption and directly work with time averages. We call a function V a Lyapunov/potential function if it maps each possible system state to a nonnegative real number.

General framework of the drift method. Here we briefly describe the general framework of the drift method when applied to our problem. The goal is to construct a Lyapunov function V such that

$$(\mathbf{C1}) \quad R_N^{\text{rel}} - R(\pi, \mathbf{S}_0) \leq C_1 \mathbb{E}[V(\mathbf{X}_\infty)] / N + O(1/\sqrt{N}) \text{ for some constant } C_1;$$

(C2) (Drift condition) $\mathbb{E}[V(\mathbf{X}_{t+1}) \mid \mathbf{X}_t] - V(\mathbf{X}_t) \leq -C_2 V(\mathbf{X}_t) + O(\sqrt{N})$
for a constant C_2 .

The drift condition requires that on average, the value of V approximately decreases (ignoring the additive $O(\sqrt{N})$) after a time step. The drift condition implies a bound on $\mathbb{E}[V(\mathbf{X}_\infty)]$. To see this, let $\mathbf{X}_t$ follow the steady-state distribution, which means $\mathbf{X}_{t+1}$ also follows the steady-state distribution, and take expectations on both sides of the inequality. Then we get $0 = \mathbb{E}[V(\mathbf{X}_{t+1})] - \mathbb{E}[V(\mathbf{X}_t)] \leq -C_2 \mathbb{E}[V(\mathbf{X}_t)] + O(\sqrt{N})$, which implies $\mathbb{E}[V(\mathbf{X}_\infty)] = \mathbb{E}[V(\mathbf{X}_t)] = O(\sqrt{N})$. Combining this with (C1) proves the desired $O(1/\sqrt{N})$ upper bound on the optimality gap.

Key challenge: constructing Lyapunov function. We highlight this challenge by contrasting the homogeneous setting and the heterogeneous setting. In the *homogeneous* setting, there is only one optimal state distribution, μ^* . The Lyapunov function in Chapter 4 [153] is defined based on the distance between the *empirical state distribution* across arms and μ^* . Specifically, it is based on a set of functions $(h(\mathbf{X}_t, D))_{D \subseteq [N]}$ defined as:

$$h(\mathbf{X}_t, D) = \|\mathbf{X}_t(D) - m(D)\mu^*\|, \quad (6.13)$$

where $\mathbf{X}_t(D) = (X_t(D, s))_{s \in \mathcal{S}}$ denotes within D , the number of arms in each state s , divided by N ; $m(D) = |D|/N$; and the norm $\|\cdot\|$ is a properly defined norm. The idea is that if all arms in D follow the optimal single-armed policy, the state distribution of each arm in D gets closer to μ^* , and thus $\mathbf{X}_t(D)$ gets closer to $m(D)\mu^*$ over time.

In the *heterogeneous* setting, we also want to construct a Lyapunov function $h(\mathbf{X}_t, D)$ to witness the convergence of any set of arms D if they follow the optimal single-armed policies. However, unlike the homogeneous setting, now it no longer makes sense to aggregate arm states into an empirical state distribution, since each arm's dynamics is distinct. Instead, our Lyapunov function considers $X_{i,t} - \mu_i^*$, where recall $X_{i,t}(s)$ is the indicator that arm i 's state is s at time t . A naive first attempt is to construct the Lyapunov function from the pointwise distances, $\|X_{i,t} - \mu_i^*\|$ for each arm i , with a properly defined norm $\|\cdot\|$. However, the pointwise distances are very noisy: $\|X_{i,t} - \mu_i^*\|$ could be large even when the state of arm i independently follows the distribution μ_i^* for each i , a situation when we should view the set of arms as already converged.

Intuitively, to make the Lyapunov function properly reflect the convergence of the set of arms (referred to as “the system” in the rest of the section)

following the optimal single-armed policies, we would like it to depend less strongly on the state of each individual arm and focus more on the collective properties of the whole system. Our idea is to *project* the system state onto a properly selected set of *feature vectors*, and construct the Lyapunov function based on how far these projections are from the projections of the optimal state distributions $(\mu_i^*)_{i \in [N]}$. Then what features of the system state do we need to determine whether it has converged in a proper sense? The first feature we consider is the instantaneous reward of the system, $\sum_{i \in D} \langle X_{i,t}, r_i^* \rangle$, where $r_i^* \in \mathbb{R}^S$ is the reward function of arm i under $\bar{\pi}_i^*$, and recall that the inner product is defined between two vectors whose entries correspond to states in $\mathbb{S}$. We also want to keep track of the ℓ -step ahead expected reward, $\sum_{i \in D} \langle X_{i,t} P_i^\ell, r_i^* \rangle$, for each $\ell \in \mathbb{N}_+$. Intuitively, if $\sum_{i \in D} \langle (X_{i,t} - \mu_i^*) P_i^\ell, r_i^* \rangle$ is small for each $\ell \in \mathbb{N}$, the reward of the system should remain close to that under the optimal state distributions $(\mu_i^*)_{i \in [N]}$ for a long time; conversely, if the state of each arm i independently follows μ_i^* , each of these features should be small as well. We also consider the ℓ -step ahead expected type- k cost for each $\ell \in \mathbb{N}$ and $k \in [K]$ as features, defined analogously.

Combining the above ideas, for any set of arms D , we let the Lyapunov function $h(\mathbf{X}_t, D)$ be the supremum of the differences between $\mathbf{X}_t$ and μ^* in all the features directions defined above, under proper weightings:

$$h(\mathbf{X}_t, D) = \max_{g \in \mathcal{G}} \sup_{\ell \in \mathbb{N}} \left| \sum_{i \in D} \langle (X_{i,t} - \mu_i^*) P_i^\ell / \gamma^\ell, g_i \rangle \right|, \quad (6.14)$$

where $\gamma = \exp(-1/(2\tau))$ for τ defined in Assumption 6.2; each element $g \in \mathcal{G}$ is either $g = (r_i^*)_{i \in [N]}$, or corresponds to the type- k cost for some $k \in [K]$ (See Appendix 6.E for the definition of $\mathcal{G}$). Note that *dividing* each term by powers of γ is another interesting trick, which induces a negative drift in $h(\mathbf{X}_t, D)$ under the optimal single-armed policies, as demonstrated in the proof of Lemma 6.8.

Now with the set of functions $(h(\mathbf{X}_t, D))_{D \subseteq [N]}$ defined, we generalize the idea of focus sets in Chapter 4 [153] to convert $(h(\mathbf{X}_t, D))_{D \subseteq [N]}$ into a Lyapunov function $V(\mathbf{X}_t)$. We prove that V satisfies (C1) and (C2) using the structure of $(h(\mathbf{X}_t, D))_{D \subseteq [N]}$.

Remark 6.6. The idea for constructing $h(\mathbf{X}_t, D)$ is potentially useful for analyzing other heterogeneous stochastic systems. At a high level, projecting the system state onto a set of feature vectors (and their future expectations) can be roughly viewed as aggregating system states whose relevant performance

metrics *remain close for a sufficiently long time*. This idea provides a new way to measure the distance between two system states in a heterogeneous system, and this distance notation enjoys similar properties as that in a homogeneous system, without resorting to symmetry.

6.5 Experiments

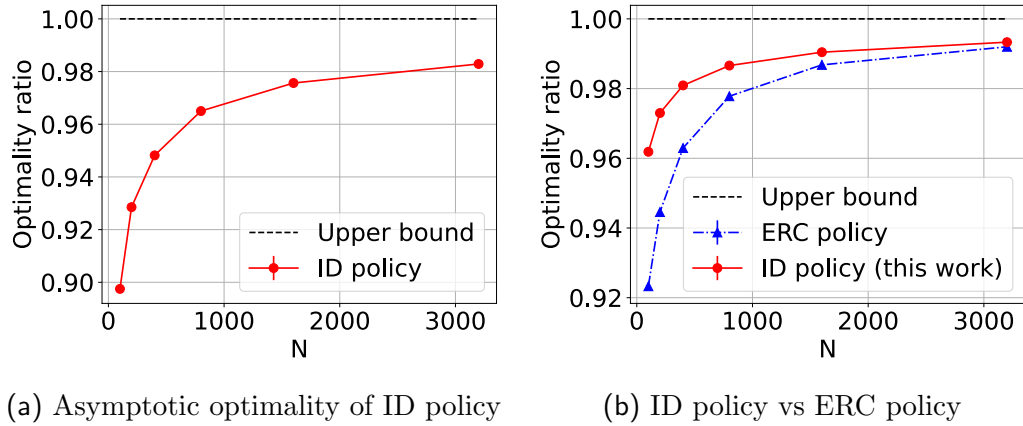


Figure 6.1: Numerical performance of the ID policy for fully heterogeneous WCMDPs.

In this section, we perform two sets of experiments to illustrate the numerical performance of the proposed ID policy for fully heterogeneous WCMDPs.

In the first set of experiments, we demonstrate the asymptotic optimality of the ID policy. We increase the number of arms as $N \in \{100, 200, 400, 800, 1600, 3200\}$. Each arm's MDP has 10 states and 4 actions, with parameters generated uniformly at random in a proper sense. The N -armed problem has 4 budget constraints, with cost functions also generated randomly. More details are provided in Appendix 6.B.1. We simulate the policy for 2×10^4 time steps over 4 replications for each N . To illustrate the performance more clearly, we measure the *optimality ratio*, defined as the ratio between the long-run average reward achieved by a policy and the LP relaxation upper bound R_N^{rel} . Confidence intervals are calculated using the batch means method with a batch size of 4000, but

they are typically too small to be visible on figures. Figure 6.1a shows that the optimality ratio of the ID policy becomes increasingly close to 1 as N increases.

In the second set of experiments, we compare the ID policy with the ERC policy proposed in [293]. As discussed in Section 6.1, only a few prior papers address heterogeneous WCMDPs. Among them, [293] considers the most general setting, but is still limited to a single-budget constraint, state-independent costs, and typed heterogeneity. To make a fair comparison, we evaluate both policies under this special case, while keeping all other settings the same as in the first set of experiments. ID policy turns out to have a slight improvement over ERC policy in some instances, one of which is shown in Figure 6.1b, and has comparable performance in others. Importantly, unlike ERC, the ID policy applies to far more general classes of WCMDPs.

6.A Detailed review on related work

In this section, we provide a more detailed, though still non-exhaustive, review of the literature. We mainly focus on theoretical work with formal performance guarantees, leaving out the extensive body of work with empirical results. We begin by surveying papers with the same reward criterion as ours, i.e., infinite-horizon average-reward criterion. Next, we turn to work that considers other reward criteria: the finite-horizon total-reward criterion and the infinite-horizon discounted-reward criterion.

Infinite-horizon average-reward WCMDPs. Work on average-reward WCMDPs remains relatively scarce, and to our knowledge, fully heterogeneous WCMDPs have yet to be addressed. Compared to RBs, a WCMDP allows multiple actions for each arm and multiple cost constraints, where each cost function is a function of both the state and the action of an arm. The line of research [121, 144, 266, 293] has generalized the action space and cost model of RBs to WCMDPs, and some of them allow for typed heterogeneity. In particular, Hodge and Glazebrook [144] generalize the Whittle index policy to homogeneous WCMDPs with a single constraint and multiple actions, where each action represents a different activation level and has a different cost. Verloop [266] extends the LP-Priority policies to typed heterogeneous WCMDPs with a single constraint and multiple actions, but requires each

action to have the same cost. Then Xiong et al. [293] propose another index policy for typed heterogeneous WCMDPs with a single constraint, and allow each action to have a different cost. In the three papers above [144, 266, 293], $o(1)$ optimality gaps have been proved under a similar set of assumptions as in most restless bandit papers, i.e., aperiodic-unichain (or irreducibility) condition, global attractor property, and a generalized indexability condition if the policy is Whittle index. Finally, Goldsztajn and Avrachenkov [121] consider homogeneous WCMDPs with multiple actions and multiple cost constraints with general cost functions, and propose a class of policies with $o(1)$ optimality gaps under a weaker-than-standard aperiodic-unichain condition.

Finite-horizon total-reward WCMDPs. Prior work on RBs has been discussed in Appendix 3.A. Since then, researchers have established asymptotic optimality in more general settings [56, 58, 87, 111, 296]. Among these papers, the most general setting is addressed by Brown and Zhang [58], where the authors consider fully heterogeneous WCMDPs; they obtain $O(1/\sqrt{N})$ optimality gaps for a naive fluid policy and a reoptimization-based fluid policy, among a few other results to be reviewed in the next paragraphs. Notably, there is also a further generalization of fully heterogeneous WCMDPs, which involves an exogenous state that affects all arms' transitions, rewards, and constraints; Brown and Zhang [57] propose this setting, where they achieve an $O(1/\sqrt{N})$ optimality gap using a dynamic fluid policy.

Similar to RBs, there has been a line of work on improving the optimality gap beyond the order $O(1/\sqrt{N})$ for WCMDPs by making an additional assumption called non-degeneracy. Gast et al. [109] and Brown and Zhang [58] establish $O(1/N)$ optimality gaps for homogeneous and typed heterogeneous WCMDPs, respectively, assuming non-degeneracy. More recently, Zhang [297] proposes a policy for fully heterogeneous WCMDPs; the optimality gap bound of the policy interpolates between $O(1/\sqrt{N})$ and $O(1/N)$ as the degree of non-degeneracy varies, unifying the performance bounds in the degenerate and non-degenerate cases.

Despite the generality of the settings and the fast diminishing rate of the optimality gaps as $N \rightarrow \infty$, most of the optimality gaps in the finite-horizon setting depend *super-linearly* on the time horizon, so they do not carry over to the infinite-horizon average-reward setting. There are two exceptions, [58, 109], which achieve optimality gaps that depend linearly on the time horizon under some special conditions: [58] requires all entries of the transition

kernels to be bounded away from zero; [109] assumes an ergodicity property, which requires two arms in any different states to synchronize in a fixed number of steps under *any sequence of actions* with a positive probability. However, without these conditions, the optimality gaps in [58, 109] depend quadratically on the time horizon. Apart from having distinct optimality gap bounds, all existing algorithms in the finite-horizon setting need to (sometimes repeatedly) solve LPs whose number of variables scales with the time horizon, so they cannot be directly adapted to the infinite-horizon average-reward setting.

Infinite-horizon discounted-reward WCMDPs. Asymptotic optimality has also been established for RBs and WCMDPs under the infinite-horizon discounted-reward criterion for homogeneous and heterogeneous settings. In addition to the prior work discussed in Appendix 3.A on homogeneous RBs, here we review prior work on RBs in the heterogeneous setting and WCMDPs in both settings. In particular, Brown and Smith [56] establish an $O(N^{\log_2(\sqrt{\gamma})})$ optimality gap for fully heterogeneous WCMDPs when $\gamma \in (1/2, 1)$. Subsequently, Ghosh et al. [111], Zhang and Frazier [301] obtain $O(1/\sqrt{N})$ optimality gaps for homogeneous and typed heterogeneous RBs, and Brown and Zhang [58] establish the same order of optimality gap for fully heterogeneous WCMDPs. Similar to the finite-horizon setting, most of these optimality gaps depend super-linearly on the effective time horizon $1/(1 - \gamma)$ unless special conditions hold [58], so they do not carry over to the infinite-horizon average-reward setting. The policies here also require solving LPs whose complexities scale with the effective time horizon.

6.B Experimental details

In this appendix, we provide details of the two WCMDP instances considered in Section 6.5, and the definition of the baseline policy, the ERC policy from [293]. The complete code for these experiments is available on GitHub [298], and all results can be reproduced within 24 hours on a standard PC (e.g., 6-Core Intel Core i7).

6.B.1 WCMDP instance generation

Details of the WCMDP instance 1 (In Figure 6.1a). $|\mathbb{S}| = 10$, $|\mathbb{A}| = 4$, $K = 4$. For each $i \in [N]$, $s \in \mathbb{S}$, and $a \in \mathbb{A}$, $r_i(s, 0) = 0$, and $r_i(s, a)$ is independently sampled from the uniform distribution over $[0, 1]$ for each $a \neq 0$; $P_i(s, a, \cdot)$ is independently sampled from the uniform distribution over the probability simplex. As for the cost function, for each $i \in [N]$, $s \in \mathbb{S}$, $a \in \mathbb{A}$, and $k \in [K]$, we have $c_{k,i}(s, 0) = 0$, and $c_{k,i}(s, a)$ is independently sampled from the uniform distribution over $[0, 1]$ for each $a \neq 0$. For each $k \in [K]$, α_k is uniformly sampled from $\{0.05, 0.1, 0.15, \dots, 0.45\}$, i.e., the uniform distribution over integer multiples of 0.05 in the interval $(0, 0.5)$.

Details of the WCMDP instance 2 (In Figure 6.1b). This instance is typed heterogeneous with 10 types, with equal fraction of arms in each type. Each arm has $|\mathbb{S}| = 10$ states, $|\mathbb{A}| = 4$ actions. For each type, the reward function and transition kernel are generated in the same way as each arm in the first instance. The budget is also sampled from the same distribution as the first instance. There is a single budget constraint. The cost function depends only on the action, consistent with the setting in [293]; the cost of each of action $a \in \mathbb{A}$ is sampled from the uniform distribution over $[0, 1]$.

6.B.2 Definition of ERC policy

The ERC policy [293] solves the same LP in (6.5) to define the single-armed policies $\bar{\pi}_i^*(a|s)$ as in (6.7). At each time step, the policy computes an index for each arm-state pair $(i, s) \in [N] \times \mathbb{S}$:

$$\mathcal{I}(i, s) = \sum_{a \in \mathbb{A}} \bar{\pi}_i^*(a|s) r_i(s, a).$$

It then iterates through the arms in descending order of these indices. For each arm, it samples an action $a \sim \bar{\pi}_i^*(\cdot | s)$ and applies it only if the budget is sufficient. If not, it defaults to action 0.

6.C Proving the LP relaxation

In this section, we prove Lemma 6.1, which shows that the linear program in (6.5) is a relaxation of the WCMDP problem. Lemma 6.1 is restated as follows.

Lemma 6.1. *The optimal value of any N -armed WCMDP problem is upper bounded by the optimal value of the corresponding linear program in (6.5), i.e.,*

$$R^*(N, \mathbf{S}_0) \leq R_N^{\text{rel}}, \quad \forall N, \forall \mathbf{S}_0.$$

Proof. To upper bound the optimal reward of the WCMDP, $R^*(N, \mathbf{S}_0)$, we observe that standard MDP theory ensures that a stationary Markovian policy achieves the optimal reward, as the WCMDP has finitely many system states and system actions [230, Theorem 9.18]. Therefore, it suffices to show that $R(\pi, \mathbf{S}_0) \leq R^{\text{rel}}$ for any stationary policy π and initial system state $\mathbf{S}_0$.

For any stationary policy π , consider the *state-action frequency under π* , given by

$$y_i^\pi(s, a) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\mathbf{1} \{S_{i,t}^\pi = s, A_{i,t}^\pi = a\}] , \quad \forall s \in \mathbb{S}, a \in \mathbb{A}, i \in [N].$$

where the limit is well-defined due to the stationarity of π . We argue that $y^\pi \triangleq (y_i^\pi(s, a))_{i \in [N], s \in \mathbb{S}, a \in \mathbb{A}}$ is a feasible solution to the LP relaxation in (6.5), with objective value being $R(\pi, \mathbf{S}_0)$. Then $R(\pi, \mathbf{S}_0) \leq R^{\text{rel}}$ follows from the optimality of R^{rel} .

To show that y^π satisfies the budget constraints of the LP relaxation (6.5b), we compute as follows: for any $s \in \mathbb{S}$, $a \in \mathbb{A}$ and constraint $k \in [K]$, we have

$$\begin{aligned} & \sum_{i \in [N]} c_{k,i}(s, a) y_i^\pi(s, a) \\ &= \sum_{i \in [N]} \sum_{s \in \mathbb{S}, a \in \mathbb{A}} c_{k,i}(s, a) \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\mathbf{1} \{S_{i,t}^\pi = s, A_{i,t}^\pi = a\}] \\ &= \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E} \left[\sum_{i \in [N]} \sum_{s \in \mathbb{S}, a \in \mathbb{A}} c_{k,i}(S_{i,t}^\pi, A_{i,t}^\pi) \right] \\ &\leq \alpha_k N, \end{aligned}$$

where the inequality follows from the fact that under a feasible N -armed policy π , $\sum_{i \in [N]} \sum_{s \in \mathbb{S}, a \in \mathbb{A}} c_{k,i}(S_{i,t}^\pi, A_{i,t}^\pi) \leq \alpha_k N$ for each budget constraint $k \in [K]$.

Then we verify that y^π satisfies the stationarity constraint of the LP relaxation (6.5c): for any state $s \in \mathbb{S}$ and arm $i \in [N]$, we have

$$\begin{aligned}
& \sum_{s' \in \mathbb{S}, a' \in \mathbb{A}} y_i^\pi(s', a') \mathbb{P}_i(s \mid s', a') \\
&= \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{s' \in \mathbb{S}, a' \in \mathbb{A}} P(S_{i,t}^\pi = s', A_{i,t}^\pi = a') \mathbb{P}_i(s \mid s', a') \\
&= \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} P(S_{i,t+1}^\pi = s) \\
&= \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T P(S_{i,t}^\pi = s) \\
&= \sum_{a \in \mathbb{A}} y_i^\pi(s, a).
\end{aligned}$$

We then argue that for each $i \in [N]$, $(y_i^\pi(s, a))_{s \in \mathbb{S}, a \in \mathbb{A}}$ is in the probability simplex of $\mathbb{S} \times \mathbb{A}$, as required by the last constraint in (6.5d), which is obvious: for any $i \in [N]$ and $s \in \mathbb{S}, a \in \mathbb{A}$, we have $y_i^\pi(s, a) \geq 0$; for any $i \in [N]$, we have

$$\sum_{s \in \mathbb{S}, a \in \mathbb{A}} y_i^\pi(s, a) = \lim_{T \rightarrow \infty} \sum_{t=0}^{T-1} \mathbb{E} \left[\sum_{s,a} \mathbf{1} \{S_{i,t}^\pi = s, A_{i,t}^\pi = a\} \right] = 1.$$

Therefore, y^π satisfies the constraints of the LP relaxation.

Finally, we show that the objective value of y^π equals $R(\pi, \mathbf{S}_0)$:

$$\begin{aligned}
& \frac{1}{N} \sum_{i \in [N]} \sum_{s \in \mathbb{S}, a \in \mathbb{A}} y_i^\pi(s, a) r_i(s, a) \\
&= \frac{1}{N} \sum_{i \in [N]} \sum_{s \in \mathbb{S}, a \in \mathbb{A}} r_i(s, a) \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\mathbf{1} \{S_{i,t}^\pi = s, A_{i,t}^\pi = a\}] \\
&= \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i \in [N]} \mathbb{E}[r_i(S_{i,t}^\pi, A_{i,t}^\pi)] \\
&= R(\pi, \mathbf{S}_0).
\end{aligned}$$

Because R^{rel} is the optimal value of the LP relaxation, we have $R^{\text{rel}} \geq R(\pi, \mathbf{S}_0)$ for any stationary policy π . Taking π to be the optimal policy finishes the proof. $\square$

6.D More details and properties of the ID reassignment algorithm

In this section, we provide more details and properties of the ID reassignment algorithm in Algorithm 11. We first explain the rationale of the algorithm at a high level, and then give a key property of it in Lemma 6.7.

Recall that for each arm $i \in [N]$ and each cost type $k \in [K]$, the expected cost under the optimal single-armed policy is defined as $C_{k,i}^* = \sum_{s \in \mathbb{S}, a \in \mathbb{A}} y_i^*(s, a) c_{k,i}(s, a)$. Based on $C_{k,i}^*$'s, we divide the cost constraints into *active* constraints and *inactive* constraints as follows. For each cost type $k \in [K]$, we say the type- k cost constraint is *active* if $\sum_{i \in [N]} C_{k,i}^* \geq \frac{\alpha_k}{2} N$, and *inactive* otherwise. Let $\mathcal{A} \subseteq [K]$ denote the set of cost types corresponding to active constraints.

Now consider a subset $D \subseteq [N]$ of arms. For each $k \in [K]$, let $C_k^*(D) = \sum_{i \in D} C_{k,i}^*$, i.e., $C_k^*(D)$ is the total expected type- k cost for arms in D under the optimal single-armed policies. In our analysis, we often need to consider a notion of *remaining* budget, defined as

$$\overline{C}_k^*(D) = \begin{cases} \alpha_k N - C_k^*(D), & \text{if } k \in \mathcal{A}, \\ \alpha_k N - C_k^*(D) - \frac{\alpha_k}{3} |D|, & \text{otherwise,} \end{cases} \quad (6.15)$$

where the $\frac{\alpha_k}{3} |D|$ is a correction term when type- k constraint is inactive. Note that the budget constraint (6.5b) of the LP relaxation implies that $C_k^*(D) \leq \alpha_k N$. Consequently, $\overline{C}_k^*(D) \geq 0$ and $C_k^*(D) + \overline{C}_k^*(D) \leq \alpha_k N$ for all $k \in [K]$ and all $D \subseteq [N]$.

The remaining budgets play a key role in the design of the ID reassignment algorithm. Roughly speaking, the goal of the ID reassignment is to ensure that when we expand a set of arms from $[n_1]$ to $[n_2]$ for some $n_1 \leq n_2$, the drop in the remaining budget of any type k , i.e., $\overline{C}_k^*([n_1]) - \overline{C}_k^*([n_2])$, is (almost) at least linear in $n_2 - n_1$. Note that this property is automatically satisfied for k if the type- k constraint is inactive. This property is formalized in Lemma 6.7, and the need for it will become clearer in Appendix 6.E when we introduce the so-called focus set in our analysis.

To achieve this desired property, we design our ID reassignment algorithm in the following way. If the set of active constraints is empty, i.e., $\mathcal{A} = \emptyset$, then there is no need to reassign the IDs. Otherwise, i.e., when $\mathcal{A} \neq \emptyset$, we first carefully choose two parameters, a positive real number δ and a positive integer d . We then divide the full ID set $[N]$ into groups of size d ,

i.e., $[d], [d+1 : 2d], [2d+1 : 3d], \dots, [(\lfloor N/d \rfloor - 1)d + 1 : \lfloor N/d \rfloor d]$, and the remainder. We ensure that after the reassignment, each group contains at least one arm i with $C_{k,i}^* \geq \delta$ for each active constraint type $k \in \mathcal{A}$.

The key here is to choose δ and d properly so such a reassignment is feasible. In particular, we choose $\delta = \alpha_{\min}/4$, where $\alpha_{\min} = \min_{k \in [K]} \alpha_k$, and let $d = \left\lceil \frac{(c_{\max} - \delta)K}{\alpha_{\min}/2 - \delta} \right\rceil$. Note that $d \geq K$ since one can verify that $c_{\max} \geq \alpha_{\min}/2$ when $\mathcal{A} \neq \emptyset$.

Lemma 6.7. *After applying the ID reassignment algorithm (Algorithm 11), for any n_1, n_2 with $1 \leq n_1 \leq n_2 \leq N$, we have*

$$\overline{C}_k^*([n_1]) - \overline{C}_k^*([n_2]) \geq \eta_c(n_2 - n_1) - M_c, \quad (6.16)$$

for all $k \in [K]$, where $\eta_c > 0$ and $M_c > 0$ are constants determined by $\alpha_{\min}$, $c_{\max}$, and K .

Further, let $\overline{C}^*(D) = \min_{k \in [K]} \overline{C}_k^*(D)$ for all $D \subseteq [N]$. Then the bound (6.16) implies that for any n_1, n_2 with $1 \leq n_1 \leq n_2 \leq N$,

$$\overline{C}^*([n_1]) - \overline{C}^*([n_2]) \geq \eta_c(n_2 - n_1) - M_c. \quad (6.17)$$

Proof. Our goal is to prove that for any n_1, n_2 with $1 \leq n_1 \leq n_2 \leq N$, we have

$$\overline{C}_k^*([n_1]) - \overline{C}_k^*([n_2]) \geq \eta_c(n_2 - n_1) - M_c, \quad (6.18)$$

for all $k \in [K]$, where

$$\eta_c = \min \left\{ \frac{\alpha_{\min}}{3}, \delta \cdot \left(\left\lceil \frac{(c_{\max} - \delta)K}{\alpha_{\min}/2 - \delta} \right\rceil \right)^{-1} \right\}, \quad M_c = 2\delta.$$

Case 1: $\mathcal{A} = \emptyset$. For any $k \in [K]$, by the definition of the remaining budget in (6.15),

$$\begin{aligned} & \overline{C}_k^*([n_1]) - \overline{C}_k^*([n_2]) \\ &= \left(\alpha_k N - C_k^*([n_1]) - \frac{\alpha_k}{3} n_1 \right) - \left(\alpha_k N - C_k^*([n_2]) - \frac{\alpha_k}{3} n_2 \right) \\ &= C_k^*([n_2]) - C_k^*([n_1]) + \frac{\alpha_k}{3} (n_2 - n_1) \\ &\geq \frac{\alpha_k}{3} (n_2 - n_1) \\ &\geq \eta_c (n_2 - n_1) - M_c. \end{aligned}$$

Case 2: $\mathcal{A} \neq \emptyset$. In this case, for any $k \notin \mathcal{A}$, following the same arguments as those in the previous paragraph, we again get $\overline{C}_k^*([n_1]) - \overline{C}_k^*([n_2]) \geq \eta_c(n_2 - n_1) - M_c$.

Now consider any $k \in \mathcal{A}$. Let $\mathcal{D}_k = \{i \in [N] : C_{k,i}^* \geq \delta\}$. We first show that

$$|\mathcal{D}_k| \geq \frac{(\alpha_k/2 - \delta)N}{c_{\max} - \delta}. \quad (6.19)$$

Note that

$$\sum_{i \in [N]} C_{k,i}^* \leq c_{\max} |\mathcal{D}_k| + \delta(N - |\mathcal{D}_k|). \quad (6.20)$$

Also since the type- k constraint is active,

$$\sum_{i \in [N]} C_{k,i}^* \geq \frac{\alpha_k}{2} N. \quad (6.21)$$

Combining these two inequalities and recalling that $\delta = \alpha_{\min}/4 < \alpha_k/2$, where $\alpha_k/2 \leq c_{\max}$ since $\alpha_k N/2 \leq \sum_{i \in [N]} C_{k,i}^* \leq c_{\max} N$, gives (6.19).

We next argue that $\mathcal{D}_k$ contains enough arms to ensure the ID reassignment steps from line 12 to line 16 in Algorithm 11 can be performed. Observe that for each $\ell = 0, 1, \dots, \lfloor N/d \rfloor - 1$, these steps remove at most K elements from each $\mathcal{D}_k$. To confirm $\mathcal{D}_k$ contains enough arms, note that

$$\begin{aligned} |\mathcal{D}_k| &\geq \frac{(\alpha_k/2 - \delta)N}{c_{\max} - \delta} \\ &\geq KN \cdot \frac{\alpha_{\min}/2 - \delta}{(c_{\max} - \delta)K} \\ &\geq KN \cdot \frac{1}{d} \\ &\geq K \lfloor N/d \rfloor, \end{aligned}$$

where we used the definition $d = \left\lceil \frac{(c_{\max} - \delta)K}{\alpha_{\min}/2 - \delta} \right\rceil$.

We are now ready to prove the inequality $\overline{C}_k^*([n_1]) - \overline{C}_k^*([n_2]) \geq \eta_c(n_2 - n_1) - M_c$ for any n_1, n_2 with $1 \leq n_1 \leq n_2 \leq N$. Consider the arms with new IDs in $[n_1 : n_2]$. Let g be the number of groups from groups of the form $\mathcal{I}(\ell) = [\ell d + 1 : (\ell + 1)d]$ with $\ell = 0, 1, \dots, \lfloor N/d \rfloor - 1$ that are completely contained in $[n_1 : n_2]$. Then it is easy to see

$$g \geq \frac{n_2 - n_1}{d} - 2.$$

Since Algorithm 11 ensures that $\sum_{i \in [N]} C_{k,i}^* \mathbf{1}\{\text{newID}(i) \in \mathcal{I}(\ell)\} \geq \delta$ for each ℓ , we know that

$$C_k^*([n_1 : n_2]) \geq \left(\frac{n_2 - n_1}{d} - 2 \right) \delta.$$

Therefore,

$$\begin{aligned} \overline{C}_k^*([n_1]) - \overline{C}_k^*([n_2]) &= \alpha_k N - C_k^*([n_1]) - (\alpha_k N - C_k^*([n_2])) \\ &= C_k^*([n_2]) - C_k^*([n_1]) \\ &= C_k^*([n_1 : n_2]) \\ &\geq \left(\frac{n_2 - n_1}{d} - 2 \right) \delta \\ &\geq \eta_c(n_2 - n_1) - M_c. \end{aligned}$$

For $\overline{C}^*(D) = \min_{k \in [K]} \overline{C}_k^*(D)$, it is straightforward to verify that

$$\begin{aligned} \overline{C}^*([n_1]) - \overline{C}^*([n_2]) &= \min_{k \in [K]} \overline{C}_k^*([n_1]) - \min_{k \in [K]} \overline{C}_k^*([n_2]) \\ &\geq \min_{k \in [K]} \left(\overline{C}_k^*([n_1]) - \overline{C}_k^*([n_2]) \right) \\ &\geq \eta_c(n_2 - n_1) - M_c, \end{aligned}$$

which completes the proof. $\square$

6.E Proof of main result (Theorem 6.3)

As outlined in the technical overview in Section 6.4, the core of our proof is the construction of a Lyapunov function. The Lyapunov function we construct is the following

$$V(\mathbf{x}) = h_{\text{ID}}(\mathbf{x}, m(\mathbf{x})) + L_h N \cdot (1 - m(\mathbf{x})). \quad (6.22)$$

In the rest of this section, we first define the functions $h_{\text{ID}}(\cdot, \cdot)$ and $m(\cdot)$, along with the constant L_h . We then proceed to analyze the Lyapunov function V to upper bound on the optimality gap.

Defining $h_{\text{ID}}(\cdot, \cdot)$ using subset Lyapunov functions. We first construct a Lyapunov function indexed by a subset of arms $D \subseteq [N]$, denoted as $h(\mathbf{x}, D)$, which is viewed as a function of the system state $\mathbf{x}$, and it is referred to as a *subset Lyapunov function*.

For each cost type $k \in [K]$, let $c_{k,i}^*(s) = \sum_{a \in \mathbb{A}} \bar{\pi}_i^*(a|s) c_{k,i}(s, a)$, and let $c_k^* = (c_{k,i}^*)_{i \in [N]}$ denote the vector of the functions $c_{k,i}^*$'s. In addition, let $r_i^*(s) = \sum_{a \in \mathbb{A}} \bar{\pi}_i^*(a|s) r_i(s, a)$, and let $r^* = (r_i^*)_{i \in [N]}$ denote the vector of the functions r_i^* 's. We combine these vectors into a set $\mathcal{G} = \{c_1^*, c_2^*, \dots, c_K^*, r^*\}$.

The subset Lyapunov function is then defined as

$$h(\mathbf{x}, D) = \max_{g \in \mathcal{G}} \sup_{\ell \in \mathbb{N}} \left| \sum_{i \in D} \langle (x_i - \mu_i^*) P_i^\ell / \gamma^\ell, g_i \rangle \right|. \quad (6.23)$$

Here recall that $\mathbf{x}$ is an $N \times |\mathbb{S}|$ matrix whose i th row, x_i , describes the state of arm i ; P_i is the transition probability matrix for arm i under the optimal single-armed policy; $\gamma = \exp(-1/(2\tau))$ for τ defined in Assumption 6.2. To build intuition for $h(\mathbf{x}, D)$, consider the term corresponding to $g = r^*$ and $\ell = 0$. In this case, the term $\sum_{i \in D} \langle x_i - \mu_i^*, g_i \rangle$ measures the difference between the reward obtained by applying the optimal single-armed policies to arms in D and the reward upper bound given by the LP relaxation. A similar interpretation holds for the differences in costs.

In Lemma 6.8 below, we show that $h(\mathbf{x}, D)$ is well-defined and establish its two key properties, which play a critical role in our analysis. The proof of Lemma 6.8 is given in Appendix 6.F.2.

Lemma 6.8. *The Lyapunov function $h(\mathbf{x}, D)$ defined in (6.23) is finite for all system state $\mathbf{x}$ and subset $D \subseteq [N]$. Moreover, $h(\mathbf{x}, D)$ has the following properties.*

1. **(Lipschitz continuity)** *There exists a Lipschitz constant L_h such that for each system state $\mathbf{x}$ and $D' \subseteq D \subseteq [N]$, we have*

$$|h(\mathbf{x}, D) - h(\mathbf{x}, D')| \leq L_h |D/D'|. \quad (6.24)$$

2. **(Drift condition)** *If each arm in D takes the action sampled from the optimal single-armed policy, i.e., $A_{i,t} \sim \bar{\pi}_i^*(\cdot | S_{i,t})$, then there exists a constant $C_h > 0$ such that*

$$\begin{aligned} & \mathbb{E} \left[(h(\mathbf{X}_{t+1}, D) - \gamma h(\mathbf{X}_t, D))^+ \mid \mathbf{X}_t, A_{i,t} \sim \bar{\pi}_i^*(\cdot | S_{i,t}), \forall i \in D \right] \\ & \leq C_h \sqrt{N}. \end{aligned} \quad (6.25)$$

Note that (6.25) implies the following more typical form of drift condition

$$\begin{aligned} & \mathbb{E}[h(\mathbf{X}_{t+1}, D) | \mathbf{X}_t, A_{i,t} \sim \bar{\pi}_i^*(S_{i,t}), \forall i \in D] - h(\mathbf{X}_t, D) \\ & \leq -(1 - \gamma)h(\mathbf{X}_t, D) + C_h\sqrt{N}. \end{aligned} \quad (6.26)$$

We are now ready to define the function $h_{\text{ID}}(\cdot, \cdot)$ used to construct the Lyapunov function V . For any system state $\mathbf{x}$ and $m \in [0, 1]_N$ (where recall that $[0, 1]_N$ is the set of integer multiples of $1/N$ within the interval $[0, 1]$), $h_{\text{ID}}(\mathbf{x}, m)$ is defined as

$$h_{\text{ID}}(\mathbf{x}, m) = \max_{m' \in [0, 1]_N : m' \leq m} h(\mathbf{x}, [Nm']). \quad (6.27)$$

That is, $h_{\text{ID}}(\mathbf{x}, m)$ is an upper envelope of the subset Lyapunov functions $h(\mathbf{x}, [Nm'])$'s. The function $h_{\text{ID}}(\mathbf{x}, m)$ has properties similar to those in Lemma 6.8, which we state as Lemma 6.12 and prove in Appendix 6.F.3.

Focus set. We next introduce the concept of the focus set, which is directly tied to the function $m(\cdot)$ in the Lyapunov function V . The focus set is a dynamic subset of arms based on the current system state. Specifically, for any system state $\mathbf{x}$, the focus set is defined as the set $[Nm(\mathbf{x})]$, where $m(\mathbf{x})$ is given by

$$m(\mathbf{x}) = \max \left\{ m \in [0, 1]_N : h_{\text{ID}}(\mathbf{x}, m) \leq \min_{k \in [K]} \bar{C}_k^*([Nm]) \right\}. \quad (6.28)$$

The focus set is introduced because it has several desirable properties that are useful for the analysis. First, under the ID policy, almost all the arms in the focus set, except for $O(\sqrt{N})$ arms, can follow the optimal single-armed policies. Additionally, as the focus set evolves over time, it is almost non-shrinking, and its size is closely related to the value of the function $h_{\text{ID}}(\cdot, \cdot)$. These properties are formalized as Lemmas 6.13, 6.14 and 6.15, which are presented in Appendix 6.G.

Bounding the optimality gap via analyzing the Lyapunov function V . With the Lyapunov function V fully defined, we now proceed to bound the optimality gap $R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0)$, where the policy π is the ID policy. As outlined in the technical overview in Section 6.4, an upper bound on the optimality gap is established via the following two lemmas.

Lemma 6.9. *Consider any N -armed WCMDP with initial system state $\mathbf{S}_0$ and assume that it satisfies Assumption 6.2. Let policy π be the ID policy. Consider the Lyapunov function V defined in (6.22). Then the optimality gap of π is bounded as*

$$R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0) \leq \frac{2r_{\max} + L_h}{L_h N} \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[V(\mathbf{X}_t)] + \frac{2r_{\max} K_{\text{conf}}}{\sqrt{N}},$$

where L_h is the Lipschitz constant in Lemma 6.8 and K_{conf} is the positive constant in Lemma 6.13.

Lemma 6.10. *Consider any N -armed WCMDP with initial system state $\mathbf{S}_0$ and assume that it satisfies Assumption 6.2. Let $\mathbf{X}_t$ be the system state at time t under the ID policy. Consider the Lyapunov function V defined in (6.22). Then its drift satisfies*

$$\mathbb{E}[V(\mathbf{X}_{t+1}) \mid \mathbf{X}_t] - V(\mathbf{X}_t) \leq -\rho_V V(\mathbf{X}_t) + K_V \sqrt{N}, \quad (6.29)$$

which further implies that

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[V(\mathbf{X}_t)] \leq \frac{K_V \sqrt{N}}{\rho_V}, \quad (6.30)$$

where ρ_V and K_V are constants whose values are given in the proof.

The proofs of Lemmas 6.9 and 6.10 are provided in Appendix 6.H. It is then straightforward to combine these two lemmas to get Theorem 6.3.

6.F Lemmas and proofs for subset Lyapunov functions

In this section, we prove lemmas for the subset Lyapunov functions $h(\mathbf{x}, D)$ and $h_{\text{ID}}(\mathbf{x}, m)$, defined in (6.23) and (6.27). In Appendix 6.F.1, we prove a preliminary lemma that helps us utilize Assumption 6.2. In Appendix 6.F.2, we prove Lemma 6.8, which addresses the properties of the subset Lyapunov functions $h(\mathbf{x}, D)$. Finally, in Appendix 6.F.3, we present and prove Lemma 6.12, which establishes properties of the function $h_{\text{ID}}(\mathbf{x}, m)$.

6.F.1 Lemma for utilizing mixing time bound of transition matrices

Recall that in Assumption 6.2, we assume that for each arm i , the mixing times of the transition matrix P_i under the optimal single-armed policy $\bar{\pi}^*$ is uniformly bounded by a constant τ . In this section, we prove a lemma that allows us to utilize this assumption.

Let Ξ_i denote a matrix whose rows are the same vector μ_i^* , where recall that μ_i^* is the stationary distribution of P_i . Lemma 6.11 stated below bounds the infinite series $\sum_{\ell=0}^{\infty} \max_{i \in [N]} \|(P_i - \Xi_i)^\ell\|_\infty \gamma^{-\ell}$ by a constant independent of N .

Lemma 6.11. *Suppose Assumption 6.2 holds, and let*

$$\gamma = \exp\left(-\frac{1}{2\tau}\right) \quad C_\tau = \frac{4e}{1 - 1/\sqrt{e}}\tau.$$

Then we have

$$\sum_{\ell=0}^{\infty} \max_{i \in [N]} \frac{\|(P_i - \Xi_i)^\ell\|_\infty}{\gamma^\ell} \leq C_\tau.$$

Here, recall that $\|A\|_\infty \triangleq \max_{s \in \mathbb{S}} \sum_{s' \in \mathbb{S}} |A(s, s')|$ for any matrix $A \in \mathbb{R}^{\mathbb{S} \times \mathbb{S}}$.

Note that Lemma 6.11 is the only place in our proofs where Assumption 6.2 is directly invoked; all of our later proofs utilize Assumption 6.2 indirectly by invoking Lemma 6.11.

Proof of Lemma 6.11. For each $\ell \in \mathbb{N}$, let $\ell_0 = \ell - \tau \lfloor \ell/\tau \rfloor$. By properties of the operator norm $\|\cdot\|_\infty$,

$$\|(P_i - \Xi_i)^\ell\|_\infty \leq \|(P_i - \Xi_i)^{\ell_0}\|_\infty \|(P_i - \Xi_i)^\tau\|_\infty^{\lfloor \ell/\tau \rfloor}. \quad (6.31)$$

Because each row of Ξ_i is μ_i^* , it is easy to verify that $(P_i - \Xi_i)^{\ell'} = P_i^{\ell'} - \Xi_i$ for any $\ell' \in \mathbb{N}$. Consequently,

$$\begin{aligned} \|(P_i - \Xi_i)^{\ell_0}\|_\infty &= \|P_i^{\ell_0} - \Xi_i\|_\infty \\ &= \max_{s \in \mathbb{S}} \sum_{s' \in \mathbb{S}} |P_i^{\ell_0}(s, s') - \mu_i^*(s')| \\ &\leq \max_{s \in \mathbb{S}} \sum_{s' \in \mathbb{S}} (P_i^{\ell_0}(s, s') + \mu_i^*(s')) \end{aligned}$$

$$= 2.$$

Moreover, by Assumption 6.2, it is not hard to see that

$$\max_{s \in \mathbb{S}} \|P_i^\tau(s, \cdot) - \mu_i^*(\cdot)\|_1 \leq \frac{1}{e},$$

so we have

$$\begin{aligned} \|(P_i - \Xi_i)^\tau\|_\infty &= \max_{s \in \mathbb{S}} \sum_{s' \in \mathbb{S}} |P_i^\tau(s, s') - \mu_i^*(s')| \\ &= \max_{s \in \mathbb{S}} \|P_i^\tau(s, \cdot) - \mu_i^*(\cdot)\|_1 \\ &\leq \frac{1}{e}. \end{aligned}$$

Substituting the bounds on $\|(P_i - \Xi_i)^\tau\|_\infty$ and $\|(P_i - \Xi_i)^{\ell_0}\|_\infty$ back to (6.31), we get

$$\begin{aligned} \|(P_i - \Xi_i)^\ell\|_\infty &\leq 2 \exp\left(-\left\lfloor \frac{\ell}{\tau} \right\rfloor\right) \\ &\leq 2 \exp\left(-\frac{\ell}{\tau} + 1\right). \end{aligned} \tag{6.32}$$

Therefore,

$$\begin{aligned} \sum_{\ell=0}^{\infty} \max_{i \in [N]} \frac{\|(P_i - \Xi_i)^\ell\|_\infty}{\gamma^\ell} &\leq \sum_{\ell=0}^{\infty} 2 \exp\left(-\frac{\ell}{\tau} + 1\right) \gamma^{-\ell} \\ &= \sum_{\ell=0}^{\infty} 2 \exp\left(-\frac{\ell}{2\tau} + 1\right) \\ &= \frac{2e}{1 - \exp(-1/(2\tau))}. \end{aligned} \tag{6.33}$$

Observe that $\tau \geq 1$, $0 < 1/(2\tau) \leq 1/2$, and $\exp(-a) \leq 1 - (1 - 1/\sqrt{e})a$ for any $0 < a \leq 1/2$, so

$$\frac{2e}{1 - \exp(-1/(2\tau))} \leq \frac{4e}{1 - 1/\sqrt{e}} \tau = C_\tau. \quad \square$$

6.F.2 Proof of Lemma 6.8

In this subsection, we prove Lemma 6.8, which is about properties of the Lyapunov function $h(\mathbf{x}, D)$:

$$h(\mathbf{x}, D) \triangleq \max_{g \in \mathcal{G}} \sup_{\ell \in \mathbb{N}} \left| \sum_{i \in D} \langle (x_i - \mu_i^*) P_i^\ell / \gamma^\ell, g_i \rangle \right|. \quad (6.23)$$

Lemma 6.8 is restated as follows.

Lemma 6.8. *The Lyapunov function $h(\mathbf{x}, D)$ defined in (6.23) is finite for all system state $\mathbf{x}$ and subset $D \subseteq [N]$. Moreover, $h(\mathbf{x}, D)$ has the following properties.*

1. (**Lipschitz continuity**) *There exists a Lipschitz constant L_h such that for each system state $\mathbf{x}$ and $D' \subseteq D \subseteq [N]$, we have*

$$|h(\mathbf{x}, D) - h(\mathbf{x}, D')| \leq L_h |D/D'|. \quad (6.24)$$

2. (**Drift condition**) *If each arm in D takes the action sampled from the optimal single-armed policy, i.e., $A_{i,t} \sim \bar{\pi}_i^*(\cdot \mid S_{i,t})$, then there exists a constant $C_h > 0$ such that*

$$\begin{aligned} & \mathbb{E} [(h(\mathbf{X}_{t+1}, D) - \gamma h(\mathbf{X}_t, D))^+ \mid \mathbf{X}_t, A_{i,t} \sim \bar{\pi}_i^*(\cdot \mid S_{i,t}), \forall i \in D] \\ & \leq C_h \sqrt{N}. \end{aligned} \quad (6.25)$$

In the proof of Lemma 6.8, we will frequently use the following equivalent form of $h(\mathbf{X}, D)$:

$$h(\mathbf{x}, D) = \max_{g \in \mathcal{G}} \sup_{\ell \in \mathbb{N}} \left| \sum_{i \in D} \langle (x_i - \mu_i^*) (P_i - \Xi_i)^\ell / \gamma^\ell, g_i \rangle \right|, \quad (6.34)$$

where Ξ_i is the matrix whose each row is the optimal state distribution of the i -th arm, μ_i^* . The equation (6.34) is equivalent to (6.23) because $(v_1 - v_2) P_i^\ell = (v_1 - v_2) (P_i - \Xi_i)^\ell$ for any $i \in [N]$, $\ell \in \mathbb{N}$, and row vectors $v_1, v_2 \in \Delta(\mathbb{S})$.

Proof. We organize the proof into three parts: we first show the finiteness of the subset Lyapunov function $h(\mathbf{x}, D)$; then, we prove the Lipschitz continuity of $h(\mathbf{x}, D)$ with respect to D (6.24); finally, we prove the drift condition for $h(\mathbf{x}, D)$ stated in (6.25).

Finiteness of $h(\mathbf{x}, D)$. To show that $h(\mathbf{x}, D)$ is finite for any system state $\mathbf{x}$ and subset $D \subseteq [N]$, we have for any $g \in \mathcal{G}$:

$$\begin{aligned}
\sup_{\ell \in \mathbb{N}} \left| \sum_{i \in D} \langle (x_i - \mu_i^*) P_i^\ell / \gamma^\ell, g_i \rangle \right| &\leq \sum_{i \in D} \sum_{\ell=0}^{\infty} |\langle (x_i - \mu_i^*) P_i^\ell / \gamma^\ell, g_i \rangle| \\
&= \sum_{i \in D} \sum_{\ell=0}^{\infty} |\langle (x_i - \mu_i^*) (P_i - \Xi_i)^\ell / \gamma^\ell, g_i \rangle| \quad (6.35) \\
&\leq \sum_{i \in D} \sum_{\ell=0}^{\infty} \|x_i - \mu_i^*\|_1 \frac{\|(P_i - \Xi_i)^\ell\|_\infty}{\gamma^\ell} \|g_i\|_\infty, \\
&\quad (6.36)
\end{aligned}$$

By Lemma 6.11, $\sum_{\ell=0}^{\infty} \|(P_i - \Xi_i)^\ell\|_\infty / \gamma^\ell$ is finite, so the expression in (6.36) is also finite. Taking maximum over $g \in \mathcal{G}$, because $\mathcal{G}$ is a finite set, we have

$$h(\mathbf{x}, D) = \max_{g \in \mathcal{G}} \sup_{\ell \in \mathbb{N}} \left| \sum_{i \in D} \langle (x_i - \mu_i^*) P_i^\ell / \gamma^\ell, g_i \rangle \right| < \infty.$$

Lipschitz continuity. For any system state $\mathbf{x}$ and subsets D, D' such that $D' \subseteq D \subseteq [N]$, we have

$$\begin{aligned}
&|h(\mathbf{x}, D) - h(\mathbf{x}, D')| \\
&= \left| \max_{g \in \mathcal{G}} \sup_{\ell \in \mathbb{N}} \left| \sum_{i \in D} \langle (x_i - \mu_i^*) P_i^\ell / \gamma^\ell, g_i \rangle \right| - \max_{g \in \mathcal{G}} \sup_{\ell \in \mathbb{N}} \left| \sum_{i \in D'} \langle (x_i - \mu_i^*) P_i^\ell / \gamma^\ell, g_i \rangle \right| \right| \\
&\leq \max_{g \in \mathcal{G}} \sup_{\ell \in \mathbb{N}} \left| \sum_{i \in D/D'} \langle (x_i - \mu_i^*) P_i^\ell / \gamma^\ell, g_i \rangle \right|. \quad (6.37)
\end{aligned}$$

Following similar arguments used to proving finiteness of $h(\mathbf{x}, D)$, we further bound the last expression as:

$$\begin{aligned}
&\max_{g \in \mathcal{G}} \sup_{\ell \in \mathbb{N}} \left| \sum_{i \in D/D'} \langle (x_i - \mu_i^*) P_i^\ell / \gamma^\ell, g_i \rangle \right| \\
&\leq \max_{g \in \mathcal{G}} \sum_{i \in D/D'} \|x_i - \mu_i^*\|_1 \|g_i\|_\infty \sum_{\ell=0}^{\infty} \frac{\|(P_i - \Xi_i)^\ell\|_\infty}{\gamma^\ell}
\end{aligned}$$

$$\leq 2 |D/D'| \max \{c_{\max}, r_{\max}\} C_{\tau}, \quad (6.38)$$

where in the last inequality, we utilize the facts that for any $g \in \mathcal{G}$ and $i \in [N]$, $\|g_i\|_{\infty} \leq \max \{c_{\max}, r_{\max}\}$, $\|x_i - \mu_i^*\|_1 \leq 2$, and $\sum_{\ell=0}^{\infty} \|(P_i - \Xi_i)^{\ell}\|_{\infty} / \gamma^{\ell} \leq C_{\tau}$. Therefore, $h(\mathbf{x}, D)$ is Lipschitz continuous in D with the Lipschitz constant $L_h = 2 \max \{c_{\max}, r_{\max}\} C_{\tau}$.

Drift condition. Next, we prove the drift condition in (6.25), which requires showing

$$\mathbb{E} [(h(\mathbf{X}_{t+1}, D) - \gamma h(\mathbf{X}_t, D))^+ | \mathbf{X}_t] = O(1/\sqrt{N}),$$

when the i -th arm follows the action generated by $\bar{\pi}_i^*$ for each $i \in D$. Because D is fixed in the rest of the proof, for simplicity, we use $h(\mathbf{x})$ as shorthand for $h(\mathbf{x}, D)$.

We first perform the following decomposition:

$$\begin{aligned} & \mathbb{E} [(h(\mathbf{X}_{t+1}) - \gamma h(\mathbf{X}_t))^+ | \mathbf{X}_t] \\ & \leq \mathbb{E} [|h(\mathbf{X}_{t+1}) - h(\mathbf{X}_t P)| | \mathbf{X}_t] + (h(\mathbf{X}_t P) - \gamma h(\mathbf{X}_t))^+, \end{aligned} \quad (6.39)$$

where $\mathbf{X}_t P \in \mathbb{R}^{N \times |\mathbb{S}|}$ denotes the matrix whose i -th row is given by $(\mathbf{X}_t P)_i \triangleq X_{i,t} P_i$, which is the state distribution of arm i after one-step of transition from $X_{i,t}$ under the policy $\bar{\pi}_i^*$. Next, we bound the two terms on the right-hand side of (6.39) separately.

We first bound the term $(h(\mathbf{X}_t P) - \gamma h(\mathbf{X}_t))^+$. Substituting $\mathbf{X}_t P$ into the definition of h , we have

$$\begin{aligned} h(\mathbf{X}_t P) &= \max_{g \in \mathcal{G}} \sup_{\ell \in \mathbb{N}} \left| \sum_{i \in D} \langle (X_{i,t} - \mu_i^*) P_i^{\ell+1} / \gamma^{\ell}, g_i \rangle \right| \\ &= \gamma \max_{g \in \mathcal{G}} \sup_{\ell \in \mathbb{N}: \ell \geq 1} \left| \sum_{i \in D} \langle (X_{i,t} - \mu_i^*) P_i^{\ell} / \gamma^{\ell}, g_i \rangle \right| \\ &\leq \gamma h(\mathbf{X}_t), \end{aligned} \quad (6.40)$$

so $(h(\mathbf{X}_t P) - \gamma h(\mathbf{X}_t))^+ = 0$ with probability 1.

Next, we bound the term $\mathbb{E} [|h(\mathbf{X}_{t+1}) - h(\mathbf{X}_t P)| | \mathbf{X}_t]$. Let $\epsilon_{i,t} \in \mathbb{R}^{\mathbb{S}}$ be the random vector given by $\epsilon_{i,t} = X_{i,t+1} - X_{i,t} P_i$ for $i \in D$. Then for each state $s \in \mathbb{S}$, $\epsilon_{i,t}(s)$ conditioned on $\mathbf{X}_t$ is a Bernoulli distribution with mean 0. Consequently,

$$|h(\mathbf{X}_{t+1}) - h(\mathbf{X}_t P)|$$

$$\begin{aligned}
&\leq \max_{g \in \mathcal{G}} \sup_{\ell \in \mathbb{N}} \left| \sum_{i \in D} \langle (X_{i,t+1} - \mu_i^*) P_i^\ell / \gamma^\ell, g_i \rangle \right| - \left| \sum_{i \in D} \langle (X_{i,t} P_i - \mu_i^*) P_i^\ell / \gamma^\ell, g_i \rangle \right| \\
&\leq \max_{g \in \mathcal{G}} \sup_{\ell \in \mathbb{N}} \left| \sum_{i \in D} \langle \epsilon_{i,t} P_i^\ell / \gamma^\ell, g_i \rangle \right| \\
&= \max_{g \in \mathcal{G}} \sup_{\ell \in \mathbb{N}} \left| \sum_{i \in D} \langle \epsilon_{i,t} (P_i - \Xi_i)^\ell / \gamma^\ell, g_i \rangle \right| \\
&= \max_{g \in \mathcal{G}} \sup_{\ell \in \mathbb{N}} \gamma^{-\ell} \left| \sum_{i \in D} \langle \epsilon_{i,t}, (P_i - \Xi_i)^\ell g_i \rangle \right|. \tag{6.41}
\end{aligned}$$

Taking the conditional expectation of $|h(\mathbf{X}_{t+1}) - h(\mathbf{X}_t P)|$ given $\mathbf{X}_t$, then (6.41) implies that

$$\begin{aligned}
\mathbb{E} [|h(\mathbf{X}_{t+1}) - h(\mathbf{X}_t P)| \mid \mathbf{X}_t] &\leq \mathbb{E} \left[\max_{g \in \mathcal{G}} \sup_{\ell \in \mathbb{N}} \gamma^{-\ell} \left| \sum_{i \in D} \langle \epsilon_{i,t}, (P_i - \Xi_i)^\ell g_i \rangle \right| \mid \mathbf{X}_t \right] \\
&\leq \sum_{g \in \mathcal{G}} \sum_{\ell=0}^{\infty} \gamma^{-\ell} \mathbb{E} \left[\left| \sum_{i \in D} \langle \epsilon_{i,t}, (P_i - \Xi_i)^\ell g_i \rangle \right| \mid \mathbf{X}_t \right]. \tag{6.42}
\end{aligned}$$

We bound the conditional expectation in (6.42) as follows: for any $\ell \in \mathbb{N}$ and $g \in \mathcal{G}$, we have

$$\begin{aligned}
&\mathbb{E} \left[\left| \sum_{i \in D} \langle \epsilon_{i,t}, (P_i - \Xi_i)^\ell g_i \rangle \right| \mid \mathbf{X}_t \right] \\
&\leq \sqrt{\mathbb{E} \left[\left(\sum_{i \in D} \langle \epsilon_{i,t}, (P_i - \Xi_i)^\ell g_i \rangle \right)^2 \mid \mathbf{X}_t \right]} \\
&= \sqrt{\sum_{i \in D} \mathbb{E} \left[\langle \epsilon_{i,t}, (P_i - \Xi_i)^\ell g_i \rangle^2 \mid \mathbf{X}_t \right]} \tag{6.43}
\end{aligned}$$

$$\begin{aligned}
&\leq \sqrt{\sum_{i \in D} \mathbb{E} \left[\left(\|\epsilon_{i,t}\|_1 \|(P_i - \Xi_i)^\ell\|_\infty \|g_i\|_\infty \right)^2 \mid \mathbf{X}_t \right]} \\
&\leq 2\sqrt{|D|} \max_{i \in D} \|g_i\|_\infty \|(P_i - \Xi_i)^\ell\|_\infty, \tag{6.44}
\end{aligned}$$

where (6.43) uses the facts that $\epsilon_{i,t}$ are independent across $i \in D$ and $\langle \epsilon_{i,t}, (P_i - \Xi_i)^\ell g_i \rangle$ has zero mean conditioned on $\mathbf{X}_t$; (6.44) is because $\|\epsilon_{i,t}\|_1 = \|X_{i,t+1} - X_{i,t}P_i\|_1 \leq 2$ for each $i \in D$.

Substituting the bound in (6.44) back to (6.42), we get

$$\begin{aligned}
& \mathbb{E} [|h(\mathbf{X}_{t+1}) - h(\mathbf{X}_t P)| \mid \mathbf{X}_t] \\
& \leq 2\sqrt{|D|} \sum_{\ell=0}^{\infty} \sum_{g \in \mathcal{G}} \gamma^{-\ell} \max_{i \in D} \|g_i\|_{\infty} \|(P_i - \Xi_i)^\ell\|_{\infty} \\
& = 2\sqrt{|D|} \sum_{\ell=0}^{\infty} \sum_{g \in \mathcal{G}} \max_{i \in D} \|g_i\|_{\infty} \cdot \max_{i \in D} \gamma^{-\ell} \|(P_i - \Xi_i)^\ell\|_{\infty} \\
& \leq 2\sqrt{|D|} (Kc_{\max} + r_{\max}) \sum_{\ell=0}^{\infty} \max_{i \in D} \frac{\|(P_i - \Xi_i)^\ell\|_{\infty}}{\gamma^{\ell}} \\
& \leq 2\sqrt{N} (Kc_{\max} + r_{\max}) C_{\tau}, \tag{6.45}
\end{aligned}$$

where the inequality in (6.45) is due to $\sum_{\ell=0}^{\infty} \max_{i \in D} \|(P_i - \Xi_i)^\ell\|_{\infty} / \gamma^{\ell} \leq C_{\tau}$ (Lemma 6.11).

Combining the above calculations, we get:

$$\begin{aligned}
& \mathbb{E} [(h(\mathbf{X}_{t+1}) - \gamma h(\mathbf{X}_t))^+ \mid \mathbf{X}_t] \\
& \leq \mathbb{E} [|h(\mathbf{X}_{t+1}) - h(\mathbf{X}_t P)| \mid \mathbf{X}_t] + (h(\mathbf{X}_t P) - \gamma h(\mathbf{X}_t))^+ \\
& \leq 2\sqrt{N} (Kc_{\max} + r_{\max}) C_{\tau}.
\end{aligned}$$

Therefore, $\mathbb{E} [(h(\mathbf{X}_{t+1}) - \gamma h(\mathbf{X}_t))^+ \mid \mathbf{X}_t] \leq C_h \sqrt{N}$ with $C_h = 2(Kc_{\max} + r_{\max})C_{\tau}$. $\square$

6.F.3 Properties of $h_{\text{ID}}(\cdot, \cdot)$

Lemma 6.12. *The Lyapunov function $h_{\text{ID}}(\mathbf{x}, m)$ defined in (6.27) has the following properties:*

1. (**Lipschitz continuity**) For each system state $\mathbf{x}$ and $m, m' \in [0, 1]_N$, we have

$$|h_{\text{ID}}(\mathbf{x}, m) - h_{\text{ID}}(\mathbf{x}, m')| \leq N L_h |m - m'|, \tag{6.46}$$

where $L_h > 0$ is the Lipschitz constant given in Lemma 6.8.

2. **(Drift condition)** For each $m \in [0, 1]_N$, if all arms in $[Nm]$ follow the optimal single-armed policies, we have:

$$\mathbb{E} \left[(h_{\text{ID}}(\mathbf{X}_{t+1}, m) - \gamma h_{\text{ID}}(\mathbf{X}_t, m))^+ \mid \mathbf{X}_t, A_{i,t} \sim \bar{\pi}_i^*(\cdot \mid S_{i,t}), \forall i \in [Nm] \right] \leq 2C_h \sqrt{N},$$

where $C_h > 0$ is the constant given in Lemma 6.8.

Proof. We first prove the Lipschitz continuity of $h_{\text{ID}}(\mathbf{x}, m)$ with respect to m . Because $h_{\text{ID}}(\mathbf{x}, m)$ is non-decreasing in m , it suffices to demonstrate that for any $m, m' \in [0, 1]_N$ such that $m > m'$,

$$h_{\text{ID}}(\mathbf{x}, m) - h_{\text{ID}}(\mathbf{x}, m') \leq NL_h(m - m'). \quad (6.47)$$

Denote $m_1 = \arg\max_{m_1 \in [0, 1]_N: m_1 \leq m} h(\mathbf{x}, [Nm_1])$. Then, by the definition of h_{ID} , we have $h_{\text{ID}}(\mathbf{x}, m) = h(\mathbf{x}, [Nm_1])$ and

$$h_{\text{ID}}(\mathbf{x}, m) - h_{\text{ID}}(\mathbf{x}, m') = h(\mathbf{x}, [Nm_1]) - h_{\text{ID}}(\mathbf{x}, m'). \quad (6.48)$$

If $m_1 \leq m'$, the right-hand side of (6.48) is non-positive, so (6.47) follows. If $m' < m_1 \leq m$, because $h_{\text{ID}}(\mathbf{x}, m') \geq h(\mathbf{x}, [Nm'])$, (6.48) implies that

$$\begin{aligned} h_{\text{ID}}(\mathbf{x}, m) - h_{\text{ID}}(\mathbf{x}, m') &\leq h(\mathbf{x}, [Nm_1]) - h(\mathbf{x}, [Nm']) \\ &\leq L_h(Nm_1 - Nm') \\ &\leq NL_h(m - m'), \end{aligned} \quad (6.49)$$

where (6.49) is due to the Lipschitz continuity of $h(x, D)$ with respect to D , as established in Lemma 6.8. We have thus proved (6.47).

Next, we prove the drift condition. We will assume $A_{i,t} \sim \bar{\pi}_i^*(\cdot \mid S_{i,t})$ for all $i \in [Nm]$ in the rest of the proof, without explicitly writing it in the conditional probabilities each time. We start by bounding the following expression:

$$\begin{aligned} &h_{\text{ID}}(\mathbf{X}_{t+1}, m) - \gamma h_{\text{ID}}(\mathbf{X}_t, m) \\ &= \max_{m' \in [0, 1]_N: m' \leq m} h(\mathbf{X}_{t+1}, [Nm']) - \max_{m' \in [0, 1]_N: m' \leq m} \gamma h(\mathbf{X}_t, [Nm']) \\ &\leq \max_{m' \in [0, 1]_N: m' \leq m} (h(\mathbf{X}_{t+1}, [Nm']) - \gamma h(\mathbf{X}_t, [Nm'])) \\ &\leq \max_{m' \in [0, 1]_N: m' \leq m} (h(\mathbf{X}_{t+1}, [Nm']) - h(\mathbf{X}_t P, [Nm'])) \end{aligned} \quad (6.50)$$

$$\leq \max_{m' \in [0,1]_N : m' \leq m} \max_{g \in \mathcal{G}} \sup_{\ell \in \mathbb{N}} \gamma^{-\ell} \left| \sum_{i \in [Nm']} \langle \epsilon_{i,t}, (P_i - \Xi_i)^\ell g_i \rangle \right|, \quad (6.51)$$

where (6.50) follows from (6.40); $\epsilon_{i,t}$ is a $|\mathbb{S}|$ -dimensional random vector given by $\epsilon_{i,t} \triangleq X_{i,t+1} - X_{i,t}P_i$, which satisfies $\mathbb{E}[\epsilon_{i,t}(s) \mid \mathbf{X}_t] = 0$ for any $s \in \mathbb{S}$ and $\|\epsilon_{i,t}\|_1 \leq 2$; (6.51) follows (6.41). Now, we take the expectations of the positive parts of the inequality (6.51) conditioned on $\mathbf{X}_t$, which yields

$$\begin{aligned} & \mathbb{E}[(h_{\text{ID}}(\mathbf{X}_{t+1}, m) - \gamma h_{\text{ID}}(\mathbf{X}_t, m))^+ \mid \mathbf{X}_t] \\ & \leq \mathbb{E} \left[\max_{m' \in [0,1]_N : m' \leq m} \max_{g \in \mathcal{G}} \sup_{\ell \in \mathbb{N}} \gamma^{-\ell} \left| \sum_{i \in [Nm']} \langle \epsilon_{i,t}, (P_i - \Xi_i)^\ell g_i \rangle \right| \mid \mathbf{X}_t \right] \\ & \leq \sum_{\ell=0}^{\infty} \gamma^{-\ell} \sum_{g \in \mathcal{G}} \mathbb{E} \left[\max_{m' \in [0,1]_N : m' \leq m} \left| \sum_{i \in [Nm']} \langle \epsilon_{i,t}, (P_i - \Xi_i)^\ell g_i \rangle \right| \mid \mathbf{X}_t \right] \\ & \leq \sum_{\ell=0}^{\infty} \gamma^{-\ell} \sum_{g \in \mathcal{G}} \mathbb{E} \left[\max_{m' \in [0,1]_N : m' \leq m} \left| \sum_{i \in [Nm']} \langle \epsilon_{i,t}, (P_i - \Xi_i)^\ell g_i \rangle \right|^2 \mid \mathbf{X}_t \right]^{1/2}. \quad (6.52) \end{aligned}$$

Because $\epsilon_{i,t}$ are independent across $i \in [Nm]$ and $\langle \epsilon_{i,t}, (P_i - \Xi_i)^\ell g_i \rangle$ has zero mean conditioned on $\mathbf{X}_t$, the partial sum $\sum_{i \in [n]} \langle \epsilon_{i,t}, (P_i - \Xi_i)^\ell g_i \rangle$ is a martingale in n and $\left| \sum_{i \in [n]} \langle \epsilon_{i,t}, (P_i - \Xi_i)^\ell g_i \rangle \right|$ is a submartingale in n for $n \in [Nm]$. Then it follows from Doob's L_2 inequality (86, Theorem 5.4.3) that

$$\begin{aligned} & \mathbb{E} \left[\max_{m' \in [0,1]_N : m' \leq m} \left| \sum_{i \in [Nm']} \langle \epsilon_{i,t}, (P_i - \Xi_i)^\ell g_i \rangle \right|^2 \mid \mathbf{X}_t \right]^{1/2} \\ & \leq 2 \mathbb{E} \left[\left(\sum_{i \in [Nm]} \langle \epsilon_{i,t}, (P_i - \Xi_i)^\ell g_i \rangle \right)^2 \mid \mathbf{X}_t \right]^{1/2} \\ & = 2 \left(\sum_{i \in [Nm]} \mathbb{E} \left[\langle \epsilon_{i,t}, (P_i - \Xi_i)^\ell g_i \rangle^2 \mid \mathbf{X}_t \right] \right)^{1/2} \end{aligned}$$

$$\begin{aligned}
&\leq 2 \left(\sum_{i \in [Nm]} \mathbb{E} \left[\left(\|\epsilon_{i,t}\|_1 \|(P_i - \Xi_i)^\ell\|_\infty \|g_i\|_\infty \right)^2 \mid \mathbf{X}_t \right] \right)^{1/2} \\
&\leq 4\sqrt{Nm} \max_{i \in [Nm]} \|g_i\|_\infty \|(P_i - \Xi_i)^\ell\|_\infty, \tag{6.53}
\end{aligned}$$

where (6.53) uses the fact that $\|\epsilon_{i,t}\|_1 \leq 2$. Substituting the bound in (6.53) back to (6.52), we get

$$\begin{aligned}
&\mathbb{E} [(h_{\text{ID}}(\mathbf{X}_{t+1}, m) - \gamma h_{\text{ID}}(\mathbf{X}_t, m))^+ \mid \mathbf{X}_t] \\
&\leq 4\sqrt{Nm} \sum_{\ell=0}^{\infty} \gamma^{-\ell} \sum_{g \in \mathcal{G}} \max_{i \in [Nm]} \|g_i\|_\infty \|(P_i - \Xi_i)^\ell\|_\infty \\
&= 4\sqrt{N}(Kc_{\max} + r_{\max}) \sum_{\ell=0}^{\infty} \gamma^{-\ell} \max_{i \in [N]} \|(P_i - \Xi_i)^\ell\|_\infty \\
&\leq 4\sqrt{N}(Kc_{\max} + r_{\max})C_\tau \\
&= 2C_h\sqrt{N}. \tag{6.54}
\end{aligned}$$

□

6.G Lemmas for focus set

In this section, we present and prove three lemmas about properties of the focus set. Recall that for any system state $\mathbf{x}$, the focus set is defined as the set $[Nm(\mathbf{x})]$, where $m(\mathbf{x})$ is given by

$$m(\mathbf{x}) = \max \left\{ m \in [0, 1]_N : h_{\text{ID}}(\mathbf{x}, m) \leq \min_{k \in [K]} \overline{C}_k^*([Nm]) \right\}. \tag{6.54}$$

Consider the system state process under the ID policy, $(\mathbf{S}_t, t \in \mathbb{N})$ and its equivalent representation $(\mathbf{X}_t, t \in \mathbb{N})$. We often consider the focus set corresponding to the current system state, i.e., $m(\mathbf{X}_t)$. A closely related quantity is the number of arms that follow their optimal single-armed policies under the ID policy, which we refer to as the *conforming number*. With the system state $\mathbf{S}_t$, the conforming number is denoted as N_t^* , and it can be written as

$$N_t^* = \max \left\{ n \in [N] : \sum_{i=1}^n c_{k,i}(S_{i,t}, \hat{A}_{i,t}) \leq \alpha_k N, \forall k \in [K] \right\}, \tag{6.55}$$

where $\hat{A}_{i,t}$'s are the ideal actions sampled from the optimal single-armed policies by the ID policy.

Below we state the three lemmas, and we prove them in the subsections.

Lemma 6.13 (Majority conformity). *Let $(\mathbf{X}_t, t \in \mathbb{N})$ be the system state process under the ID policy. The size of the focus set, $Nm(\mathbf{X}_t)$, satisfies*

$$\frac{1}{N} \mathbb{E}[(Nm(\mathbf{X}_t) - N_t^*)^+ \mid \mathbf{X}_t] \leq \frac{K_{conf}}{\sqrt{N}}, \quad \text{with probability 1,}$$

for some constant $K_{conf} > 0$ independent of N .

Lemma 6.13 implies that almost all the arms in the focus set, except for $O(\sqrt{N})$ arms, can follow the optimal single-armed policies.

Lemma 6.14 (Almost non-shrinking). *Let $(\mathbf{X}_t, t \in \mathbb{N})$ be the system state process under the ID policy. Then the change in the size of the focus set over time satisfies*

$$\mathbb{E}[(m(\mathbf{X}_t) - m(\mathbf{X}_{t+1}))^+ \mid \mathbf{X}_t] \leq \frac{K_{mono}}{\sqrt{N}}, \quad \text{with probability 1,}$$

for some constant $K_{mono} > 0$ independent of N .

Lemma 6.14 implies that the size of the focus set is almost non-shrinking on average over time, or more specifically, it shrinks by at most $O(\sqrt{N})$ on average over time.

Lemma 6.15 (Sufficient coverage). *Let $(\mathbf{X}_t, t \in \mathbb{N})$ be the system state process under the ID policy. Then*

$$1 - m(\mathbf{X}_t) \leq \frac{1}{\eta_c N} h_{ID}(\mathbf{X}_t, m(\mathbf{X}_t)) + \frac{K_{cov}}{N}, \quad \text{with probability 1,}$$

for some constant $K_{cov} > 0$ independent of N .

Lemma 6.15 relates the size of the complement of the focus set to the value of the function $h_{ID}(\mathbf{X}_t, m(\mathbf{X}_t))$.

6.G.1 Proof of Lemma 6.13 (Majority conformity)

Proof. First, we claim that the conforming number N_t^* can be lower bounded using our slack budget function $\bar{C}^*$ and Lyapunov function as follows: for any time step $t \geq 0$,

$$N_t^* \geq N \max \left\{ m \in [0, 1]_N : \bar{C}^*([Nm]) - h_{ID}(\mathbf{X}_t, m) \geq \Delta_t \right\}, \quad (6.56)$$

where Δ_t is the random variable given by

$$\Delta_t = \max_{k \in [K]} \max_{m \in [0,1]_N} \left| \sum_{i \in [Nm]} \left(c_{k,i}(S_{i,t}, \hat{A}_{i,t}) - \langle X_{i,t}, c_{k,i}^* \rangle \right) \right|, \quad (6.57)$$

which captures the difference between the expected cost and the actual cost.

To prove the claim, we invoke the definition of N_t^* and the fact that $\bar{C}^*([n]) \leq \alpha_k N - C_k([n])$ for any $n \in [N]$ and $k \in [K]$:

$$\begin{aligned} N_t^* &= \max \left\{ n \in [N] : \sum_{i \in [n]} c_{k,i}(S_{i,t}, \hat{A}_{i,t}) \leq \alpha_k N \quad \forall k \in [K] \right\} \\ &\geq \max \left\{ n \in [N] : \max_{k \in [K]} \sum_{i \in [n]} \left(c_{k,i}(S_{i,t}, \hat{A}_{i,t}) - C_k([n]) \right) \leq \bar{C}^*([n]) \right\}. \end{aligned} \quad (6.58)$$

To further lower bound (6.58), we identify a subset of the set in (6.58) and reduce the task to upper bounding the following expression:

$$\begin{aligned} &\max_{k \in [K]} \sum_{i \in [n]} \left(c_{k,i}(S_{i,t}, \hat{A}_{i,t}) - C_k([n]) \right) \\ &\leq \max_{k \in [K]} \left| \sum_{i \in [n]} \langle X_{i,t}, c_{k,i}^* \rangle - C_k([n]) \right| + \max_{k \in [K]} \left| \sum_{i \in [n]} c_{k,i}(S_{i,t}, \hat{A}_{i,t}) - \sum_{i \in [n]} \langle X_{i,t}, c_{k,i}^* \rangle \right| \\ &\leq h_{\text{ID}}(\mathbf{X}_t, n/N) + \max_{k \in [K]} \left| \sum_{i \in [n]} \left(c_{k,i}(S_{i,t}, \hat{A}_{i,t}) - \langle X_{i,t}, c_{k,i}^* \rangle \right) \right| \end{aligned} \quad (6.59)$$

$$\leq h_{\text{ID}}(\mathbf{X}_t, n/N) + \Delta_t, \quad (6.60)$$

where the inequality (6.59) follows from the argument that

$$\begin{aligned} h_{\text{ID}}(\mathbf{X}_t, n/N) &= \max_{n' \in [n]} \max_{g \in \mathcal{G}} \sup_{\ell \in \mathbb{N}} \left| \sum_{i \in [n']} \langle (X_{i,t} - \mu_i^*) P_i^\ell / \gamma^\ell, g_i \rangle \right| \\ &\geq \max_{k \in [K]} \left| \sum_{i \in [n]} \langle X_{i,t} - \mu_i^*, c_{k,i}^* \rangle \right| \end{aligned}$$

$$= \max_{k \in [K]} \left| \sum_{i \in [n]} \langle X_{i,t}, c_{k,i}^* \rangle - C_k([n]) \right|,$$

and (6.60) follows from the definition of Δ_t in (6.57) by taking $m = n/N$ in the maximum. Substituting the bound in (6.60) into (6.58), we get

$$N_t^* \geq \max \left\{ n \in [N] : h_{\text{ID}}(\mathbf{X}_t, n/N) + \Delta_t \leq \overline{C}^*([n]) \right\},$$

which is equivalent to the claim in (6.56).

We now use (6.56) to prove the lemma. Observe that by the definition of $m(\mathbf{X}_t)$,

$$\overline{C}^*([Nm(\mathbf{X}_t)]) - h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)) \geq 0.$$

and $h_{\text{ID}}(\mathbf{X}_t, m)$ is non-decreasing in m . Consequently, for any $m \in [0, 1]_N$ such that $m \leq m(\mathbf{X}_t)$,

$$\begin{aligned} \overline{C}^*([Nm]) - h_{\text{ID}}(\mathbf{X}_t, m) &\geq \overline{C}^*([Nm]) - h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)) \\ &\geq \overline{C}^*([Nm]) - \overline{C}^*([Nm(\mathbf{X}_t)]) \\ &= \min_{k \in [K]} \overline{C}_k^*([Nm]) - \min_{k \in [K]} \overline{C}_k^*([Nm(\mathbf{X}_t)]) \\ &\geq \min_{k \in [K]} \left(\overline{C}_k^*([Nm]) - \overline{C}_k^*([Nm(\mathbf{X}_t)]) \right) \\ &\geq \eta_c N(m(\mathbf{X}_t) - m) - M_c, \end{aligned} \quad (6.61)$$

where (6.61) follows from the strict slope of the remaining budget $\overline{C}^*([Nm])$ (Lemma 6.7). Therefore, choosing $m = m(\mathbf{X}_t) - (\Delta_t + M_c)/(\eta_c N)$, we obtain

$$\overline{C}^*([Nm]) - h_{\text{ID}}(\mathbf{X}_t, m) \geq \Delta_t.$$

Recalling the lower bound of N_t^* established in (6.56), we arrive at

$$\begin{aligned} N_t^* &\geq N \max \left\{ m \in [0, 1]_N : \overline{C}^*([Nm]) - h_{\text{ID}}(\mathbf{X}_t, m) \geq \Delta_t \right\} \\ &\geq Nm(\mathbf{X}_t) - \frac{\Delta_t + M_c}{\eta_c}. \end{aligned}$$

Rearranging the terms and taking the conditional expectations, we establish that

$$\mathbb{E} \left[(Nm(\mathbf{X}_t) - N_t^*)^+ \mid \mathbf{X}_t \right] \leq \mathbb{E} \left[\frac{\Delta_t + M_c}{\eta_c} \mid \mathbf{X}_t \right]. \quad (6.62)$$

It remains to upper bound the conditional expectation of Δ_t given $\mathbf{X}_t$. We define the random variable $\xi_{k,i} \triangleq c_{k,i}(S_{i,t}, \hat{A}_{i,t}) - \langle X_{i,t}, c_{k,i}^* \rangle$ for each arm $i \in [N]$ and cost type $k \in [K]$. Subsequently, $\mathbb{E}[\Delta_t \mid \mathbf{X}_t]$ can be rewritten and bounded as

$$\begin{aligned} \mathbb{E}[\Delta_t \mid \mathbf{X}_t] &= \mathbb{E} \left[\max_{k \in [K]} \max_{m \in [0,1]^N} \left| \sum_{i \in [Nm]} \xi_{k,i} \right| \mid \mathbf{X}_t \right] \\ &\leq \sum_{k \in [K]} \mathbb{E} \left[\max_{n \in [N]} \left| \sum_{i \in [n]} \xi_{k,i} \right| \mid \mathbf{X}_t \right] \\ &\leq \sum_{k \in [K]} \mathbb{E} \left[\left(\max_{n \in [N]} \left| \sum_{i \in [n]} \xi_{k,i} \right| \right)^2 \mid \mathbf{X}_t \right]^{1/2}, \end{aligned} \quad (6.63)$$

where (6.63) follows from the Cauchy–Schwarz inequality. Next, we argue that the sequence $\left\{ \left| \sum_{i \in [n]} \xi_{k,i} \right| \right\}_{n \in [N]}$ is a submartingale, which allows us to apply Doob’s L_2 inequality (86, Theorem 5.4.3) to bound the expression in (6.63). Observe that, for each cost type k , conditioned on $\mathbf{X}_t$, the sequence of random variables $\{\xi_{k,i}\}_{i \in [N]}$ are independent and have zero conditional means. Consequently, the sequence of partial sums $\left\{ \sum_{i \in [n]} \xi_{k,i} \right\}_{n \in [N]}$ forms a martingale, which becomes a submartingale upon taking the absolute value. Thus, by applying Doob’s L_2 inequality and utilizing the bound $\xi_{k,i} \leq c_{\max}$, we obtain

$$\begin{aligned} \mathbb{E}[\Delta_t \mid \mathbf{X}_t] &\leq \sum_{k \in [K]} 2 \mathbb{E} \left[\left(\sum_{i \in [N]} \xi_{k,i} \right)^2 \mid \mathbf{X}_t \right]^{1/2} \\ &= \sum_{k \in [K]} 2 \left(\sum_{i \in [N]} \mathbb{E}[\xi_{k,i}^2 \mid \mathbf{X}_t] \right)^{1/2} \\ &\leq 2K c_{\max} \sqrt{N}. \end{aligned} \quad (6.64)$$

Combining (6.64) with our earlier calculations, we get

$$\mathbb{E} \left[(Nm(\mathbf{X}_t) - N_t^*)^+ \mid \mathbf{X}_t \right] \leq \mathbb{E} \left[\frac{\Delta_t + M_c}{\eta_c} \mid \mathbf{X}_t \right]$$

$$\begin{aligned}
&\leq \frac{2Kc_{\max}}{\eta_c} \sqrt{N} + \frac{M_c}{\eta_c} \\
&\leq K_{\text{conf}} \sqrt{N}.
\end{aligned}$$

where $K_{\text{conf}} \triangleq (2Kc_{\max} + M_c)/\eta_c$. $\square$

6.G.2 Proof of Lemma 6.14 (Almost non-shrinking)

We first state and prove a supporting lemma below, which will be used in the proof of Lemma 6.14.

Lemma 6.16. *Under the ID policy, we have*

$$\mathbb{E} \left[(h_{\text{ID}}(\mathbf{X}_{t+1}, m(\mathbf{X}_t)) - \gamma h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)))^+ \mid \mathbf{X}_t \right] \leq (2 + K_{\text{conf}}) C_h \sqrt{N},$$

where $C_h > 0$ is the positive constant given in Lemma 6.8.

Proof. We upper bound $\mathbb{E} [(h_{\text{ID}}(\mathbf{X}_{t+1}, m(\mathbf{X}_t)) - \gamma h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)))^+ \mid \mathbf{X}_t]$ by coupling $\mathbf{X}_{t+1}$ with a random element $\mathbf{X}'_{t+1}$ constructed as follows: Let $\mathbf{X}'_{t+1}$ be the system state at step $t+1$ if we were able to set $A_{i,t} = \hat{A}_{i,t}$ for all $i \in [N]$. From the drift condition of the Lyapunov function $h_{\text{ID}}(\cdot, D)$, as established in Lemma 6.12, we obtain:

$$\mathbb{E} [(h_{\text{ID}}(\mathbf{X}'_{t+1}, m(\mathbf{X}_t)) - \gamma h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)))^+ \mid \mathbf{X}_t] \leq 2C_h \sqrt{N}. \quad (6.65)$$

We couple $\mathbf{X}'_{t+1}$ and $\mathbf{X}_{t+1}$ such that $\mathbf{X}'_{i,t+1} = \mathbf{X}_{i,t+1}$ for all $i \leq \min \{N_t^*, Nm(\mathbf{X}_t)\}$. Let $\mathcal{U}_t = [Nm(\mathbf{X}_t)] \setminus [N_t^*]$. Then we have

$$\begin{aligned}
&\mathbb{E} \left[(h_{\text{ID}}(\mathbf{X}_{t+1}, m(\mathbf{X}_t)) - \gamma h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)))^+ \mid \mathbf{X}_t \right] \\
&\quad - \mathbb{E} \left[(h_{\text{ID}}(\mathbf{X}'_{t+1}, m(\mathbf{X}_t)) - \gamma h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)))^+ \mid \mathbf{X}_t \right] \\
&\leq \mathbb{E} \left[(h_{\text{ID}}(\mathbf{X}_{t+1}, m(\mathbf{X}_t)) - h_{\text{ID}}(\mathbf{X}'_{t+1}, m(\mathbf{X}_t)))^+ \mid \mathbf{X}_t \right] \\
&\leq \mathbb{E} \left[\max_{\substack{m' \in [0,1]_N \\ m' \leq m(\mathbf{X}_t)}} (h(\mathbf{X}_{t+1}, [Nm']) - h(\mathbf{X}'_{t+1}, [Nm']))^+ \mid \mathbf{X}_t \right] \\
&\leq \mathbb{E} \left[\max_{\substack{m' \in [0,1]_N \\ m' \leq m(\mathbf{X}_t)}} \max_{g \in \mathcal{G}} \sup_{\ell \in \mathbb{N}} \left| \sum_{i \in [Nm']} \langle (X'_{i,t+1} - X_{i,t+1}) P_i^\ell \gamma^{-\ell}, g_i \rangle \right| \mid \mathbf{X}_t \right]
\end{aligned}$$

$$\begin{aligned}
&= \mathbb{E} \left[\max_{\substack{m' \in [0,1]_N \\ m' \leq m(\mathbf{X}_t)}} \max_{g \in \mathcal{G}} \sup_{\ell \in \mathbb{N}} \left| \sum_{i \in [Nm']} \langle (X'_{i,t+1} - X_{i,t+1})(P_i - \Xi_i)^\ell \gamma^{-\ell}, g_i \rangle \right| \middle| \mathbf{X}_t \right] \\
&= \mathbb{E} \left[\max_{\substack{m' \in [0,1]_N \\ m' \leq m(\mathbf{X}_t)}} \max_{g \in \mathcal{G}} \sup_{\ell \in \mathbb{N}} \gamma^{-\ell} \left| \sum_{i \in [Nm'] \cap \mathcal{U}_t} \langle (X'_{i,t+1} - X_{i,t+1})(P_i - \Xi_i)^\ell, g_i \rangle \right| \middle| \mathbf{X}_t \right] \\
&\leq \mathbb{E} \left[\sum_{i \in \mathcal{U}_t} \sum_{g \in \mathcal{G}} \sum_{\ell=0}^{\infty} \gamma^{-\ell} |\langle (X'_{i,t+1} - X_{i,t+1})(P_i - \Xi_i)^\ell, g_i \rangle| \middle| \mathbf{X}_t \right] \\
&\leq \mathbb{E} \left[\sum_{i \in \mathcal{U}_t} \sum_{g \in \mathcal{G}} \sum_{\ell=0}^{\infty} \gamma^{-\ell} \|X_{i,t+1} - X'_{i,t+1}\|_1 \|(P_i - \Xi_i)^\ell\|_\infty \|g_i\|_\infty \middle| \mathbf{X}_t \right] \\
&\leq \mathbb{E} \left[\sum_{i \in \mathcal{U}_t} 2 \sum_{\ell=0}^{\infty} \gamma^{-\ell} \|(P_i - \Xi_i)^\ell\|_\infty \sum_{g \in \mathcal{G}} \|g_i\|_\infty \middle| \mathbf{X}_t \right] \\
&\leq \mathbb{E} \left[\sum_{i \in \mathcal{U}_t} 2C_\tau(Kc_{\max} + r_{\max}) \middle| \mathbf{X}_t \right] \\
&\leq 2C_\tau(Kc_{\max} + r_{\max}) \mathbb{E} [(Nm(\mathbf{X}_t) - N_t^*)^+ \mid \mathbf{X}_t] \tag{6.66} \\
&\leq 2C_\tau(Kc_{\max} + r_{\max}) K_{\text{conf}} \sqrt{N}, \tag{6.67}
\end{aligned}$$

where we have applied Lemma 6.13 to bound the expression in (6.66).

By combining (6.65) and (6.67), we obtain:

$$\begin{aligned}
&\mathbb{E} [(h_{\text{ID}}(\mathbf{X}_{t+1}, m(\mathbf{X}_t)) - \gamma h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)))^+ \mid \mathbf{X}_t] \\
&\leq (2C_h + 2C_\tau(Kc_{\max} + r_{\max})K_{\text{conf}}) \sqrt{N} \\
&= (2C_h + C_h K_{\text{conf}}) \sqrt{N},
\end{aligned}$$

where the last equality is because $C_h = 2C_\tau(Kc_{\max} + r_{\max})$. $\square$

We now give the proof of Lemma 6.14.

Proof of Lemma 6.14. First, we claim that

$$m(\mathbf{X}_{t+1}) \geq m(\mathbf{X}_t) - \frac{1}{\eta_c N} (h_{\text{ID}}(\mathbf{X}_{t+1}, m(\mathbf{X}_t)) - h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)))^+ - \frac{M_c}{\eta_c N}. \tag{6.68}$$

To prove the claim, by the maximality of $m(\mathbf{X}_{t+1})$ it suffices to show that for any $\bar{m} \in [0, 1]_N$ s.t.

$$\bar{m} \leq m(\mathbf{X}_t) - \frac{1}{\eta_c N} (h_{\text{ID}}(\mathbf{X}_{t+1}, m(\mathbf{X}_t)) - h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)))^+ - \frac{M_c}{\eta_c N}, \quad (6.69)$$

we have $h_{\text{ID}}(\mathbf{X}_{t+1}, \bar{m}) \leq \bar{C}^*([N\bar{m}])$. For any $\bar{m}$ satisfying (6.69), Lemma 6.7 implies that

$$\begin{aligned} & \bar{C}^*([N\bar{m}]) - \bar{C}^*([Nm(\mathbf{X}_t)]) \\ & \geq \eta_c N (m(\mathbf{X}_t) - \bar{m}) - M_c \\ & \geq \eta_c N \left(\frac{1}{\eta_c N} (h_{\text{ID}}(\mathbf{X}_{t+1}, \bar{m}) - h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)))^+ + \frac{M_c}{\eta_c N} \right) - M_c \\ & \geq (h_{\text{ID}}(\mathbf{X}_{t+1}, \bar{m}) - h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)))^+ . \end{aligned}$$

Since $\bar{C}^*(m([N\mathbf{X}_t])) \geq h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t))$ by the definition of $m(\mathbf{X}_t)$, we thus have

$$\begin{aligned} \bar{C}^*([N\bar{m}]) & \geq \bar{C}^*([Nm(\mathbf{X}_t)]) + (h_{\text{ID}}(\mathbf{X}_{t+1}, \bar{m}) - h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)))^+ \\ & \geq h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)) + (h_{\text{ID}}(\mathbf{X}_{t+1}, \bar{m}) - h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)))^+ \\ & \geq h_{\text{ID}}(\mathbf{X}_{t+1}, \bar{m}) , \end{aligned}$$

which proves the claim in (6.68).

Taking the conditional expectations in (6.68) and rearranging the terms, we get

$$\begin{aligned} & \mathbb{E} [(m(\mathbf{X}_t) - m(\mathbf{X}_{t+1}))^+ \mid \mathbf{X}_t] \\ & \leq \frac{1}{\eta_c N} \mathbb{E} [(h_{\text{ID}}(\mathbf{X}_{t+1}, m(\mathbf{X}_t)) - h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)))^+ \mid \mathbf{X}_t] + \frac{M_c}{N\eta_c} , \end{aligned}$$

where the right-hand side can be further bounded using Lemma 6.16, which states that

$$\mathbb{E} [(h_{\text{ID}}(\mathbf{X}_{t+1}, m(\mathbf{X}_t)) - h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)))^+ \mid \mathbf{X}_t] \leq (2 + K_{\text{conf}})C_h\sqrt{N} .$$

Therefore, we have

$$\mathbb{E} [(m(\mathbf{X}_t) - m(\mathbf{X}_{t+1}))^+ \mid \mathbf{X}_t] \leq \frac{(2 + K_{\text{conf}})C_h}{\eta_c\sqrt{N}} + \frac{M_c}{\eta_c N} ,$$

which implies

$$\mathbb{E} [(m(\mathbf{X}_t) - m(\mathbf{X}_{t+1}))^+ \mid \mathbf{X}_t] \leq \frac{K_{\text{mono}}}{\sqrt{N}},$$

with $K_{\text{mono}} \triangleq ((2 + K_{\text{conf}})C_h + M_c)/\eta_c$. $\square$

6.G.3 Proof of Lemma 6.15 (Sufficient coverage)

Proof. Observe that it suffices to focus on the case when $m(\mathbf{X}_t) \neq 1$. Recall that for any system state $\mathbf{x}$, $m(\mathbf{x})$ is defined as

$$m(\mathbf{x}) = \max \left\{ m \in [0, 1]_N : h_{\text{ID}}(\mathbf{x}, m) \leq \overline{C}^*([Nm]) \right\}. \quad (6.28)$$

Because $m(\mathbf{X}_t) \neq 1$, we have $m(\mathbf{X}_t) + 1/N \in [0, 1]_N$. Then the maximality of $m(\mathbf{X}_t)$ implies that

$$h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t) + 1/N) > \overline{C}^*([Nm(\mathbf{X}_t) + 1]). \quad (6.70)$$

We can upper bound the left-hand side of (6.70) using the Lipschitz continuity of h :

$$h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t) + 1/N) \leq h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)) + L_h. \quad (6.71)$$

We then lower bound the right-hand side of (6.70) using Lemma 6.7:

$$\begin{aligned} \overline{C}^*([Nm(\mathbf{X}_t) + 1]) &\geq \overline{C}^*([Nm(\mathbf{X}_t) + 1]) - \overline{C}^*([N]) \\ &\geq \eta_c(N - Nm(\mathbf{X}_t) - 1) - M_c \\ &= \eta_c N(1 - m(\mathbf{X}_t)) - \eta_c - M_c. \end{aligned} \quad (6.72)$$

Comparing (6.70), (6.71) and (6.72), we have:

$$h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)) \geq \eta_c N(1 - m(\mathbf{X}_t)) - \eta_c - M_c - L_h,$$

which, after rearranging the terms, implies

$$1 - m(\mathbf{X}_t) \leq \frac{1}{\eta_c N} h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)) + \frac{K_{\text{cov}}}{N},$$

with $K_{\text{cov}} \triangleq (\eta_c + M_c + L_h)/\eta_c$. $\square$

6.H Proofs of Lemma 6.9 and Lemma 6.10

In this section, we provide two final lemmas, Lemma 6.9 and Lemma 6.10, which together imply Theorem 6.3. We prove these two lemmas in Sections 6.H.1 and 6.H.2, respectively.

6.H.1 Proof of Lemma 6.9

Lemma 6.9. *Consider any N -armed WCMDP with initial system state $\mathbf{S}_0$ and assume that it satisfies Assumption 6.2. Let policy π be the ID policy. Consider the Lyapunov function V defined in (6.22). Then the optimality gap of π is bounded as*

$$R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0) \leq \frac{2r_{\max} + L_h}{L_h N} \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[V(\mathbf{X}_t)] + \frac{2r_{\max} K_{\text{conf}}}{\sqrt{N}},$$

where L_h is the Lipschitz constant in Lemma 6.8 and K_{conf} is the positive constant in Lemma 6.13.

Proof. We can bound the optimality gap as the following long-run average:

$$\begin{aligned} R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0) &\leq R^{\text{rel}}(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0) \\ &= R^{\text{rel}}(N, \mathbf{S}_0) - \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N} \sum_{i \in [N]} \mathbb{E}[r_i(S_{i,t}, A_{i,t})] \\ &= \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \left(R^{\text{rel}}(N, \mathbf{S}_0) - \frac{1}{N} \sum_{i \in [N]} \mathbb{E}[r_i(S_{i,t}, A_{i,t})] \right). \end{aligned}$$

To bound $R^{\text{rel}}(N, \mathbf{S}_0) - \sum_{i \in [N]} \mathbb{E}[r_i(S_{i,t}, A_{i,t})]/N$, we calculate that

$$\begin{aligned} &R^{\text{rel}}(N, \mathbf{S}_0) - \frac{1}{N} \sum_{i \in [N]} \mathbb{E}[r_i(S_{i,t}, A_{i,t})] \\ &= \frac{1}{N} \sum_{i \in [N]} \sum_{s \in \mathbb{S}, a \in \mathbb{A}} r_i(s, a) y_i^*(s, a) - \frac{1}{N} \sum_{i \in [N]} \mathbb{E}[r_i(S_{i,t}, A_{i,t})] \\ &\leq \frac{1}{N} \sum_{i \in [N]} \sum_{s \in \mathbb{S}, a \in \mathbb{A}} r_i(s, a) y_i^*(s, a) - \frac{1}{N} \sum_{i \in [N]} \mathbb{E}[r_i(S_{i,t}, \hat{A}_{i,t})] \end{aligned}$$

$$\begin{aligned}
& + \frac{2r_{\max}}{N} \sum_{i \in [N]} \mathbb{E} \left[\mathbf{1} \left\{ A_{i,t} \neq \hat{A}_{i,t} \right\} \right] \\
& = \frac{1}{N} \sum_{i \in [N]} \langle \mu_i^* - \mathbb{E}[X_{i,t}], r_i^* \rangle + \frac{2r_{\max}}{N} \sum_{i \in [N]} \mathbb{E} \left[\mathbf{1} \left\{ A_{i,t} \neq \hat{A}_{i,t} \right\} \right] \\
& \leq \frac{1}{N} \sum_{i \in [N]} \langle \mu_i^* - \mathbb{E}[X_{i,t}], r_i^* \rangle + 2r_{\max} \mathbb{E} \left[1 - \frac{N_t^*}{N} \right] \tag{6.73}
\end{aligned}$$

$$\leq \frac{1}{N} \sum_{i \in [N]} \langle \mu_i^* - \mathbb{E}[X_{i,t}], r_i^* \rangle + 2r_{\max} \mathbb{E} [1 - m(\mathbf{X}_t)] + \frac{2r_{\max} K_{\text{conf}}}{\sqrt{N}} \tag{6.74}$$

$$\leq \frac{1}{N} \mathbb{E} [h_{\text{ID}}(\mathbf{X}_t, 1)] + 2r_{\max} \mathbb{E} [1 - m(\mathbf{X}_t)] + \frac{2r_{\max} K_{\text{conf}}}{\sqrt{N}}, \tag{6.75}$$

where (6.73) is due to the definition of N_t^* ; (6.74) is due to the bound on $\mathbb{E}[(Nm(\mathbf{X}_t) - N_t^*)^+]$ in Lemma 6.13, and (6.75) directly follows from the definition of $h_{\text{ID}}(\mathbf{x}, 1)$. We further bound the first two expressions in (6.75) in terms of $V(\mathbf{X}_t)$ as follows:

$$h_{\text{ID}}(\mathbf{X}_t, 1) \leq h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)) + L_h N(1 - m(\mathbf{X}_t)) = V(\mathbf{X}_t) \tag{6.76}$$

$$1 - m(\mathbf{X}_t) \leq \frac{1}{L_h N} V(\mathbf{X}_t), \tag{6.77}$$

where (6.76) is due to the Lipschitz continuity of $h_{\text{ID}}(\mathbf{x}, m)$ with respect to the parameter m (Lemma 6.12), and (6.77) is due to the definition of $V(\mathbf{x})$.

Combining the above calculations, we get

$$\begin{aligned}
R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0) & \leq \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \left(\frac{1}{N} \mathbb{E} [h_{\text{ID}}(\mathbf{X}_t, 1)] + 2r_{\max} \mathbb{E} [1 - m(\mathbf{X}_t)] \right. \\
& \quad \left. + \frac{2r_{\max} K_{\text{conf}}}{\sqrt{N}} \right) \\
& \leq \frac{2r_{\max} + L_h}{L_h N} \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [V(\mathbf{X}_t)] + \frac{2r_{\max} K_{\text{conf}}}{\sqrt{N}}.
\end{aligned}$$

□

6.H.2 Proof of Lemma 6.10

Lemma 6.10. *Consider any N -armed WCMDP with initial system state $\mathbf{S}_0$ and assume that it satisfies Assumption 6.2. Let $\mathbf{X}_t$ be the system state at*

time t under the ID policy. Consider the Lyapunov function V defined in (6.22). Then its drift satisfies

$$\mathbb{E}[V(\mathbf{X}_{t+1}) \mid \mathbf{X}_t] - V(\mathbf{X}_t) \leq -\rho_V V(\mathbf{X}_t) + K_V \sqrt{N}, \quad (6.29)$$

which further implies that

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[V(\mathbf{X}_t)] \leq \frac{K_V \sqrt{N}}{\rho_V}, \quad (6.30)$$

where ρ_V and K_V are constants whose values are given in the proof.

Proof. We derive a recurrence relation between $\mathbb{E}[V(\mathbf{X}_{t+1})]$ and $\mathbb{E}[V(\mathbf{X}_t)]$, by bounding $\mathbb{E}[V(\mathbf{X}_{t+1}) \mid \mathbf{X}_t] - V(\mathbf{X}_t)$. Specifically, observe that by the Lipschitz continuity of $h_{\text{ID}}(\mathbf{X}, D)$ with respect to D , we have

$$\begin{aligned} V(\mathbf{X}_{t+1}) &= h_{\text{ID}}(\mathbf{X}_{t+1}, m(\mathbf{X}_{t+1})) + L_h N(1 - m(\mathbf{X}_{t+1})) \\ &\leq h_{\text{ID}}(\mathbf{X}_{t+1}, m(\mathbf{X}_t)) + L_h N(1 - m(\mathbf{X}_t)) \\ &\quad + 2L_h N(m(\mathbf{X}_t) - m(\mathbf{X}_{t+1}))^+. \end{aligned}$$

Consequently, we have

$$\begin{aligned} &\mathbb{E}[V(\mathbf{X}_{t+1}) \mid \mathbf{X}_t] - V(\mathbf{X}_t) \\ &\leq \mathbb{E}[h_{\text{ID}}(\mathbf{X}_{t+1}, m(\mathbf{X}_t)) \mid \mathbf{X}_t] - h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)) \\ &\quad + 2L_h N \mathbb{E}[(m(\mathbf{X}_t) - m(\mathbf{X}_{t+1}))^+ \mid \mathbf{X}_t] \\ &\leq \mathbb{E}[h_{\text{ID}}(\mathbf{X}_{t+1}, m(\mathbf{X}_t)) \mid \mathbf{X}_t] - h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)) + 2L_h K_{\text{mono}} \sqrt{N}, \end{aligned} \quad (6.78)$$

where the last inequality follows from Lemma 6.14.

To bound $\mathbb{E}[h_{\text{ID}}(\mathbf{X}_{t+1}, m(\mathbf{X}_t)) \mid \mathbf{X}_t]$, observe that by Lemma 6.13, all but $O(\sqrt{N})$ arms in $[Nm(\mathbf{X}_t)]$ follow the optimal single-armed policies, so the drift condition of h_{ID} applies to this set of arms. As formalized in Lemma 6.16, we can thus show that

$$\mathbb{E}[h_{\text{ID}}(\mathbf{X}_{t+1}, m(\mathbf{X}_t)) \mid \mathbf{X}_t] \leq \gamma h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)) + (2 + K_{\text{conf}}) C_h \sqrt{N}. \quad (6.79)$$

Plugging (6.79) back to (6.78), we get

$$\begin{aligned} \mathbb{E}[V(\mathbf{X}_{t+1}) \mid \mathbf{X}_t] - V(\mathbf{X}_t) &\leq -(1 - \gamma) h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)) \\ &\quad + (2C_h + C_h K_{\text{conf}} + L_h K_{\text{mono}}) \sqrt{N}. \end{aligned} \quad (6.80)$$

To further bound $h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t))$ in (6.80) in terms of $V(\mathbf{X}_t)$, we apply Lemma 6.15 to get:

$$\begin{aligned} V(\mathbf{X}_t) &= h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)) + L_h N(1 - m(\mathbf{X}_t)) \\ &\leq \left(1 + \frac{L_h}{\eta_c}\right) h_{\text{ID}}(\mathbf{X}_t, m(\mathbf{X}_t)) + L_h K_{\text{cov}}. \end{aligned} \quad (6.81)$$

Substituting (6.81) into (6.80) and rearranging the terms, we get:

$$\mathbb{E}[V(\mathbf{X}_{t+1}) \mid \mathbf{X}_t] - V(\mathbf{X}_t) \leq -\rho_V V(\mathbf{X}_t) + K_V \sqrt{N}, \quad (6.82)$$

where $\rho_V = (1 - \gamma)/(1 + \frac{L_h}{\eta_c})$, and

$$K_V = 2C_h + C_h K_{\text{conf}} + 2L_h K_{\text{mono}} + \frac{\rho_V L_h K_{\text{cov}}}{\sqrt{N}}.$$

Taking the expectation in (6.82) and unrolling the recursion of $\mathbb{E}[V(\mathbf{X}_t)]$, we get

$$\mathbb{E}[V(\mathbf{X}_t)] \leq (1 - \rho_V)^t \mathbb{E}[V(\mathbf{X}_0)] + \frac{K_V \sqrt{N}}{\rho_V}.$$

Averaging over T time steps, we have

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[V(\mathbf{X}_t)] \leq \frac{\mathbb{E}[V(\mathbf{X}_0)]}{\rho_V T} + \frac{K_V \sqrt{N}}{\rho_V}. \quad (6.83)$$

Taking the limit of $T \rightarrow \infty$, we can bound the long-run-averaged expectation of $\mathbb{E}[V(\mathbf{x})]$ as

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[V(\mathbf{X}_t)] \leq \frac{K_V \sqrt{N}}{\rho_V}, \quad (6.84)$$

which completes the proof. $\square$

By examining the proof, we can get a more explicit form of the constant C_{ID} in Theorem 6.3:

- If $c_{\text{max}} < \alpha_{\text{min}}$, the budget constraints are vacuous and all arms will follow the ideal actions under the ID policy. Consequently, $R^*(N, \mathbf{S}_0) = R(\pi, \mathbf{S}_0)$ and we can simply take $C_{\text{ID}} = 0$.

- If $c_{\max} \geq \alpha_{\min}$, we combine the optimality gap bound in Lemma 6.9 with (6.84) to get

$$R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0) \leq \frac{(2r_{\max} + L_h)K_V}{L_h \rho_V \sqrt{N}} + \frac{2r_{\max}K_{\text{conf}}}{\sqrt{N}}.$$

We therefore obtain $R^*(N, \mathbf{S}_0) - R(\pi, \mathbf{S}_0) \leq C_{\text{ID}}/\sqrt{N}$, where C_{ID} is given by

$$\begin{aligned} C_{\text{ID}} &= \frac{(2r_{\max} + L_h)(1/L_h + 1/\eta_c)}{1 - \gamma} (2C_h + C_h K_{\text{conf}} + 2L_h K_{\text{mono}}) \\ &\quad + (2r_{\max} + L_h)K_{\text{cov}} + 2r_{\max}K_{\text{conf}} \\ &= O\left(\frac{K^5 \max\{r_{\max}, c_{\max}\}^7 \tau^4}{\alpha_{\min}^6}\right). \end{aligned}$$

In the last step above, we substituted in all intermediate constants from the proofs; one can verify that each of these constants depends on the problem parameters as follows:

$$\begin{aligned} L_h &= O(\max\{r_{\max}, c_{\max}\}\tau) \\ C_h &= O(K \max\{r_{\max}, c_{\max}\}\tau) \\ \frac{1}{1 - \gamma} &= O(\tau) \\ \frac{1}{\eta_c} &= O\left(\frac{K c_{\max}}{\alpha_{\min}^2}\right) \\ K_{\text{conf}} &= O\left(\frac{K^2 c_{\max}^2}{\alpha_{\min}^2}\right) \\ K_{\text{mono}} &= O\left(\frac{K^4 c_{\max}^3 \max\{r_{\max}, c_{\max}\}}{\alpha_{\min}^4} \tau\right) \\ K_{\text{cov}} &= O\left(\frac{K c_{\max} \max\{r_{\max}, c_{\max}\}}{\alpha_{\min}^2} \tau\right). \end{aligned}$$

6.I Proof of the finite-time bound

Now we briefly outline the proof of Proposition 6.5, which states that under the same conditions as Theorem 6.3, for any $T \geq 1$, we have

$$R^*(N, \mathbf{S}_0) - \frac{1}{TN} \sum_{t=0}^{T-1} \sum_{i \in [N]} \mathbb{E}[r_i(S_{i,t}^\pi, A_{i,t}^\pi)] \leq \frac{C_{\text{ID}}}{\sqrt{N}} + \frac{C_{\text{finite}}}{T}, \quad (6.85)$$

where C_{finite} is another positive constant independent of N . This follows from two steps:

- First, it is not hard to adapt the proof of Lemma 6.9 in Appendix 6.H.1 to show that

$$\begin{aligned} R^*(N, \mathbf{S}_0) - \frac{1}{TN} \sum_{t=0}^{T-1} \sum_{i \in [N]} \mathbb{E} [r_i(S_{i,t}^\pi, A_{i,t}^\pi)] \\ \leq \left(1 + \frac{2r_{\max}}{L_h}\right) \frac{1}{TN} \sum_{t=0}^{T-1} \mathbb{E}[V(\mathbf{X}_t)] + \frac{2r_{\max}K_{\text{conf}}}{\sqrt{N}}. \end{aligned} \quad (6.86)$$

Specifically, substituting (6.77) and (6.76) into (6.75) implies that for any $t \geq 0$, we have

$$R^{\text{rel}} - \frac{1}{N} \sum_{i \in [N]} \mathbb{E} [r_i(S_{i,t}^\pi, A_{i,t}^\pi)] \leq \left(1 + \frac{2r_{\max}}{L_h}\right) \frac{1}{N} \mathbb{E}[V(\mathbf{X}_t)] + \frac{2r_{\max}K_{\text{conf}}}{\sqrt{N}}. \quad (6.87)$$

Averaging the above inequality over $t = 0, 1, 2, \dots, T-1$ and applying $R^*(N, \mathbf{S}_0) \leq R^{\text{rel}}$, we have (6.86).

- Second, we apply (6.83) in the proof of Lemma 6.10 to get

$$\begin{aligned} R^*(N, \mathbf{S}_0) - \frac{1}{TN} \sum_{t=0}^{T-1} \sum_{i \in [N]} \mathbb{E} [r_i(S_{i,t}^\pi, A_{i,t}^\pi)] \\ \leq \left(1 + \frac{2r_{\max}}{L_h}\right) \left(\frac{\mathbb{E}[V(\mathbf{X}_0)]}{\rho_V TN} + \frac{K_V}{\rho_V \sqrt{N}} \right) + \frac{2r_{\max}K_{\text{conf}}}{\sqrt{N}} \\ = \left(1 + \frac{2r_{\max}}{L_h}\right) \frac{\mathbb{E}[V(\mathbf{X}_0)]}{\rho_V TN} + \frac{C_{\text{ID}}}{\sqrt{N}} \\ \leq \left(1 + \frac{2r_{\max}}{L_h}\right) \frac{L_h}{\rho_V T} + \frac{C_{\text{ID}}}{\sqrt{N}}. \end{aligned} \quad (6.88)$$

where the last inequality is because of $V(\mathbf{X}_0) \leq NL_h$ for any system state $\mathbf{X}_0$. Letting $C_{\text{finite}} = \left(1 + \frac{2r_{\max}}{L_h}\right) \frac{L_h}{\rho_V}$, we get (6.85).

Part II

One-Dimensional Structure: Multiserver Queues

This part of the thesis develops the *one-dimensional structure*, where a key quantity of the system can be approximated by a one-dimensional process even when the full state is high- or infinite-dimensional. Applied to multiserver queues, this yields a universal bound for $G/G/n$ (Chapter 7), near-optimality of the Gittins policy for $G/G/n$ with setup times (Chapter 8), and the best achievable delay together with a near-optimal policy for the multiserver-job model (Chapter 9). Below, we give an overview of each chapter in this part, in the order above. Each overview consists of a motivation followed by a discussion of results and impact.

Before turning to the individual chapters, we clarify the sense in which we measure the *order* of a quantity in this part, both in the big-O notation and in certain notions of asymptotic optimality. In each case, order describes how a quantity grows with the system’s scale parameters: the load ρ , the number of servers n , and, in the multiserver-job model, the server needs. What differs across the three chapters is whether these variables are left arbitrary or driven to a limit.

- In Chapter 7, ρ and n are *arbitrary*: our bound holds for every load $\rho < 1$ and every number of servers n , with an explicit constant factor that depends on the interarrival- and service-time distributions but not on ρ or n . Here $O(1/(1 - \rho))$ is merely shorthand that hides this constant; no limit is taken.
- In Chapter 8, both the big-O bounds and the asymptotic optimality result are stated in the limit $\rho \rightarrow 1$ with the number of servers n held fixed.
- In Chapter 9, both the big-O bounds and the asymptotic optimality result are stated in a joint limit where $\rho \rightarrow 1$ and $n \rightarrow \infty$, with the server needs also growing with n at a sublinear rate.

Universal bound for multiserver queues

Motivation

What is the problem and why do we care? The multiserver queue is one of the most foundational models in queueing theory: a stream of jobs

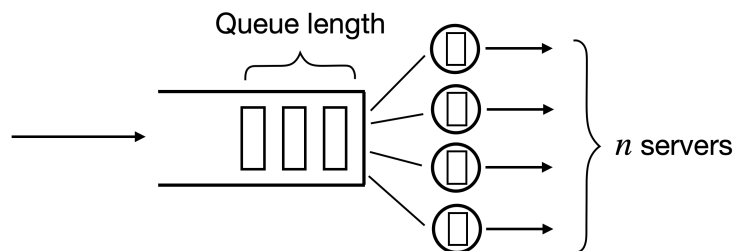


Figure II.1: The $G/G/n$ queue: jobs (rectangles) arrive over time and wait in a shared queue until one of the n servers (circles) becomes available. The quantity of interest is the mean queue length, i.e., the average number of waiting jobs, equivalently the mean waiting time.

arrives over time and is served in First-Come-First-Served order by a pool of n parallel servers (Figure II.1), and we ask how long the jobs wait, or equivalently, how many jobs are in the queue. It models systems ranging from call centers and hospitals to the server pools of a data center, and is a basic building block for more complex queueing models. Yet a fundamental question remains open: can we bound the mean queue length of the $G/G/n$ queue (more rigorously, the $GI/GI/n$ queue), the n -server queue with *general*, independent interarrival- and service-time distributions, by a *simple, accurate formula* that uses only a few parameters?

When the interarrival and service times are exponentially distributed (the $M/M/n$ queue), the queue is a one-dimensional Markov chain and is well understood. However, real systems rarely follow exponential distributions, and the waiting time can depend in complex ways on the interarrival- and service-time distributions. For the single-server queue, the gold standard is *Kingman's bound*, which controls the mean queue length using only the first two moments of the interarrival and service times. A multiserver analogue with the same simplicity would give practitioners a lightweight, interpretable formula for capacity planning, without simulation or heavy computation.

Prior work and what makes the problem hard Despite decades of effort, a Kingman-style bound for $G/G/n$ that holds *universally*, across all loads and server counts, has remained elusive. Most prior bounds hold only in specific limiting regimes, where the server count n and the load ρ scale together in a prescribed way. A recent exception [180] establishes a universal

bound of the right order, but with a constant so large that it is of little practical use.

The main challenge lies in the dimensionality of $G/G/n$. With general service times, the state of the system must track the remaining service time of every job in service, so the state is a continuous-discrete hybrid whose dimension is the number of jobs being served; when there are many servers, this becomes a high-dimensional process. This is in sharp contrast to the $M/M/n$ queue mentioned above, which is a one-dimensional Markov chain, and to the single-server queue, whose workload can be approximated by a one-dimensional reflected Brownian motion. It has been unclear whether any such one-dimensional simplification survives under general distributions when the number of servers is large.

Results and impact

Contribution 1: a new universal bound We establish a new bound on the mean queue length of the $G/G/n$ queue that is of the right order $O(1/(1 - \rho))$ and holds universally, for every load $\rho < 1$ and every number of servers n . Unlike prior universal bounds, ours carries small, interpretable constants, and for light-tailed service times it is often strikingly simple and tight: for Poisson arrivals and a broad class of service distributions, for instance, the mean queue length is at most $1/(1 - \rho)$, and for common families such as gamma-distributed service times the bound is close to exact. We refer the reader to Chapter 7 for the precise statements.

Contribution 2: a simpler proof Beyond the sharper constants, our analysis gives a much simpler and more transparent proof of the $O(1/(1 - \rho))$ universal bound for the mean queue length in the light-tailed case than was previously available.

What is new in our approach? Our key observation is that even when the number of servers is large, the mean queue length of the $G/G/n$ queue is well approximated by that of a certain $G/G/1$ queue. We prove this by establishing a certain *asymptotic independence* between the queue length and the servers' service processes. Our tool is a *leave-one-out coupling*, inspired by the leave-one-out arguments used in high-dimensional statistics.

Theoretical impact Our result makes substantial progress on a long-standing open problem: whether a simple, universal Kingman-style bound exists for the $G/G/n$ queue, a goal conjectured in much prior work and made explicit in Li and Goldberg [180, Chapter 3]. We prove such a bound, of the conjectured order $O(1/(1 - \rho))$ and with small constants. The one remaining gap is that our constant depends on the service distribution beyond its first two moments (roughly, on how light its tail is), whereas Kingman’s bound uses only the first two moments.

Practical impact The bound is directly useful for *capacity provisioning*, i.e., choosing the smallest number of servers that meets a target mean waiting time, because it is simple and, for many distributions, tight. Compared with simulation, the bound offers a qualitative advantage: it provides a *provable guarantee*, holding with certainty and for every load and server count simultaneously, rather than an estimate at only the finitely many configurations one simulates. Beyond this guarantee, the bound is also more efficient to apply than simulation in two ways. First, evaluating a closed-form bound is immediate, whereas simulation requires many long runs, especially for a very noisy system like $G/G/n$. Second, and more fundamentally, the bound depends on the interarrival- and service-time distributions only through a few parameters; when the distributions are unknown and must be learned from data, one then needs to estimate only these few parameters. This is far more data-efficient than fitting an entire distribution or collecting a long trace to simulate.

Future work There are many directions for future work. Within the $G/G/n$ model, two natural questions are: First, one would like bounds on higher moments and on the tail of the queue length, again with small constants; existing tail bounds with the correct $O(1/(1 - \rho))$ order carry the same extremely large constants that our mean bound avoids [180]. Second, our analysis currently requires light-tailed service times, and extending it to heavy-tailed service times is open.

Our work lays the foundation for more complex queueing models with multiple servers and general distributions. For instance, most analyses of load-balancing models assume exponential interarrival- and service-time distributions [e.g., 45, 47, 89, 90]. Non-exponential distributions have also been studied, but some of these works are restricted to specific distributions [e.g.,

189], while others are valid or tight only in particular limiting regimes [e.g., 6, 42, 126, 233]. Obtaining a universal bound for load-balancing models under general distributions remains open, and it would be interesting to see whether our techniques transfer.

Gittins policy for $G/G/n$ with setup times

Motivation

What is the problem and why do we care? The previous section bounds the mean queue length of the $G/G/n$ queue under First-Come-First-Served (FCFS), the policy that serves jobs in their arrival order. But FCFS is rarely the best choice: by simply modifying the *order* in which jobs are served, one can often reduce delay. The classic example arises when each job’s size (its service time on a unit-speed server) is known: the Shortest-Remaining-Processing-Time (SRPT) policy, which serves the job of least remaining size, minimizes the mean number of jobs in single-server queues. But real systems rarely match this ideal: job sizes are seldom known in advance; instead, one typically knows only the distribution they are drawn from, and learns more about a job’s remaining size only as it is served. In this case, SRPT no longer applies, and scheduling under unknown job sizes has become an active research topic. In this section, we study scheduling in complex queueing systems, specifically those with multiple servers, general (non-Poisson) arrivals, and server setup times.

A well-known policy for the unknown-size setting is the *Gittins index policy* (or simply “Gittins policy”). Gittins assigns each job a scalar *index* computed from its remaining-size distribution and serves the job of highest index. Gittins is known to be optimal in the $M/G/1$ queue, the single-server queue with Poisson arrivals [see, e.g., 112, 114, 243]. This optimality makes it a natural candidate for the more complex systems above, raising the question this section addresses: how well does Gittins perform beyond the $M/G/1$?

Prior work and what makes the problem hard Beyond the $M/G/1$, optimal scheduling is generally intractable, so the goal shifts from exact optimality to *heavy-traffic optimality*: a policy is heavy-traffic optimal if its optimality gap becomes negligible relative to the optimum as the load $\rho \rightarrow 1$ with the number of servers n held fixed. Prior work established that Gittins

is heavy-traffic optimal in the $M/G/n$ queue, showing that having multiple servers, on its own, does not spoil its performance. But having multiple servers is only one way in which real systems depart from the $M/G/1$. Two others are just as common: *general, non-Poisson arrivals*, and *setup times*, the warm-up delay a server incurs when it transitions from idle to busy. A realistic system may exhibit any combination of these three factors.

The challenge of analyzing $G/G/n$ with setup times lies in proving a *work-decomposition law*, which, roughly speaking, characterizes the steady-state mean work (the total server-time needed to complete all jobs in the system). Two complications arise when we generalize the work-decomposition law from the $M/G/n$ to the $G/G/n$ with setup times:

- First, with general interarrival times, the instantaneous arrival rate is no longer constant, but instead depends on the time elapsed since the previous arrival.
- Second, with setup times, a server may be warming up rather than serving, so the servers need not all work at full rate even when there are enough jobs to occupy them.

Because of these two complications, the work is no longer a self-contained Markov process: its dynamics also depend on the time until the next arrival and the servers' setup status. Such a multi-dimensional process is much harder to analyze and rarely admits a clean closed-form characterization.

Results and impact

Contribution 1: heavy-traffic optimality with all three factors We give the first analysis of Gittins that handles any combination of (a) multiple servers, (b) general arrivals, and (c) setup times, covering the $G/G/1$ and $G/G/n$ queues, with and without setup times. We bound the optimality gap in each case, and the bound decomposes additively: each of the three factors contributes its own term. Crucially, all of these terms are negligible in heavy traffic, so Gittins is heavy-traffic optimal in every one of these systems.

Contribution 2: exact optimality in the $M/G/1$ with setup times As a special case, our analysis yields a clean and previously unknown result that holds at *all* loads: Gittins is optimal among non-idling policies in the $M/G/1$ with setup times.

What is new in our approach? Our starting point is an identity known as WINE, which expresses the mean number of jobs in the system in terms of the *r-work* (a generalization of work) in the system, reducing the problem of bounding delay to the more tractable one of bounding *r-work*. Carrying out this strategy with general arrivals requires a *new work-decomposition law for general arrivals*, which is our main technical contribution.

The key step in proving the new work-decomposition law is to define a quantity that is almost a one-dimensional Markov process: the *r-work* corrected by a fraction of the residual arrival time. This quantity has nearly identical dynamics in the $G/G/n$ and the $G/G/1$, yielding the desired work-decomposition law.

The idea of tracking a workload-like quantity compensated by a residual timing term is not unique to our analysis. It traces back to Miyazawa’s stationary diffusion-approximation arguments [212] and the subsequent BAR approach [54], and closely related compensations underlie several recent analyses of queues with general distributions, including the universal bound for the $G/G/n$ queue under FCFS in Chapter 7, the diffusion-approximation-error analysis of the $G/G/1$ by Braverman and Scully [50], and the RESET technique of Grosz et al. [130]. Our contribution is to turn this idea into a work-decomposition law for general arrivals, which is what makes the Gittins analysis possible.

Theoretical impact Our work is the first to analyze optimal scheduling beyond the $M/G/n$, reaching systems with general arrivals and setup times. The new work-decomposition law for general arrivals could be of independent interest beyond this model.

Practical impact We identify a provably near-optimal policy, i.e., the Gittins index policy, for scheduling under unknown job sizes, multiple servers, general interarrival-time distributions, and server setup times. This offers practitioners a way to reduce delay for a wide range of practical systems, and requires only computing the Gittins index for the given job-size distribution.

Broader applicability The three factors we treat are only examples, and our proof framework can be extended to incorporate further complexities into the multiserver queue: *multiserver jobs*, which must be served by several servers simultaneously and are common in cloud computing; *batch arrivals*,

in which jobs arrive in groups rather than one at a time; and *generalized vacations*, which capture the many ways a server can become temporarily unavailable.

Multiserver jobs

Motivation

What is the problem and why do we care? So far in this part, each job has occupied a single server. We now turn to *multiserver jobs*, which occupy several servers *simultaneously* throughout their service. Such jobs are ubiquitous in modern cloud computing clusters. For example, a large parallel computation may reserve many CPUs or GPUs at once and hold them until it finishes. The *server needs* (the number of servers that the jobs require) of the jobs are highly heterogeneous, spanning many orders of magnitude. This heterogeneity makes multiserver-job systems behave very differently from the single-server-job queues studied earlier in this part. It is then natural to ask how to schedule these jobs. Specifically, we study the following three questions: how well does the standard First-Come-First-Served (FCFS) policy perform? Can any policy do better? What is the best achievable mean waiting time?

Model and goal The model consists of n servers with a central queue. Jobs of several types arrive over time; each type is associated with an arrival rate, a service rate, and a server need; a job in service occupies its server-need servers simultaneously until it completes. We study this system in the *large-system regime*, in which the number of servers n grows large while the arrival rates and server needs scale with it. This regime captures the two defining features of large clusters: many servers, and jobs whose server needs vary over a wide range. Our goal is to characterize the *mean waiting time* under different policies, and to find a policy that is *order-wise optimal* as n scales — that is, whose mean waiting time matches, up to a constant factor, the best achievable by any policy.

Prior work and what makes the problem hard Most prior work on multiserver jobs considers *stability* and *throughput*, or focuses on very special cases such as the two-server case. The *delay* of multiserver-job models in

the *large-system regime* is relatively understudied. Specifically, the two most closely related works each capture only one of the two ingredients we need: delay, or the large-system regime. Concurrent work [128, 129] does study multiserver-job *delay*, but in the *traditional* heavy-traffic regime, where the number of servers is fixed and the load approaches one, in contrast to our large-system regime, where the number of servers and the server needs scale along with the load. Wang et al. [271] also study the large-system regime, but they focus on the *queueing probability* (the probability that an arriving job has to wait at all), which cannot distinguish the performance of different policies.

The challenge is rooted in the heterogeneity among job types, in both their service rates and their server needs. This heterogeneity makes the system dynamics multidimensional: unlike a classical queue, neither the number of jobs in service nor the number of busy servers determines the rate at which jobs depart, so the system does not reduce to a simple low-dimensional Markov chain. Indeed, even the classical multi-class $M/M/n$ queue, which has heterogeneous service rates but unit server needs, is already hard to schedule optimally, and heterogeneous server needs add a further layer of intricacy.

Results and impact

Contribution 1: the mean waiting time under FCFS We first analyze the First-Come-First-Served (FCFS) policy and pin down the precise order of its mean waiting time as the system scales.

Contribution 2: a lower bound for all policies We then prove a lower bound on the mean waiting time that holds for *every* scheduling policy.

Contribution 3: a simple order-wise-optimal policy Finally, we exhibit a policy that attains the lower bound: *Smallest-Need-First (SNF)*, a simple static priority that preemptively serves the jobs of smallest server need first. Because SNF meets the lower bound, it is *order-wise optimal* in this regime. Combined with the FCFS analysis, it shows that FCFS is *strictly suboptimal*.

Intuitively, prioritizing smallest-need jobs is analogous to favoring shortest service times in the single-server setting, and it works by preventing large

jobs from blocking small ones. One might instead expect a packing heuristic that favors the *largest*-need jobs, to use the servers efficiently; but in the large-system regime we consider, servers are abundant relative to any single job’s need, so packing is not the bottleneck. What matters is the order in which jobs are served.

What is new in our approach? The first step of our analysis is to establish *state-space concentration*: in steady state, the system concentrates on a small subset of its state space. This phenomenon has been observed, case by case, in other heavy-traffic and large-system settings [see, e.g., 189, 272, 277]; we establish it for the multiserver-job model by coupling with infinite-server systems.

Restricting attention to this subset lets us find a single one-dimensional quantity, the *workload*, that admits a simple characterization. This characterization then drives all three results above.

Theoretical impact The workload bounds and the state-space-concentration arguments behind them hold for a broad class of policies, not only FCFS and SNF, and so may be of independent interest for analyzing other scheduling policies or other variants of the multiserver-job model.

Practical impact On the practical side, SNF is a simple policy that achieves order-wise lower delay than FCFS, especially when the server needs are highly heterogeneous. Because it prioritizes jobs using only their server needs, which the scheduler typically already knows, SNF is potentially useful in practice as a low-overhead alternative to the default FCFS.

Chapter 7

An Interpretable Universal Bound for Multiserver Queues via a Leave-One-Out Technique

7.1 Introduction

The $GI/GI/n$ queue is one of the most fundamental multiserver queueing models, with a wide range of applications [see, e.g., 288]. The $GI/GI/n$ queue consists of n servers and a central queue, where external jobs arrive over time and wait in the queue for service in the first-come, first-served order. When a server frees up, it picks the first job in the queue and starts serving it until the job is completed. An illustration of the $GI/GI/n$ queue is provided in Figure 7.1. The name “GI” refers to “general” and “independent”, indicating that the interarrival times and service times are two i.i.d. sequences, where elements of each sequence follow a general distribution.

In the theory of the $GI/GI/n$ queue, a natural goal is to find a simple formula that predicts the *queue length* (the number of jobs in the queue waiting for service). A prominent line of work focuses on finding accurate bounds or approximations when the system *scales up* in a certain sense. This pursuit has been remarkably successful for the single-server case, dating back to the seminal work of Kingman [171]. Kingman derives a simple upper bound for the steady-state mean queue length of the $GI/GI/1$ queue, $\mathbb{E}_\pi[Q]$:

$$\mathbb{E}_\pi[Q] \leq \frac{\rho^2 \text{Var}[\mu S] + \text{Var}[\lambda A]}{2(1 - \rho)}, \quad (7.1)$$

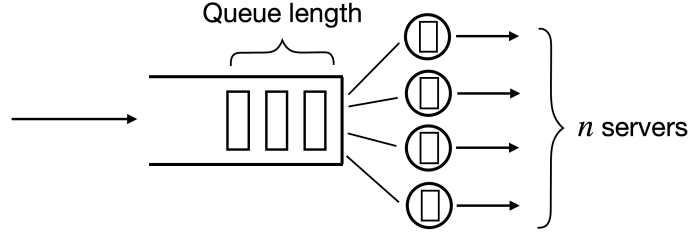


Figure 7.1: An illustration of the $GI/GI/n$ queue, where each circle denotes a server, and each rectangle denotes a job.

where S is the service time, A is the interarrival time, $\mu \triangleq 1/\mathbb{E}[S]$, $\lambda \triangleq 1/\mathbb{E}[A]$, and $\rho \triangleq \lambda/\mu$ is the *load*. Despite its simplicity, Kingman's bound is surprisingly accurate. In particular, it is asymptotically tight in the so-called *heavy-traffic* scaling regime where $\rho \uparrow 1$ [170], so it precisely captures the queue's behavior in the scenario where the queueing delay becomes significant.

In contrast to the single-server case, finding a similarly simple and accurate upper bound for the $GI/GI/n$ queue is a much greater challenge. This complexity arises from a rich set of combinations of two basic scalings: the *heavy-traffic* scaling, where the load $\rho \uparrow 1$, and the *many-server* scaling, where the number of servers $n \rightarrow \infty$. In practice, these two parameters often scale jointly to balance utilization and delay [19, 39, 104, 199, 280]. Different joint scaling rates give rise to dramatically different asymptotic behaviors. For example, in the classical heavy-traffic regime ($\rho \uparrow 1$ with fixed n), the limiting queue length distribution of the $GI/GI/n$ queue is similar to that of $GI/GI/1$, depending only on the first two moments of the input distributions [see, e.g., 173]. In contrast, in the Halfin-Whitt (HW) regime ($n \rightarrow \infty$ and $\rho = 1 - \Theta(1/\sqrt{n})$) or the Non-Degenerate Slowdown (NDS) regime ($n \rightarrow \infty$ and $\rho = 1 - \Theta(1/n)$), the limiting behavior is far more complex and can depend intricately on the full input distributions [see, e.g., 7, 20, 49, 232].

A long-standing challenge has been to establish a queue length bound of the order $O(1/(1 - \rho))$ that holds *universally* for all $\rho < 1$ and n . Achieving this is significant because this order is known to be tight in a broad class of important regimes, including the classical heavy-traffic, NDS, and Halfin-Whitt [7, 20, 173]. However, most prior results are tailored to specific scaling regimes, making them either invalid or loose outside of the specific regimes. A universal bound of the order $O(1/(1 - \rho))$ has remained elusive until the

recent breakthrough by Li and Goldberg [180], which provides the first such result, stated in their Corollary 2 as:

$$\begin{aligned} \mathbb{E}_\pi[Q] &\leq \frac{2.1 \times 10^{21} \mathbb{E}[(\mu S)^2] \left(\mathbb{E}[(\mu S)^2]^{1+\epsilon} + \mathbb{E}[(\mu S)^{2+\epsilon}] \right) \left(\frac{1}{\epsilon} \right)^4 + 49 \mathbb{E}[(\Lambda A)^2]}{1 - \rho}, \end{aligned} \quad (7.2)$$

where $\Lambda = 1/\mathbb{E}[A]$ and $\epsilon \in (0, 0.5)$. In addition to this bound on the mean queue length, Li and Goldberg [180] also establishes several high-order moment and tail bounds of the same order.

Although Li and Goldberg [180] shows the existence of a universal bound of the correct order, $O(1/(1 - \rho))$, the numerical constant in their bound is astronomically large. This stands in stark contrast to Kingman's bound for the single-server case, suggesting room for improvement. Indeed, Li and Goldberg [180] themselves conjecture that Kingman's bound may hold for any $M/GI/n$ queue. This leads to the central question of our chapter: *Can we prove a universal $O(1/(1 - \rho))$ bound for the $GI/GI/n$ queue with a small and interpretable constant factor?*

Summary of contribution We answer this question affirmatively. Our main contributions are as follows.

- *A new technique.* We introduce a *leave-one-out coupling* technique that overcomes a long-standing obstacle to analyzing the $GI/GI/n$ queue.
- *A simple and transparent proof.* Our proof is significantly simpler than that of Li and Goldberg [180]: it relies on a short core argument and makes the origin of each term in the bound transparent.
- *A small and interpretable constant.* For light-tailed service times, our leading constant is interpretable and orders of magnitude smaller than that of prior work. A comparison of the constant factors is given in Table 7.1, with the details derived in Section 7.4.
- *Heterogeneous servers.* We extend the method to the $GI/GI/n$ queue with *fully heterogeneous* service-time distributions, a setting not addressed by prior universal bounds (Section 7.8).

Our results We prove that, in the $GI/GI/n$ queue, the steady-state mean queue length satisfies

$$\mathbb{E}_\pi[Q] \leq \frac{2R_s^{\max} + \rho \text{Var}[\Lambda A] - \rho}{2(1 - \rho)} + \left(\frac{1}{2} \mathbb{E}[(\Lambda A)^2] - R_a^{\min} \right), \quad (7.3)$$

where $R_s^{\max} \triangleq \sup_{t \geq 0} \mathbb{E}[\mu S - t \mid \mu S \geq t]$ (finite for a large class of light-tailed distributions), and the second term $\mathbb{E}[(\Lambda A)^2]/2 - R_a^{\min}$ is a constant determined by the unitized arrival-time distribution. This bound holds assuming $\mathbb{E}[A^2] < \infty$, $R_s^{\max} < \infty$ (Assumption 7.1), S is non-lattice (Assumption 7.2), and a certain positive Harris recurrence condition (Assumption 7.3). When we consider the $M/GI/n$ queue, the bound in (7.7) can be further simplified into:

$$\mathbb{E}_\pi[Q] \leq \frac{R_s^{\max}}{1 - \rho}. \quad (7.4)$$

The above two bounds are given in Section 7.4 as direct corollaries of a more refined bound (Theorem 7.5), which has a better leading constant when $n = o(1/(1 - \rho))$. The refined bound mirrors the structure of Kingman's classical single-server bound. Its leading term recovers Kingman's bound for the $GI/GI/1$ queue that has interarrival time A and service time S/n . The remaining terms, governed by the maximal mean residual service time $R_s^{\max}$, quantify the multiserver boundary effect. We elaborate on this decomposition in Section 7.4.

Our bounds hold universally for all $\rho < 1$ and n and have the order $O(1/(1 - \rho))$. Compared with the bounds in [180], our bounds require stronger assumptions on the service-time distribution, but often have much smaller constants when the assumptions are satisfied. For example, in the $M/GI/n$ queue whose service times follow the gamma distribution with the shape parameter α (i.e., with the squared coefficient of variation $1/\alpha$), $R_s^{\max} = \max(1, 1/\alpha)$, so (7.4) implies $\mathbb{E}_\pi[Q] \leq \max(1, 1/\alpha)/(1 - \rho)$. Table 7.1 previews these leading constants for several standard distribution classes, including NBUE service (whose mean residual life never exceeds its mean, a class containing all increasing-failure-rate distributions), phase-type, and bounded service times; we derive them in Section 7.4.

Our results also generalize to the $GI/GI/n$ queue with *fully heterogeneous* servers. The details can be found in Section 7.8.

Proof techniques On an intuitive level, our proof proceeds by quantifying the discrepancy between the $GI/GI/n$ queue and a $GI/GI/1$ queue. This

Table 7.1: Leading constants for the $M/GI/n$ queue.

Service distribution	Li and Goldberg [180]	This chapter
Exponential ($M/M/n$)	$2 \times 10^{21} \kappa$	1
NBUE	$2 \times 10^{21} \kappa$	1
Gamma with shape α	$2 \times 10^{21} \kappa$	$\max(1, 1/\alpha)$
Phase-type with d phases	$2 \times 10^{21} \kappa$	$\mu \max_{k \in [d]} \tau_k$
Bounded by $S_{\max}$	$2 \times 10^{21} \kappa$	$\mu S_{\max}$

Note. The leading constant C in the bound $\mathbb{E}_\pi[Q] \leq C/(1-\rho)$ for the $M/GI/n$ queue. Our column follows from (7.4) (and Section 7.4); for phase-type service, τ_k is the mean residual service time in phase k . The Li and Goldberg [180] column is the coefficient of $1/(1-\rho)$ in (7.2), where κ is a distribution-dependent factor (differing across rows) involving the moments of μS , ΛA , and a parameter $\epsilon \in (0, 0.5)$. For every distribution listed, the factor κ alone exceeds the corresponding entry in the last column. The bounds of Li and Goldberg [180] also hold under weaker (finite-moment) assumptions; the table compares the resulting leading constants where both apply.

$GI/GI/1$ queue has an interarrival time distribution A and a service time distribution S/n ; intuitively, it serves the same sequence of jobs using a “super-server” that is n times faster than each server in $GI/GI/n$. Observe that the $GI/GI/n$ matches the “speed” of the corresponding $GI/GI/1$ when all of its n servers are busy, and is slower otherwise. Because this loss of speed is restricted to periods when $GI/GI/n$ has idle servers and an empty queue, one might expect the queue length of $GI/GI/n$ to be comparable to that of the corresponding $GI/GI/1$.

The formal proof consists of three steps:

- First, we construct a modified $GI/GI/n$ queue whose queue length stochastically dominates that of the original system.
- Second, we derive an identity that decomposes the mean queue length of this modified system into two components: the mean queue length of the $GI/GI/1$ and a set of “covariance” terms that capture the consequence of the “loss of speed” when the queue length hits zero.
- Finally, we bound these covariance terms.

The first two steps are based on known techniques. The modified $GI/GI/n$ is a standard construction in the literature, recently adopted by Li and Goldberg [180] to obtain the first $O(1/(1-\rho))$ bound. In the second step, the decomposition of the mean queue length is derived using the Basic Adjoint

Relationship (BAR) [see, e.g., 52, 54, 210, 212]. Our technical innovation lies in the third step, where we introduce a novel *leave-one-out coupling* to control the covariance terms. At a high level, this technique hypothetically removes one server and compares the resulting counterfactual queue length with the original. This comparison quantifies the correlation between the queue length and the removed server's service times. We note that similar covariance terms have appeared in previous analyses of $GI/GI/n$ and have been a major obstacle to obtaining a universal $O(1/(1 - \rho))$ bound [128, 145, 148, 245]. The name of our technique reflects an analogy with leave-one-out arguments in statistics and learning theory [see, e.g., 81, 175, 250].

A detailed proof sketch is provided in Section 7.5, and a roadmap to the rigorous proof can be found in Section 7.5.4.

Chapter organization We first review the related work in Section 7.2. In Section 7.3, we formally set up the model, introduce the modified $GI/GI/n$ queue, and state the assumptions about these systems. In Section 7.4, we state our main results and provide some examples. In Section 7.5, we provide an informal overview and a roadmap for the formal proof. In Section 7.6, we formally set up the analysis framework and state some supporting lemmas. In Section 7.7, we prove the main results. In Section 7.8, we generalize our results to the $GI/GI/n$ queue with heterogeneous servers. Finally, we conclude the chapter in Section 7.9.

We organize the proof so that its ideas and its formal development can be read independently. Section 7.5 is self-contained, so readers mainly seeking these ideas may read it and skip the formal development in Sections 7.6 and 7.7.

General notation For any two real numbers a, b , we use $a \wedge b$ to denote their minimum. For any positive integer k , we let $[k] \triangleq \{1, 2, \dots, k\}$. We let $\mathbb{R}$ be the set of real numbers, let $\mathbb{R}_{\geq 0}$ be the set of non-negative real numbers, let $\mathbb{Z}_{\geq 0}$ be the set of non-negative integers, and let $\mathbb{Z}_{\geq 1}$ be the set of positive integers. We use standard asymptotic notation for non-negative functions f and g : $f = O(g)$ if $\limsup f/g < \infty$; $f = o(g)$ if $f/g \rightarrow 0$; $f = \Omega(g)$ if $g = O(f)$; $f = \omega(g)$ if $f/g \rightarrow \infty$; and $f = \Theta(g)$ if $f = O(g)$ and $g = O(f)$. Unless otherwise stated, these limits are taken as $\rho \rightarrow 1$ or $n \rightarrow \infty$, with the underlying unitized distributions $S/\mathbb{E}[S]$ and $A/\mathbb{E}[A]$ fixed. We denote probability, expectation, and variance by $\mathbb{P}[\cdot]$, $\mathbb{E}[\cdot]$, and $\text{Var}[\cdot]$, respectively.

The subscripts of these three symbols specify the underlying probability distributions: for example, $\mathbb{P}_\nu[\cdot]$ and $\mathbb{E}_\nu[\cdot]$ are taken under the distribution ν . When the distribution is clear from context, we omit the subscript.

7.2 Related work

General prior work on the $GI/GI/n$ queue There has been a rich body of literature on the queue-length scaling of the $GI/GI/n$ queue. Prior to Li and Goldberg [180], most papers focus on a certain representative or useful scaling regime of ρ and n , without providing general bounds that hold for arbitrary ρ and n . In the following, we briefly review these papers and refer the readers to [180] for a comprehensive review.

The most well-studied regime is the classical heavy-traffic regime, where n is fixed, and $\rho \uparrow 1$ [e.g., 38, 128, 145, 169, 173, 174, 192, 213]. In this regime, relatively simple bounds or approximations are available, and the queue length is known to be $O(1/(1-\rho))$ under proper moment assumptions. Some of these papers contain bounds that hold beyond the classical heavy-traffic regime, but they grow faster than order $1/(1-\rho)$ when n is large [e.g., 128, 145, 174, 192, 213]. For example, the bound proved by [128] has order $O(1/(1-\rho) + n)$.

More recent papers also study the Halfin-Whitt regime (HW), where $n \rightarrow \infty$ and $\rho = 1 - \Theta(1/\sqrt{n})$ [e.g., 7, 44, 48, 49, 51, 53, 71, 75, 97–99, 133, 135, 161–164, 198, 199, 229, 232], or the Non-Degenerate Slowdown (NDS) regime, where $n \rightarrow \infty$ and $\rho = 1 - \Theta(1/n)$ [e.g., 19, 20, 133]. The queue length in these regimes has been shown to be of the order $O(1/(1-\rho))$. However, a key difference from the classical heavy-traffic regime is that the analysis for the HW and NDS regimes is typically asymptotic, focusing on the system’s limiting behavior as $n \rightarrow \infty$. Although some work provides convergence rates [e.g. 49] or proves the tightness of properly scaled queue lengths [e.g. 75, 97–99], the bounds are generally *implicit* except for certain special cases (e.g., exponential service times [48, 51, 99, 231], deterministic service times [161, 163], or lattice service times [98]). It is thus difficult to translate these results into the type of explicit, non-asymptotic performance bounds that are the focus of our work.

Some papers also consider scaling regimes with load further lighter than HW, i.e. $\rho = 1 - \omega(1/\sqrt{n})$ [e.g. 48, 51, 133, 148, 231, 271]. Such regimes are referred to as *sub-Halfin-Whitt regimes* (sub-HW). In the sub-HW, the

mean queue length of the $GI/GI/n$ queue could be exponentially smaller than $1/(1 - \rho)$, although the precise order of the mean queue length is only known for very special models, such as the $M/M/n$ queue [see, e.g. 48, 231].

Finally, apart from [180], some papers also provide results that hold across regimes, but under highly specialized distributional assumptions. Some of them assume exponential service times and provide approximations with universal error bounds [e.g., 48, 51, 133, 231]. Chapter 9 [148] provides an $O(1/(1 - \rho))$ bound for some regimes between HW and NDS, assuming Poisson arrivals and hyperexponential service times.

Basic Adjoint Relationship and similar methods Basic Adjoint Relationship (BAR), also known as the *stationary equation* in the earlier literature [211], is a set of equations that characterize the stationary distribution of Markov processes. BAR allows us to extract information about the stationary distribution of a Markov process by taking suitable test functions. BAR and BAR-based analytical techniques have been studied in the literature under different names, such as the drift method [see, e.g., 89] or the rate conservation law [see, e.g., 210, 212]. In this chapter, we use BAR to refer to the particular framework in [52, 54], which handles Markov processes with both continuous and discrete state changes, utilizing the Palm distribution. This BAR framework has recently been applied to various queueing systems with general interarrival-time and service-time distributions under heavy traffic [e.g., 50, 52, 54, 70, 72, 73, 130, 131, 146, 212].

Some prior work has derived queue length bounds for the $GI/GI/n$ queue based on BAR or equivalent approaches [e.g. 128, 145, 148, 245]. These bounds typically have small constant factors in front of the $1/(1 - \rho)$ -order term, but they also contain additive error terms that depend on n . These additive error terms cause these bounds to grow faster than order $1/(1 - \rho)$ when n is large, reflecting a common technical challenge in analyzing the $GI/GI/n$ queue using BAR-based approaches. We will discuss this challenge in more detail in Section 7.5.

Leave-one-out techniques Our core technical innovation, the leave-one-out technique, is similar in spirit to the leave-one-out cross-validation in statistics [175], as well as the related methods in learning theory [250] and matrix completion [81]. On a high level, these approaches quantify the

influence of a single input on a stochastic system by analyzing the sensitivity of the outcome to that component's removal or replacement.

7.3 Problem setup

In this section, we define two queueing systems with homogeneous servers: the $GI/GI/n$ queue, whose queue length we want to upper bound, and a proxy system named *modified $GI/GI/n$ queue*, which will be the focus of our analysis.

7.3.1 $GI/GI/n$ queue

We consider the multiserver queue with general independent interarrival-time and service-time distributions, also known as the $GI/GI/n$ queue, under the first-come, first-served (FCFS) scheduling policy. As illustrated in Figure 7.1, this queueing system consists of a central queue and n identical servers, serving externally arriving jobs with independent and identically distributed (i.i.d.) interarrival times. Each job joins the end of the queue if all servers are busy, or starts service immediately if there are any idle servers. After entering the service, the job picks an idle server uniformly at random, occupies the server for a random amount of service time, and leaves the system after completing the service. The service times of the jobs are also i.i.d. When a server completes a job, it either immediately takes the next job from the head of the queue to serve it if the queue is non-empty, or becomes idle if the queue is empty. If multiple servers complete the service simultaneously, we assign the head-of-line jobs to these servers uniformly at random.

The $GI/GI/n$ system starts running at the time $t = 0$, possibly with some initial jobs in the queue and in service. We allow arbitrary initial queue length (denoted as $Q(0)$), time until first arrival, the set of initially busy servers, and the residual service times of the busy servers. These data describe the status of the system at the time $t = 0$, and we refer to them collectively as the *initial state*.

The $GI/GI/n$ queue is driven by two independent i.i.d. sequences of random variables, which represent the interarrival times and service times. Specifically, under the FCFS policy, we can index all jobs by the order in which they begin service. For the k -th job that begins service after time 0, we let S^k denote its service time and let A^k denote the interarrival time between

the k -th and $(k+1)$ -th jobs. Note that $\{A^k\}_{k=1}^{Q(0)}$ correspond to jobs already in the queue at time 0, so they do not play a role in the analysis and can take arbitrary values. By the definition of the $GI/GI/n$ queue, the interarrival times $\{A^k\}_{k=Q(0)+1}^{\infty}$ and the service times $\{S^k\}_{k=1}^{\infty}$ are independent, and each sequence is i.i.d. with its own distribution. We represent the distributions of A^k and S^k using two generic random variables A and S , respectively, and simply refer to them as the *interarrival time* and the *service time*. The *arrival rate* Λ and *service rate* μ of the system can therefore be defined as $\Lambda \triangleq 1/\mathbb{E}[A]$ and $\mu \triangleq 1/\mathbb{E}[S]$. We also define the *scaled arrival rate* λ to be $\lambda \triangleq \Lambda/n$, which represents the average arrival rate processed by each server. To characterize the tails of the unitized service time μS and the unitized interarrival time ΛA , we consider the supremum of μS 's mean residual time $R_s^{\max} \triangleq \sup_{t \geq 0} \mathbb{E}[\mu S - t \mid \mu S \geq t]$, and the infimum of ΛA 's mean residual time $R_a^{\min} = \inf_{t \geq 0} \mathbb{E}[\Lambda A - t \mid \Lambda A \geq t]$.

Next, we state two assumptions on the distributions of A and S .

Assumption 7.1. We assume that $\mathbb{E}[A^2] < \infty$ and $R_s^{\max} < \infty$.

Assumption 7.2. We assume that the distribution of S is non-lattice, i.e., the support of S is not a subset of $\{\delta, 2\delta, 3\delta, \dots\}$ for any $\delta > 0$.

Assumption 7.1 requires that the interarrival time has a finite second moment, and the service time S has a bounded mean residual time. The assumption $R_s^{\max} < \infty$ has also been used in [85, 128, 213]. As shown in [85], $R_s^{\max} < \infty$ implies that the tail CDF of S decays at least exponentially and hence that all moments of S are finite; for completeness, we also provide a proof in Appendix 7.A. The mean residual time is bounded for many common classes of light-tailed distributions, including the phase-type, Increasing Failure Rate (IFR), New Better Than Used in Expectation (NBUE), gamma distributions, etc. As previewed in Table 7.1, we compute $R_s^{\max}$ for several of these examples in Section 7.4.

Assumption 7.2 can be roughly viewed as a continuous-time analogue of the aperiodicity condition for renewal processes with the inter-event distribution S . The non-lattice condition is required for some common versions of renewal theorem [16, Theorem V.4.3], which we will use in the proof of Lemma 7.15.

The $GI/GI/n$ queue is a Markov process whose *state* at time t can be represented as

$$X(t) \triangleq (R_{s,1}(t), R_{s,2}(t), \dots, R_{s,n}(t), R_a(t), Q(t)), \quad (7.5)$$

where $Q(t)$ denotes the *queue length* (not including the jobs in service), $R_a(t)$ denotes the *residual arrival time*, and $R_{s,i}(t)$ denotes the *residual service time*

of the i -th server, for $i = 1, 2, \dots, n$. We use the convention that all the state variables have right-continuous sample paths, and let $R_{s,i}(t) = 0$ if the i -th server is empty at time t . The set of all possible realizations of $X(t)$ constitutes the *state space* of the $GI/GI/n$ queue, denoted as $\mathcal{X} \triangleq \mathbb{R}_{\geq 0}^{n+1} \times \mathbb{Z}_{\geq 0}$.

The load of the $GI/GI/n$ system is defined as $\rho \triangleq \Lambda/(n\mu) = \lambda/\mu$. We assume that $\rho < 1$, which implies that the system is positive Harris recurrent [16, Theorem XII.2.2].

7.3.2 Modified $GI/GI/n$ queue

Now we set up the *modified $GI/GI/n$ queue*, which we will focus on in our analysis as a proxy of the original $GI/GI/n$ queue. On a high-level, the modified $GI/GI/n$ queue adds virtual jobs to the $GI/GI/n$ queue to keep the servers busy. Specifically, when the queue is empty and a server is about to idle, the server immediately begins to serve a virtual job with a service time independently sampled from the distribution of S . The server then works non-preemptively on this virtual job, becoming available to serve the next job only after completing the virtual job. We initialize the modified $GI/GI/n$ queue with the same set of jobs as the original system, supplemented by the virtual jobs assigned to any servers that would otherwise be idle at the time $t = 0$.

The modified $GI/GI/n$ queue is a well-established construction in the queueing literature [38, 60, 135, 159, 165, 281]. Under the coupling described above, the queue length distribution of the modified system stochastically dominates that of the original $GI/GI/n$ at all times $t \geq 0$. This dominance, which was rigorously established in [97, Proposition 1], ensures that any upper bound derived for the modified queue length applies automatically to the original system. Following the approach of Li and Goldberg [180], we therefore focus our analysis exclusively on the modified $GI/GI/n$. For completeness, we provide a self-contained proof of this stochastic dominance in Appendix 7.C.

A nice property of the modified $GI/GI/n$ queue is that its sample path is *determined* by $(n+1)$ mutually independent renewal processes [97, Corollary 1]: the service-completions of each server are a renewal process with the inter-event distribution S , and the arrivals are a renewal process with the inter-event distribution A .

The modified $GI/GI/n$ queue is a Markov process whose state can be represented by the same tuple $X(t)$ as the original $GI/GI/n$ queue (7.5),

with the state space $\mathcal{X} = \mathbb{R}_{\geq 0}^{n+1} \times \mathbb{Z}_{\geq 0}$. The dynamics of this Markov process is summarized below:

- **Continuous dynamics** At any time t , the residual times (i.e., forward recurrence times) of the renewal processes, $R_a(t)$ and $R_{s,i}(t)$ for $i \in [n]$, decrease at the constant rate 1.
- **Discrete events** When one or more residual times reach zero at time t , discrete events are triggered. Each type of event corresponds to a subroutine that maps an input state, $X^- = (R_{s,1}^-, \dots, R_{s,n}^-, R_a^-, Q^-)$, to an output state, $X^+ = (R_{s,1}^+, \dots, R_{s,n}^+, R_a^+, Q^+)$, according to the following rules:
 - (i) **Arrival event** ($R_a(t-) = 0$): $R_a^+ = R_a^- + A$ for some interarrival time A independent of X^- ; $Q^+ = Q^- + 1$; other coordinates of the state remain unchanged.
 - (ii) **Completion event at server i** ($R_{s,i}(t-) = 0$): $R_{s,i}^+ = R_{s,i}^- + S$ for some service time S independent of X^- ; $Q^+ = Q^- - \mathbf{1}\{Q^- > 0\}$; other coordinates of the state remain unchanged.

Typically, only one event happens at a time, and the post-event state $X(t)$ is generated by applying the corresponding subroutine to the pre-event state $X(t-)$. In rare cases where multiple events occur simultaneously, the subroutines are applied sequentially. Different orders of applying the subroutines result in different system dynamics, so we stick to a fixed order to avoid ambiguities: we first apply the arrival subroutine if there is an arrival event, and then any completion subroutines in increasing order of their server indices.

Next, we make a stability assumption for the modified $GI/GI/n$ queues. Note that in addition to the main modified $GI/GI/n$ queue under study, we will also construct a few more auxiliary instances of modified $GI/GI/n$ queues (Section 7.6.1); these auxiliary systems have different numbers of servers but same interarrival-time and service-time distributions as the main modified $GI/GI/n$ queue under study. Our assumption should also hold for all these systems; the load of each of these systems is defined as the ratio of its arrival rate and the total service rate of its servers.

Assumption 7.3. Each modified $GI/GI/n$ queue with homogeneous servers considered in this chapter is positive Harris recurrent if its load is less than 1. In addition, let ν be its unique stationary distribution, then $\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \mathbb{E}[Q(t)] dt = \mathbb{E}_\nu[Q] < \infty$.

Remark 7.4. Assumption 7.3 has two parts: the positive Harris recurrence of the modified $GI/GI/n$ queue, and the finiteness of its steady-state mean queue length. Essentially, only the first part is restrictive, as the finiteness of the steady-state mean queue length could be implied by the existing queue length bounds for modified $GI/GI/n$ queues. In particular, Corollary 2 in [180] holds for any modified $GI/GI/n$ queue with load less than 1 that satisfies positive Harris recurrent, aperiodicity, and some standard finite moment assumptions implied by Assumption 7.1. We will discuss Assumption 7.3 in more detail in Appendix 7.B, where we argue that certain regularity conditions on the distributions of A and S are likely to imply Assumption 7.3.

7.4 Main results

In this section, we state our main results for the $GI/GI/n$ queue with homogeneous servers under Assumptions 7.1, 7.2 and 7.3, and provide some examples. We will also compare our results with those of [180]. We will present generalizations to the heterogeneous setting in Section 7.8.

We begin with our main bound, presented in Theorem 7.5 below. This bound is a more refined version of the bounds (7.3) and (7.4) from the introduction.

Theorem 7.5. *Consider the $GI/GI/n$ queue with homogeneous servers under the FCFS policy, whose service-time and interarrival-time distributions satisfy Assumptions 7.1, 7.2 and 7.3. When the load $\rho < 1$, let π be the unique stationary distribution of the system. Then the steady-state mean queue length, $\mathbb{E}_\pi[Q]$, is bounded as*

$$\begin{aligned} \mathbb{E}_\pi[Q] &\leq \frac{\rho \text{Var}[\Lambda A] + \text{Var}[\mu S] + 1 - \rho}{2(1 - \rho)} + \min\left(n, \frac{1}{1 - \rho}\right) \left(R_s^{\max} - \frac{1}{2} \mathbb{E}[(\mu S)^2] \right) \\ &\quad + \left(\frac{1}{2} \mathbb{E}[(\Lambda A)^2] - R_a^{\min} \right). \end{aligned} \tag{7.6}$$

We first comment on the order of the bound in (7.6). Observe that this bound only depends on the distributions of the *unitized* service time μS and the unitized interarrival time ΛA . Thus, once these distributions are fixed, this bound remains $O(1/(1 - \rho))$ universally for any load $\rho < 1$ and server count n . When specialized to the Halfin-Whitt regime [135] and the

Non-Degenerate Slowdown regime [19], the bound has the orders $O(\sqrt{n})$ and $O(n)$, respectively, which match the correct order in these two regimes. The “min” in the second term of (7.6) leads to a smaller leading constant when $n = o(1/(1 - \rho))$, which includes the classical heavy-traffic regime with n fixed.

As we might notice, the bound in (7.6) shares some similarities to Kingman’s approximation for the $GI/GI/1$ queue [170], so it is natural to wonder what each term in (7.6) represents and where it comes from. In the following, we provide some high-level comments and leave more detailed discussions to the technical overview (Section 7.5).

1. The first term in (7.6) matches Kingman’s approximation for the $GI/GI/1$ queue whose server has interarrival time A and service time S/n , up to low-order errors; this $GI/GI/1$ queue can be interpreted as having a “super-server” that is n -times faster than each server in the $GI/GI/n$ queue. Because the first term in (7.6) is dominant as $\rho \uparrow 1$ with n fixed, it reflects the shared asymptotical behavior of $GI/GI/n$ and $GI/GI/1$ in the classical heavy-traffic regime. This connection between $GI/GI/n$ and $GI/GI/1$ is well known in the literature [see, e.g., 128, 145, 245]. In our analysis, this comparison between $GI/GI/n$ and $GI/GI/1$ is performed *implicitly* using the BAR technique.
2. The second and third terms in (7.6) bound the difference between the (modified) $GI/GI/n$ queue and the $GI/GI/1$ queue. In particular, the second term in (7.6) is associated with the conditional distribution of the residual service time given $Q = 0$ in the modified $GI/GI/n$ queue, reflecting a certain “boundary effect” when the queue is empty. Note that this term vanishes when S is exponentially distributed, since $\mathbb{E}[(\mu S)^2]/2$ is the expectation of μS ’s excess distribution, and recall that $R_s^{\max} = \sup_{t \geq 0} \mathbb{E}[\mu S - t \mid \mu S \geq t]$.
3. The third term in (7.6) is associated with the conditional distribution of the residual interarrival time given $Q = 0$ in the modified $GI/GI/n$ queue, and it vanishes when A is exponentially distributed.

Theorem 7.5 implies the following simpler bound after relaxing the minimum in (7.6) to $1/(1 - \rho)$:

$$\mathbb{E}_\pi[Q] \leq \frac{\rho \text{Var}[\Lambda A] - \rho + 2R_s^{\max}}{2(1 - \rho)} + \frac{1}{2} \mathbb{E}[(\Lambda A)^2] - R_a^{\min}. \quad (7.7)$$

When we consider the $M/GI/n$ queue, the bound in (7.7) can be further simplified into:

$$\mathbb{E}_\pi[Q] \leq \frac{R_s^{\max}}{1 - \rho}. \quad (7.8)$$

Next, we apply the bound in (7.8) to some examples, deriving the leading constants previewed in Table 7.1. We assume that the examples under consideration satisfy Assumption 7.3.

Example 7.6 (NBUE distribution). Consider the $M/GI/n$ queue whose service-time distribution is non-lattice and New Better Than Used in Expectation (NBUE). NBUE is a large class of distributions that contains all Increasing Failure Rate distributions; a distribution is NBUE if its mean residual time conditioned on any age is no larger than its original mean. Then by definition, we have $R_s^{\max} \leq \mathbb{E}[\mu S] = 1$, so (7.8) implies $\mathbb{E}_\pi[Q] \leq 1/(1 - \rho)$.

Example 7.7 (Gamma distribution). Consider the $M/GI/n$ queue with gamma-distributed service times. Suppose S follows the gamma distribution with shape parameter α and scale parameter θ , then the unitized service time μS is also gamma-distributed with shape α and scale $1/\alpha$. When $\alpha > 1$, μS has increasing failure rates, so $R_s^{\max} = \mathbb{E}[\mu S] = 1$. When $\alpha < 1$, direct calculation shows that $R_s^{\max} = 1/\alpha$. Because the gamma distribution is always non-lattice, (7.8) implies $\mathbb{E}_\pi[Q] \leq \max(1/\alpha, 1)/(1 - \rho)$.

Example 7.8 (Phase-type distribution). Consider the $M/GI/n$ queue whose service-time distribution follows the phase-type distribution with d phases. For $k \in [d]$, let τ_k be the expected residual time of S conditioned on it being in the phase k , as in Table 7.1. Then it is not hard to see that $R_s^{\max} \leq \mu \max_{k \in [d]} \tau_k < \infty$. Moreover, phase-type distributions are non-lattice. Therefore, (7.8) implies $\mathbb{E}_\pi[Q] \leq \mu \max_{k \in [d]} \tau_k/(1 - \rho)$.

Example 7.9 (Bounded distribution). Consider the $M/GI/n$ queue whose service-time distribution is non-lattice and bounded by some constant $S_{\max} > 0$. Then we have $R_s^{\max} \leq \mu S_{\max}$, so (7.8) implies $\mathbb{E}_\pi[Q] \leq \mu S_{\max}/(1 - \rho)$.

Comparing with Li and Goldberg [180] Li and Goldberg [180] has proved a wide range of queue length bounds of the order $O(1/(1 - \rho))$, which primarily focus on the tail of the steady-state queue length distribution (Theorems 1, 2, and 3). These tail bounds imply first-moment bounds as corollaries. For example:

Theorem (Adapted from Corollary 2 of [180]). *Consider the $GI/GI/n$ queue with homogeneous servers under the FCFS policy. Assume that (1)*

$\mathbb{E}[A^2] < \infty$; (2) there exists $\epsilon \in (0, 0.5)$ s.t. $\mathbb{E}[S^{2+\epsilon}] < \infty$; (3) $\rho < 1$; (4) the system has a unique limiting distribution π . Then

$$\begin{aligned} \mathbb{E}_\pi[Q] &\leq \frac{2.1 \times 10^{21} \mathbb{E}[(\mu S)^2] \left(\mathbb{E}[(\mu S)^2]^{1+\epsilon} + \mathbb{E}[(\mu S)^{2+\epsilon}] \right) \left(\frac{1}{\epsilon} \right)^4 + 49 \mathbb{E}[(\Lambda A)^2]}{1 - \rho}, \end{aligned} \quad (7.9)$$

The bound in (7.9) achieves the order $O(1/(1 - \rho))$, with the constant factor only depending on $\mathbb{E}[(\Lambda A)^2]$, $\mathbb{E}[(\mu S)^2]$ and $\mathbb{E}[(\mu S)^{2+\epsilon}]$. However, this bound involves a huge numerical constant. Similar large numerical constants also appear in all other $O(1/(1 - \rho))$ bounds in [180]. Our bounds improve upon the bounds in [180] when $R_s^{\max}$ is reasonably small.

7.5 An informal overview of the proof, with a roadmap

This section overviews the proof of Theorem 7.5. We will consider four cases with gradually increasing complexity: (1) $M/M/1$ queue, (2) $GI/GI/1$ queue, (3) $GI/GI/n$ queue with $n \leq 1/(1 - \rho)$, and (4) $GI/GI/n$ queue with $n > 1/(1 - \rho)$. Each case builds upon the previous one and requires at most one additional technique.

Cases (1)–(3) consist of *known results*, and the tutorial-style proof sketch provided here contains all important arguments of the proof at an informal level. Case (4) represents the main technical novelty of this work. At the end of the section, we provide a roadmap for the proof of this last hard case, which we will focus on in the next two sections. Readers already familiar with BAR may proceed directly to the $GI/GI/n$ case (Section 7.5.3).

7.5.1 $M/M/1$: quadratic test function

Takeaway: the mean queue length of $M/M/1$ can be read off from the stationarity of a *quadratic* test function.

Consider $M/M/1$ queue with the arrival rate λ and the service rate μ . The queue length Q jumps at the rate $\lambda + \mu$, and when a jump happens, the post- and pre-jump states Q^+ and Q^- satisfy:

$$Q^+ = \begin{cases} Q^- + 1 & \text{with probability } \lambda/(\lambda + \mu) \\ Q^- - \mathbb{1}\{Q^- > 0\} & \text{with probability } \mu/(\lambda + \mu). \end{cases} \quad (7.10)$$

In steady state, both Q^+ and Q^- follow the stationary distribution π due to PASTA; the update rule (7.10) thus establishes an invariance relation that allows us to derive equalities or inequalities about π . In particular, to compute the first moment $\mathbb{E}_\pi[Q]$, we use the stationarity of the second moment, i.e., $\mathbb{E}[(Q^+)^2] = \mathbb{E}[(Q^-)^2]$. Denote $\Delta Q \triangleq Q^+ - Q^-$, then we have

$$\begin{aligned} 0 &= \mathbb{E}[(Q^- + \Delta Q)^2 - (Q^-)^2] \\ &= \mathbb{E}[2(\Delta Q)Q^-] + \mathbb{E}[(\Delta Q)^2]. \end{aligned} \quad (7.11)$$

The key observation is that $\mathbb{E}[\Delta Q \mid Q^-] = (\lambda - \mu \mathbb{1}\{Q^- > 0\})/(\lambda + \mu)$ is almost a constant, making the first term in (7.11) proportional to $\mathbb{E}[Q^-]$:

$$\mathbb{E}[2(\Delta Q)Q^-] = \frac{2(\lambda - \mu)}{\lambda + \mu} \mathbb{E}[Q^-].$$

In addition, one can verify that $\mathbb{E}[(\Delta Q)^2] = 2\lambda/(\lambda + \mu)$. Substituting these two terms back into (7.11), we get $\mathbb{E}_\pi[Q] = \mathbb{E}[Q^-] = \lambda/(\mu - \lambda) = \rho/(1 - \rho)$.

Intuitively, the queue length of $M/M/1$ drifts towards zero at a constant mean rate, with noisy transitions that occasionally cause increases. The steady-state mean $\mathbb{E}_\pi[Q]$ is fully determined by the rate at which the queue length drifts towards zero, $\mathbb{E}[\Delta Q \mid Q^-]$, and the second moment of the noise, $\mathbb{E}[(\Delta Q)^2]$, via (7.11).

7.5.2 $GI/GI/1$: “smoothed” quadratic test function

Takeaway: a “smoothed” variant of the quadratic test function extends the argument to general service times.

While the quadratic test function argument may not appear necessary for $M/M/1$ whose stationary distribution is available in closed form, it turns out to be incredibly useful for $GI/GI/1$: one can get an equation similar to (7.11) and manipulate it into a simple bound on $\mathbb{E}_\pi[Q]$, without figuring out everything about the stationary distribution π . Below, we walk through

the $GI/GI/1$ analysis in detail. We first present a naive adaptation of the $M/M/1$ argument, explain why it fails, and then present a small refinement to fix the argument. The techniques presented here serve as the foundation for our analysis of $GI/GI/n$.

To derive an upper bound for the queue length of $GI/GI/1$, we focus on its modified version (Section 7.3.2). Let A and S denote generic interarrival and service times, and let $\lambda \triangleq 1/\mathbb{E}[A]$ and $\mu \triangleq 1/\mathbb{E}[S]$ be the arrival and completion rates. Then the queue length of the modified $GI/GI/1$ is modulated by two renewal processes, the arrival and completion processes, with inter-event time distributions A and S , respectively. The jump in queue length $\Delta Q = Q^+ - Q^-$ is given by:

$$\text{At arrival:} \quad \Delta Q = 1 \quad (7.12)$$

$$\text{At completion:} \quad \Delta Q = -\mathbb{1}\{Q^- > 0\}. \quad (7.13)$$

We first consider the naive quadratic test function Q^2 . Intuitively, the long-run average rate of change in Q^2 caused by arrivals and completions should sum to zero. This intuition is captured by the following equation:

$$0 = \lambda \mathbb{E}_a[(Q^- + \Delta Q)^2 - (Q^-)^2] + \mu \mathbb{E}_s[(Q^- + \Delta Q)^2 - (Q^-)^2] \quad (7.14)$$

$$= 2\lambda \mathbb{E}_a[Q^-] - 2\mu \mathbb{E}_s[Q^-] + \lambda + \mu \mathbb{P}_s[Q^- > 0]. \quad (7.15)$$

Here, $\mathbb{E}_a[\cdot]$ denotes the expectation with respect to the system's state distribution observed at a random arrival event in the long run, so the first term $\lambda \mathbb{E}_a[(Q^- + \Delta Q)^2 - (Q^-)^2]$ is the long-run average rate of changes in Q^2 witnessed by arrivals; $\mathbb{E}_s[\cdot]$ is defined analogously for the completions; $\mathbb{E}_a[\cdot]$ and $\mathbb{E}_s[\cdot]$ are known as *Palm expectations* and will be formalized in Section 7.6.2.

Equations (7.14)–(7.15) are *not* useful here because $\mathbb{E}_a[Q^-]$ and $\mathbb{E}_s[Q^-]$ are generally different from $\mathbb{E}_\pi[Q]$. Intuitively, the queue observed right before a completion should generally be longer than average, whereas the queue observed before an arrival should generally be shorter.

To fix this, the idea is to consider a “smoothed” proxy for Q which (1) *continuously* drifts towards zero at a *constant* mean rate, as in the case of $M/M/1$, and (2) has zero expected jump sizes at arrival and completion events to avoid introducing $\mathbb{E}_a[Q^-]$ and $\mathbb{E}_s[Q^-]$. This smoothed proxy is given by

$$V \triangleq Q + \mu R_s - \lambda R_a, \quad (7.16)$$

where R_s and R_a are the residual service and arrival times. Let $\Delta V \triangleq V^+ - V^-$ denote V 's jump size at a generic event, and let $\dot{V}$ denote V 's continuous rate of change, then

$$\text{At an arrival:} \quad \Delta V = 1 - \lambda A \quad (7.17)$$

$$\implies \mathbb{E}[\Delta V \mid Q^-, R_a^-, R_s^-] = 0$$

$$\text{At a completion:} \quad \Delta V = -\mathbb{1}\{Q^- > 0\} + \mu S \quad (7.18)$$

$$\implies \mathbb{E}[\Delta V \mid Q^-, R_a^-, R_s^-] = \mathbb{1}\{Q^- = 0\}$$

$$\text{At any time:} \quad \dot{V} = -\mu + \lambda \quad (7.19)$$

Here, A denotes the new interarrival time sampled at the arrival time and is independent of the pre-jump state (Q^-, R_a^-, R_s^-) ; S is defined analogously.

Consider the quadratic test function V^2 . The long-run average rate of change in V^2 being zero leads to the next equation; compared to (7.14), there is an additional term $\mathbb{E}_\pi[2\dot{V}V]$ corresponding to continuous changes:

$$0 = \mathbb{E}_\pi[2\dot{V}V] + \lambda \mathbb{E}_a[(V^- + \Delta V)^2 - (V^-)^2] \quad (7.20)$$

$$+ \mu \mathbb{E}_s[(V^- + \Delta V)^2 - (V^-)^2]$$

$$= \mathbb{E}_\pi[2\dot{V}V] + \lambda \mathbb{E}_a[2(\Delta V)V^-] + \mu \mathbb{E}_s[2(\Delta V)V^-] \quad (7.21)$$

$$+ \lambda \mathbb{E}_a[(\Delta V)^2] + \mu \mathbb{E}_s[(\Delta V)^2]$$

$$= -2(\mu - \lambda)\mathbb{E}_\pi[V] + 2\mu \mathbb{E}_s[\mathbb{1}\{Q^- = 0\}V^-] \quad (7.22)$$

$$+ \lambda \mathbb{E}_a[(\Delta V)^2] + \mu \mathbb{E}_s[(\Delta V)^2],$$

where we expand the squares in (7.21) and substitute the values of $\dot{V}$ and $\mathbb{E}[\Delta V \mid V^-]$ to get (7.22). We move $-2(\mu - \lambda)\mathbb{E}_\pi[V]$ in (7.22) to the left-hand side and substitute $\mathbb{E}_a[(\Delta V)^2] = \text{Var}[\lambda A]$ and $\mathbb{E}_s[(\Delta V)^2] = \text{Var}[\mu S] + 1 - \rho$, which yields

$$2(\mu - \lambda)\mathbb{E}_\pi[V] = 2\mu \mathbb{E}_s[\mathbb{1}\{Q^- = 0\}V^-] + \lambda \text{Var}[\lambda A] + \mu(\text{Var}[\mu S] + 1 - \rho).$$

Because $Q = V - (\mu R_s - \lambda R_a)$, further rearranging the terms yields

$$\begin{aligned} \mathbb{E}_\pi[Q] &= \frac{\rho \text{Var}[\lambda A] + \text{Var}[\mu S] + 1 - \rho}{2(1 - \rho)} \\ &\quad + \frac{1}{1 - \rho} \mathbb{E}_s[\mathbb{1}\{Q^- = 0\}(\mu R_s^- - \lambda R_a^-)] - \mathbb{E}_\pi[\mu R_s - \lambda R_a]. \end{aligned} \quad (7.23)$$

It is not hard to see that the second and third terms above are low-order terms: in particular, the second term is non-positive since $R_s^- = 0$ under $\mathbb{E}_s[\cdot]$; the third term is bounded by $O(1)$. We skip presenting the explicit bounds for these terms.

Equations like (7.14) and (7.20) follow from the Basic Adjoint Relationship (BAR); we will formalize them in Section 7.6.2, following the framework in [47]. The “smoothed” test function V^2 is adapted from Miyazawa [212]; this test function and its variants have been the core of many recent advances in the steady-state analysis of queueing systems with general distributions [see, e.g., 50, 52, 54, 70, 72, 73, 130, 131, 146, 212].

7.5.3 $GI/GI/n$

Now we are ready to consider the $GI/GI/n$ queue. For the purpose of deriving an upper bound for $\mathbb{E}_\pi[Q]$, we focus on the modified $GI/GI/n$ system and refer to it simply as $GI/GI/n$ for brevity. We omit the superscript “—” on random variables within the Palm expectations $\mathbb{E}_s[\cdot]$ or $\mathbb{E}_a[\cdot]$ when the context is clear.

The $GI/GI/n$ queue can be viewed as “partly” a $GI/GI/1$: when all servers are busy ($Q > 0$), the total service rate is constant, and the n servers act as a single “super-server”. However, $GI/GI/n$ deviates from the super-server ideal at the boundary ($Q = 0$), during which virtual jobs could be added. Consequently, upper-bounding the queue length of $GI/GI/n$ boils down to quantifying how much this boundary effect could increase the queue length in the steady state.

To quantify the difference between $GI/GI/n$ and $GI/GI/1$ in queue lengths, we generalize the test function used for $GI/GI/1$ and substitute it in the BAR of $GI/GI/n$. Specifically, consider a smoothed proxy for Q given by:

$$V \triangleq Q + \sum_{i=1}^n \mu R_{s,i} - \Lambda R_a, \quad (7.24)$$

where $R_{s,i}$ is the residual service time of server i . Expanding the BAR for the quadratic test function V^2 and following the same steps as in (7.21)–(7.22),

we get a decomposition of $\mathbb{E}_\pi[Q]$ that generalizes the single-server result (7.23) (see Section 7.7.1 for the detailed calculation):

$$\mathbb{E}_\pi[Q] = \frac{\rho \text{Var}(\Lambda A) + \text{Var}(\mu S) + 1 - \rho}{2(1 - \rho)} + \frac{1}{1 - \rho} \left(\sum_{j=1}^n \Gamma_{s,j} - \Gamma_a \right), \quad (7.25)$$

where $\Gamma_{s,j}$ and Γ_a are the boundary terms, defined as:

$$\Gamma_{s,j} \triangleq \mathbb{E}_s[\mathbf{1}\{Q = 0\} \mu R_{s,j}] - (1 - \rho) \mathbb{E}_\pi[\mu R_{s,j}], \quad (7.26)$$

$$\Gamma_a \triangleq \mathbb{E}_s[\mathbf{1}\{Q = 0\} \Lambda R_a] - (1 - \rho) \mathbb{E}_\pi[\Lambda R_a], \quad (7.27)$$

where $\mathbb{E}_s[\cdot]$ is the expectation corresponding to the long-run averaging at the completion events of *all servers*. Contrasting with (7.23), we see that the first term on the right-hand side of (7.25) is approximately the mean queue length of $GI/GI/1$ with the super server. Therefore, $\Gamma_{s,j}$ and Γ_a capture the difference between $GI/GI/n$ and the corresponding $GI/GI/1$.

A useful heuristic is to roughly view $\Gamma_{s,j}$ (or Γ_a) as the *covariance* between the event $\{Q = 0\}$ and the residual service (or arrival) time, if we assume $\mathbb{E}_s[\cdot] \approx \mathbb{E}_\pi[\cdot]$ and show that $\mathbb{P}_s[Q = 0] = 1 - \rho$.

Although these covariances are hard to calculate exactly, we will prove that they are small: by their definitions, it suffices to show that $\{Q = 0\}$ occurs with a low probability, or to show that the expected residual service (or arrival) times *conditioned on* $Q = 0$ are not significantly larger (or smaller) than the unconditional expectations.

This heuristic view of $\Gamma_{s,j}$ and Γ_a is also consistent with the following intuition about the mean queue length: the mean queue length of $GI/GI/n$ is close to that of $GI/GI/1$ if either (1) Q rarely hits zero, so the boundary effect is naturally small, or (2) the next completion time right after Q hits zero is not too far away, so Q *will not build up significantly before seeing the next completion*.

Motivated by the above discussion, we divide the rest of the proof into two cases, based on the relationship between n and the load ρ , as they determine the probability of Q hitting 0:

- Simpler case: $n \leq 1/(1 - \rho)$. This corresponds to a heavily loaded regime ($\rho \geq 1 - 1/n$) where $\{Q = 0\}$ occurs with a low probability in the steady state. This case has been studied more extensively in the literature.

- Harder case: $n > 1/(1 - \rho)$. This corresponds to a more lightly loaded regime where $\{Q = 0\}$ occurs with relatively higher probability in the steady state, so we need to more precisely understand the conditional expectation of the residual times, $\mathbb{E}_s[R_{s,j} \mid Q = 0]$ and $\mathbb{E}_s[R_a \mid Q = 0]$. Addressing this case is the *main technical challenge* of this work.

Simpler case: $GI/GI/n$ with $n \leq 1/(1 - \rho)$

In this case, it suffices to show

$$\sum_{j=1}^n \Gamma_{s,j} \leq C(1 - \rho)n, \quad (7.28)$$

for some constant C independent of ρ and n . Because $\Gamma_a \geq 0$, combined with (7.25), we get $\mathbb{E}_\pi[Q] = O(1/(1 - \rho))$. To establish (7.28), we first bound $\Gamma_{s,j}$ by $O(1 - \rho)$ for each j . Specifically, we can show that

$$\begin{aligned} \Gamma_{s,j} &= \mathbb{E}_s[\mathbb{1}\{Q = 0\}\mu R_{s,j}] - (1 - \rho)\mathbb{E}_\pi[\mu R_{s,j}] \\ &= \mathbb{E}_s[\mu R_{s,j} \mid Q = 0]\mathbb{P}_s[Q = 0] - (1 - \rho)\mathbb{E}_\pi[\mu R_{s,j}] \\ &\leq (1 - \rho)\left(R_s^{\max} - \mathbb{E}[(\mu S)^2]/2\right), \end{aligned} \quad (7.29)$$

where (7.29) follows from the facts that $\mathbb{E}_s[\mu R_{s,j} \mid Q = 0] \leq R_s^{\max}$, $\mathbb{P}_s[Q = 0] = 1 - \rho$, and $\mathbb{E}_\pi[\mu R_{s,j}] = \mathbb{E}[(\mu S)^2]/2$. This leads to the final bound:

$$\begin{aligned} \mathbb{E}_\pi[Q] &\leq \frac{\rho \text{Var}[\Lambda A] + \text{Var}[\mu S] + 1 - \rho}{2(1 - \rho)} + n \left(R_s^{\max} - \frac{\mathbb{E}[(\mu S)^2]}{2} \right) \\ &\quad + \left(\frac{\mathbb{E}[(\Lambda A)^2]}{2} - R_a^{\min} \right), \end{aligned} \quad (7.30)$$

where the $-R_a^{\min}$ term comes from a more refined upper bound of $-\Gamma_a$, obtained using an argument similar to (7.29). The bound in (7.30) holds for all n and ρ , and when $n \leq 1/(1 - \rho)$, it achieves the desired order, $O(1/(1 - \rho))$, and coincides with the inequality stated in Theorem 7.5.

Bounds with $O(1/(1 - \rho))$ order for small n have been established in the literature for $GI/GI/n$ and its variants [e.g., 128, 148, 174, 192, 213, 245]. The proof sketch presented here follows a two-step structure common to this line of work: the relevant performance metric is first decomposed into a

$1/(1 - \rho)$ term with explicit coefficients and several boundary terms, which are subsequently shown to be small. For instance, the results in [128] imply an $O(1/(1 - \rho) + n)$ bound for $M/GI/n$, with the same coefficients for the $1/(1 - \rho)$ term as in (7.30); a key step of their analysis (Lemma 3) utilizes a similar decomposition as (7.25). In the following section, we show that this two-step framework can yield an $O(1/(1 - \rho))$ bound, if we adopt a more refined analysis of the boundary terms when $n > 1/(1 - \rho)$.

Harder case: $GI/GI/n$ with $n > 1/(1 - \rho)$

We show for any $n > 1/(1 - \rho)$,

$$\sum_{j=1}^n \Gamma_{s,j} \leq C, \quad (7.31)$$

for some constant C independent of ρ and n , so $\mathbb{E}_\pi[Q] = O(1/(1 - \rho))$. By symmetry, (7.31) is equivalent to

$$\Gamma_{s,1} \triangleq \mathbb{E}_s[\mathbb{1}\{Q = 0\}\mu R_{s,1}] - (1 - \rho)\mathbb{E}_\pi[\mu R_{s,1}] \leq \frac{C}{n}, \quad (7.32)$$

where we take $j = 1$ to focus on the residual service time of the first server, $R_{s,1}$.

To bound $\Gamma_{s,1}$, the naive bound in (7.29) is no longer helpful as it only gives $\Gamma_{s,1} \leq C(1 - \rho)$, which is too loose when n is much larger than $1/(1 - \rho)$.

To prove a more refined bound, recall the intuition that $\Gamma_{s,1}$ is roughly the covariance between $\mathbb{1}\{Q = 0\}$ and $R_{s,1}$. We further observe that $\mathbb{1}\{Q = 0\}$ and $R_{s,1}$ are *asymptotically independent* when n gets large: Intuitively, the queue length Q in a modified $GI/GI/n$ queue is the joint effect of n independent completion processes, so each process should only have a $1/n$ fraction of “responsibility” for the realization of $\mathbb{1}\{Q = 0\}$.

To quantify the effect of the first server’s completion process on the queue length, consider the following motivating question:

How would the queue length change if a server had not existed?

An attempt to answer this question inspires a clean proof of (7.32). We consider the counterfactual queue length, Q_{-1} , if the first server had not existed in the modified $GI/GI/n$ queue from the beginning, but all arrivals and completions of other servers had remained the same (see Section 7.6.1 for the precise definition of Q_{-1}). Then Q_{-1} is independent of $R_{s,1}$, and $Q_{-1} \geq Q$. Consider the indicator variable $\mathbb{1}\{Q_{-1} = 0\}$, which has two key properties:

Property 1: $\mathbb{1}\{Q_{-1} = 0\}$ is independent of $R_{s,1}$.

Property 2: $\mathbb{1}\{Q = 0\} \geq \mathbb{1}\{Q_{-1} = 0\}$ and $\mathbb{E}_s[\mathbb{1}\{Q = 0\} - \mathbb{1}\{Q_{-1} = 0\}] = O(1/n)$.

In other words, $\mathbb{1}\{Q = 0\}$ is approximated by a random variable independent of $R_{s,1}$, which captures our intuition of asymptotic independence.

To formally show $\Gamma_{s,1} \leq C/n$, consider the decomposition:

$$\Gamma_{s,1} = \mathbb{E}_s \left[\left(\mathbb{1}\{Q = 0\} - \mathbb{1}\{Q_{-1} = 0\} \right) \mu R_{s,1} \right] \quad (7.33)$$

$$+ \mathbb{E}_s[\mathbb{1}\{Q_{-1} = 0\} \mu R_{s,1}] - (1 - \rho) \mathbb{E}_\pi[\mu R_{s,1}]. \quad (7.34)$$

The term on the right-hand side of (7.33) can be bounded by $R_s^{\max}/n$ using Property 2, and the terms in (7.34) can be explicitly calculated to be $-\mathbb{E}[(\mu S)^2]/(2n)$ with Property 1. Putting together, we get

$$\Gamma_{s,1} \leq \frac{1}{n} \left(R_s^{\max} - \mathbb{E}[(\mu S)^2]/2 \right),$$

which proves (7.32) and thus Theorem 7.5 when $n > 1/(1 - \rho)$. The rigorous argument can be found in Section 7.7.2.

7.5.4 Roadmap for the rigorous proof

Sections 7.6 and 7.7 provide the formal development of the arguments outlined in this overview. We focus on the $n > 1/(1 - \rho)$ case as it constitutes the main technical contribution of this chapter; we skip the $n \leq 1/(1 - \rho)$ case as all important arguments have been provided in Section 7.5.3. Note that our chapter is still self-contained, as the proof for all regimes is provided for the heterogeneous generalization (Section 7.8).

Here is a detailed breakdown of the content in Sections 7.6 and 7.7. Section 7.6 establishes preliminary definitions and supporting lemmas:

- Section 7.6.1: The definition of *leave-one-out systems*, formalizing Q_{-1} used in Section 7.5.3.
- Section 7.6.2: Basic Adjoint Relationship and relevant concepts (e.g., $\mathbb{E}_s[\cdot]$ and $\mathbb{E}_a[\cdot]$).
- Section 7.6.3: Statements of intuitive facts and supporting properties used in the proof sketch, such as the values of $\mathbb{P}_s[Q = 0]$ and $\mathbb{E}_\pi[R_a]$, the bound $\mathbb{E}_s[\mu R_{s,j} \mid Q = 0] \leq R_s^{\max}$, and the independence between Q_{-1} and $R_{s,1}$.

Section 7.7 contains the central proof, which consists of two parts:

- Section 7.7.1: Proving the decomposition of $\mathbb{E}_\pi[Q]$ stated in (7.25).
- Section 7.7.2: Formalizing the argument in Section 7.5.3 to bound $\Gamma_{s,j}$ and Γ_a .

Finally, Section 7.8 generalizes these arguments to $GI/GI/n$ with heterogeneous servers, following essentially the same set of ideas.

The appendix of the chapter contains the proofs of some supporting lemmas. Although these proofs may appear to be long, they prove intuitive facts, so the readers should be able to understand the main body of this chapter without reading the appendices.

7.6 Technical framework and supporting lemmas

In this section, we formally set up the technical framework and state the supporting lemmas for Theorem 7.5. We define the leave-one-out systems in Section 7.6.1, set up the Basic Adjoint Relationship in Section 7.6.2, prove some additional supporting lemmas in Section 7.6.3.

To simplify the presentation, we adopt the following assumption. Our results hold without this assumption, and we will outline the necessary modifications for the general case in Remark 7.13.

Assumption 7.10. In the modified $GI/GI/n$ queue, no two events (i.e., an arrival and a completion, two arrivals, or two completions) happen at the same time.

In addition, unless otherwise stated, we assume $\rho < 1 - 1/n$.

7.6.1 Leave-one-out systems

In this section, we construct a set of auxiliary systems, referred to as *leave-one-out systems*. Each leave-one-out system maintains a counterfactual queue length if a certain server were removed from the main n -server modified $GI/GI/n$ system under study (referred to as the “main system” or simply “the modified $GI/GI/n$ queue” from now on).

As illustrated in Figure 7.2, for each $j \in [n]$, the j -th leave-one-out system is a modified $GI/GI/(n-1)$ system coupled with the main system, and it shares all arrival events and the completion events of all but the j -th server

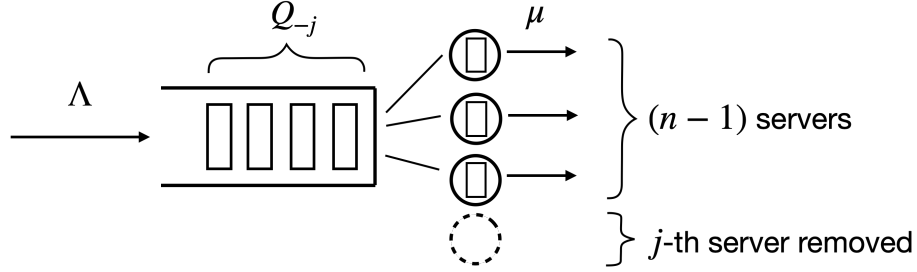


Figure 7.2: An illustration of the j -th leave-one-out system, where the j -th server is removed and does not take jobs from the queue.

with the main system. More concretely, each job arrival in the main system also increases $Q_{-j}(t)$ by 1 for each $j \in [n]$, and each completion of the i -th server reduces $Q_{-j}(t)$ by $\mathbf{1}\{Q_{-j}(t-) > 0\}$ for $i, j \in [n]$ such that $i \neq j$. We initialize $Q_{-j}(0) = Q(0)$ all $j \in [n]$, and denote the joint queue lengths of all leave-one-out systems as $\tilde{Q}(t) \triangleq (Q_{-j}(t))_{j \in [n]}$.

As one would expect, the queue length of each leave-one-out system pathwise dominates the queue length of the main system:

Lemma 7.11. *For each $j \in [n]$, the queue length of the j -th leave-one-out system pathwise upper bounds the queue length of the main system, i.e.,*

$$Q_{-j}(t) \geq Q(t) \quad \forall t \geq 0.$$

Proof. We define the stopping time $\tau = \inf\{t \geq 0: Q_{-j}(t) \leq Q(t) - 1\}$. It suffices to show that $\tau = \infty$ for all sample paths. First, because $Q_{-j}(0) = Q(0)$, by the right-continuity of $Q_{-j}(t)$ and $Q(t)$, we must have $\tau > 0$. Suppose there exists a sample path such that $\tau < \infty$, then we have $Q_{-j}(\tau-) = Q(\tau-)$ and $Q_{-j}(\tau) \leq Q(\tau) - 1$. This implies that at time τ , there must be at least a completion event that affects the j -th leave-one-out system but not the main system, or an arrival event that affects the main system but not the j -th leave-one-out system. As neither of these cases is possible, we get a contradiction. Consequently, for all sample paths, $\tau = \infty$, and thus $Q_{-j}(t) \geq Q(t)$ for all $t \geq 0$. $\square$

Finally, by Assumption 7.3, a leave-one-out system is positive Harris recurrent if $\rho < 1 - 1/n$.

7.6.2 Basic Adjoint Relationship

As outlined in Section 7.5, we adopt the Basic Adjoint Relationship (BAR) approach to perform the steady-state analysis. In this section, we formally set up the preliminaries of BAR, following the framework in [54, Section 6].

We will establish BAR for the state-augmented Markov process $\{\bar{X}(t)\}_{t \geq 0}$, defined as:

$$\bar{X}(t) \triangleq (R_{s,1}(t), R_{s,2}(t), \dots, R_{s,n}(t), R_a(t), Q(t), Q_{-1}(t), \dots, Q_{-n}(t)), \quad (7.35)$$

with the state space $\bar{\mathcal{X}} \triangleq \mathbb{R}_{\geq 0}^{n+1} \times \mathbb{Z}_{\geq 0}^{n+1}$. Since we assume $\rho < 1 - 1/n$ throughout Sections 7.6 and 7.7, Assumption 7.3 ensures that $\{\bar{X}(t)\}_{t \geq 0}$ possesses a unique stationary distribution π . From this point forward until the end of Section 7.7, we slightly abuse the notation to let $\bar{X}(0) \sim \pi$, so that $\bar{X}(t) \sim \pi$ for all $t \geq 0$.

Next, we formalize the three ways in which $\bar{X}(t)$ evolves: the continuous decrease of residual times, and the discrete jumps associated with arrivals and service completions.

To characterize the continuous changes, we define the differential operator $\mathcal{A}$ as

$$\mathcal{A}f(\bar{x}) \triangleq - \sum_{\ell=1}^{n+1} \partial_{\ell}^+ f(\bar{x}) \quad \forall \bar{x} \in \bar{\mathcal{X}}, \quad (7.36)$$

where ∂_{ℓ}^+ denotes the right partial derivative with respect to the ℓ -th coordinate, and $f: \bar{\mathcal{X}} \rightarrow \mathbb{R}$ is any function that has right partial derivatives with respect to the first $(n+1)$ coordinates (the coordinates corresponding to $R_{s,i}$'s and R_a). This operator captures the instantaneous rate of change of $f(\bar{X}(t))$ when $\bar{X}(t) = \bar{x}$.

Next, we consider changes in $\bar{X}(t)$ caused by the arrivals. Since arrivals occur at discrete points, we cannot define their rates. Instead, we use the so-called *Palm distribution*, which is a probability measure over $\bar{\mathcal{X}}^2$. The Palm distribution captures the steady-state probability of the pair of states before and after a random jump. Formally, the *Palm distribution at arrivals* $\mathbb{P}_a$ is defined as

$$\mathbb{P}_a[B] = \frac{1}{\Lambda h} \mathbb{E} \left[\sum_{k=1}^{\infty} \mathbb{1} \left\{ \left(\bar{X}(\tau_a^k), \bar{X}(\tau_a^k -) \right) \in B \right\} \mathbb{1} \left\{ \tau_a^k \leq h \right\} \right], \quad (7.37)$$

where B is any measurable subset of $\bar{\mathcal{X}}^2$ and τ_a^k is the time of the k -th arrival event; h is any positive number, and $\mathbb{P}_a$ does not depend on the choice of h

due to the stationarity of $\bar{X}(t)$ [16, Theorem VII.6.3]. Roughly speaking, $\mathbb{P}_a$ represents the long-run average distribution evaluated at arrival events; this can be seen by taking a large h , and note that Λ is the long-term average frequency of arrival events. We refer the readers to [16, Section VII.6] or [54, Section 6] for a detailed background of Palm distributions induced by point processes.

Given the Palm distribution $\mathbb{P}_a$, we can define the corresponding Palm expectation $\mathbb{E}_a[\cdot]$: For any measurable function $g: \bar{\mathcal{X}}^2 \rightarrow \mathbb{R}$, we have

$$\mathbb{E}_a[g(\bar{X}^+, \bar{X}^-)] \triangleq \frac{1}{\Lambda h} \mathbb{E} \left[\sum_{k=1}^{\infty} g(\bar{X}(\tau_a^k), \bar{X}(\tau_a^k -)) \mathbf{1}_{\{\tau_a^k \leq h\}} \right], \quad (7.38)$$

where $(\bar{X}^+, \bar{X}^-)$ on the left-hand side is a pair of dummy variables following the distribution $\mathbb{P}_a$, and h is any positive number.

We similarly define $\mathbb{P}_{s,i}$ and $\mathbb{E}_{s,i}[\cdot]$ for the completion events of the i -th server for $i \in [n]$.

Now we are ready to state the main lemma for the BAR approach, which connects the stationary distribution π with the Palm distributions $\mathbb{P}_a$ and $\mathbb{P}_{s,i}$ defined above, and characterize these Palm distributions. The proof of Lemma 7.12 is provided in Appendix 7.D.2, which is adapted from [54, Section 6.1].

Lemma 7.12 (Basic Adjoint Relationship for $GI/GI/n$). *Consider the modified $GI/GI/n$ system satisfying Assumptions 7.1, 7.2 and 7.3, with load $\rho < 1 - 1/n$. For any function $f: \bar{\mathcal{X}} \rightarrow \mathbb{R}$ that is bounded, locally Lipschitz continuous, and partially differentiable from the right with respect to its first $(n+1)$ coordinates, we have*

$$\mathbb{E}_\pi[\mathcal{A}f(\bar{X})] + \Lambda \mathbb{E}_a[f(\bar{X}^+) - f(\bar{X}^-)] + \mu \sum_{i=1}^n \mathbb{E}_{s,i}[f(\bar{X}^+) - f(\bar{X}^-)] = 0, \quad (7.39)$$

where $\bar{X}, \bar{X}^+, \bar{X}^- \in \bar{\mathcal{X}}$ are random elements following the distributions specified in the subscripts of expectations. Moreover, the Palm distributions $\mathbb{P}_a$ and $\mathbb{P}_{s,i}$ satisfy the following with probability 1:

- (i) If $(\bar{X}^+, \bar{X}^-) \sim \mathbb{P}_a$, then $R_a^- = 0$ and the coordinates of $\bar{X}^+$ and $\bar{X}^-$ satisfy the relations:

$$R_a^+ = R_a^- + A \quad (7.40)$$

$$Q^+ = Q^- + 1 \quad (7.41)$$

$$Q_{-j}^+ = Q_{-j}^- + 1 \quad \forall j \in [n], \quad (7.42)$$

where A is a generic interarrival time independent of $\bar{X}^-$; other coordinates of $\bar{X}^+$ are identical to $\bar{X}^-$. (Here and throughout, any superscript applied to an element of $\bar{X}$, such as $\bar{X}^-$ or $\bar{X}^+$, propagates to all of its coordinates.)

(ii) For each $i \in [n]$, if $(\bar{X}^+, \bar{X}^-) \sim \mathbb{P}_{s,i}$, then $R_{s,i}^- = 0$ and the coordinates of $\bar{X}^+$ and $\bar{X}^-$ satisfy the relations:

$$R_{s,i}^+ = R_{s,i}^- + S \quad (7.43)$$

$$Q^+ = Q^- - \mathbf{1}\{Q^- > 0\} \quad (7.44)$$

$$Q_{-j}^+ = Q_{-j}^- - \mathbf{1}\{Q_{-j}^- > 0\} \quad \forall j \in [n] \setminus \{i\}, \quad (7.45)$$

where S is a generic service time independent of $\bar{X}^-$; other coordinates of $\bar{X}^+$ are identical to $\bar{X}^-$.

We now provide a brief interpretation of the statement in Lemma 7.12.

Equation (7.39) states that the long-run average rate of change of $f(\bar{X}(t))$ is zero. In this expression, $\mathbb{E}_\pi[\mathcal{A}f(\bar{X})]$ represents the average rate of change due to the continuous drift of residual times, while the remaining terms capture the discrete changes associated with arrivals and service completions.

Furthermore, Equations (7.40)–(7.45) characterize the conditional Palm distribution of the post-jump state $\bar{X}^+$ given the pre-jump state $\bar{X}^-$. The specific relations between $\bar{X}^+$ and $\bar{X}^-$ defined by these identities are consistent with the construction of the modified $GI/GI/n$ in Section 7.3.2.

Next, we define some convenient notational shorthand. We define the Palm distribution $\mathbb{P}_s[\cdot] \triangleq (1/n) \sum_{i=1}^n \mathbb{P}_{s,i}[\cdot]$ and the corresponding expectation $\mathbb{E}_s[g(\bar{X}^+, \bar{X}^-)] \triangleq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{s,i}[g(\bar{X}^+, \bar{X}^-)]$. Under this notation, (7.39) can be compactly rewritten as

$$\mathbb{E}_\pi[\mathcal{A}f(\bar{X})] + \Lambda \mathbb{E}_a[f(\bar{X}^+) - f(\bar{X}^-)] + n\mu \mathbb{E}_s[f(\bar{X}^+) - f(\bar{X}^-)] = 0. \quad (7.46)$$

We also adopt the simplification that, when a Palm expectation only involves the pre-event state, $\bar{X}^-$, we omit its superscript and simply write $\bar{X}$.

Remark 7.13 (Simultaneous events). We have adopted the assumption that no events occur simultaneously (Assumption 7.10) to simplify the presentation. Nevertheless, our results still hold without this assumption, and we outline the necessary changes as follows.

The main changes are in the definitions of the Palm distributions. When $M(t) \geq 1$ events occur simultaneously at time t , we let $\bar{X}(t, 0), \dots, \bar{X}(t, M(t))$ denote the intermediate states produced by sequentially applying the state-updating subroutines (see Section 7.3.2); in particular, $\bar{X}(t, 0) \triangleq \bar{X}(t-)$ and $\bar{X}(t, M(t)) \triangleq \bar{X}(t)$. Consider the sequence of all pairs (t, m) such that $\bar{X}(t, m)$ is an intermediate state immediately preceding an arrival update. We order these pairs lexicographically and let (τ_a^k, m_a^k) denote the k -th element in this sequence. The definition of $\mathbb{P}_a$ is then generalized as:

$$\mathbb{P}_a[B] = \frac{1}{\Lambda h} \mathbb{E} \left[\sum_{k=1}^{\infty} \mathbb{1} \left\{ \left(\bar{X}(\tau_a^k, m_a^k + 1), \bar{X}(\tau_a^k, m_a^k) \right) \in B \right\} \mathbb{1} \{ \tau_a^k \leq h \} \right]. \quad (7.47)$$

The Palm distributions $\mathbb{P}_{s,i}$ for $i \in [n]$ should be generalized analogously. Intuitively, $\mathbb{P}_a$ and $\mathbb{P}_{s,i}$ still represent the average state distributions observed by arrival events or completion events, although the states at some time points could be counted multiple times or involved in multiple Palm distributions. The same definition has also been used in [54] to deal with simultaneous events.

With the new definitions of $\mathbb{P}_a$ and $\mathbb{P}_{s,i}$, the statement of Lemma 7.12 is valid without requiring any other changes. The other lemmas, including those for heterogeneous $GI/GI/n$ in Section 7.8, also hold verbatim. We leave the detailed verifications to readers.

Finally, we mention an interesting detail. As we can see in Section 7.3.2, when constructing the modified $GI/GI/n$, the order of updating the state in the presence of simultaneous events could affect the system dynamics. However, our generalized proof should work for any order of updates. Intuitively, whether we apply the arrival subroutine before or after a completion subroutine would affect the distribution of R_a^- under $\mathbb{P}_{s,i}$, but not those facts stated in Lemma 7.12.

7.6.3 Other supporting lemmas

In this section, we state the remaining supporting lemmas for the proof of Theorem 7.5. These lemmas formalize several intuitive properties of the system; we briefly discuss their intuitions and defer their proofs to Appendix 7.D.

The first lemma, Lemma 7.14, bounds the conditional mean residual times given the queue lengths, Q and $\tilde{Q} \triangleq (Q_{-j})_{j \in [n]}$. It is proved in Appendix 7.D.3.

Lemma 7.14 (Bounded mean residual times). *Consider the modified GI/GI/n queue satisfying Assumptions 7.1 and 7.3, with the load $\rho < 1 - 1/n$. For any $j \in [n]$,*

$$\mu \mathbb{E}_\nu [R_{s,j} \mid Q, \tilde{Q}] \leq R_s^{\max} \quad (7.48)$$

$$\Lambda \mathbb{E}_\nu [R_a \mid Q, \tilde{Q}] \geq R_a^{\min} \quad (7.49)$$

where $\mathbb{E}_\nu[\cdot]$ can denote either $\mathbb{E}_\pi[\cdot]$ or $\mathbb{E}_{s,i}[\cdot]$ for $i \in [n] \setminus \{j\}$.

The intuition behind Lemma 7.14 is that $R_s^{\max}$ is defined to be an upper bound on the mean residual service time conditioned on any *age* (the time elapsed since the last service completion). Because the service times are i.i.d., this age is a sufficient statistic for the mean residual service time. Consequently, Q and $\tilde{Q}$, which are functions of past arrival and completion times, provide no additional predictive information once the age is known; thus, conditioning on them does not change the bound. A similar logic applies to the lower bound $R_a^{\min}$ for the arrival process.

The next lemma establishes that for any $i, j \in [n]$ such that $i \neq j$, $R_{s,j}$ and $\mathbb{1}\{Q_{-j} = 0\}$ are uncorrelated under the Palm distribution $\mathbb{P}_{s,i}$. It is proved in Appendix 7.D.4.

Lemma 7.15 (Uncorrelation lemma). *Consider the modified GI/GI/n queue satisfying Assumptions 7.1, 7.2, and 7.3, with the load $\rho < 1 - 1/n$. For any $i, j \in [n]$ such that $i \neq j$, we have*

$$\mathbb{E}_{s,i}[\mathbb{1}\{Q_{-j} = 0\} R_{s,j}] = \mathbb{P}_{s,i}[Q_{-j} = 0] \mathbb{E}_\pi[R_{s,j}]. \quad (7.50)$$

Intuitively, $R_{s,j}$ depends only on the j -th server's completion process, whereas Q_{-j} and $\mathbb{P}_{s,i}$ are determined by the other n renewal processes; these $n + 1$ renewal processes are mutually independent. The formal proof of Lemma 7.15 also utilizes a mixing argument for the completion processes, which relies on S being non-lattice (Assumption 7.2). Notably, this lemma is the only part of our analysis where the non-lattice assumption is required.

The next lemma shows several basic equalities about the residual times of the renewal processes and the idle probability of the queue. This lemma is proved in Appendix 7.D.5.

Lemma 7.16. *Consider the modified GI/GI/n queue satisfying Assumptions 7.1 and 7.3. If $\rho < 1$, for $i, j \in [n]$ such that $i \neq j$, we have*

$$\mathbb{E}_\pi[R_{s,i}] = \frac{1}{2} \mu \mathbb{E}[S^2] \quad (7.51)$$

$$\mathbb{E}_\pi[R_a] = \frac{1}{2}\Lambda\mathbb{E}[A^2] \quad (7.52)$$

$$\mathbb{E}_{s,i}[R_{s,j}] + \mathbb{E}_{s,j}[R_{s,i}] = \mu\mathbb{E}[S^2] \quad (7.53)$$

$$\mathbb{E}_{s,i}[R_a] + \mathbb{E}_a[R_{s,i}] = \frac{1}{2}\mu\mathbb{E}[S^2] + \frac{1}{2}\Lambda\mathbb{E}[A^2] \quad (7.54)$$

$$\mathbb{P}_s[Q = 0] = 1 - \rho \quad (7.55)$$

In addition, if $\rho < 1 - 1/n$, for each $j \in [n]$, the j -th leave-one-out system satisfies

$$\frac{1}{n} \sum_{i: i \neq j} \mathbb{P}_{s,i}[Q_{-j} = 0] = 1 - \rho - \frac{1}{n}. \quad (7.56)$$

Intuitively, (7.51) and (7.52) characterize the mean residual times *under the steady-state distribution*, while (7.53) and (7.54) characterize these means *observed by independent renewal processes*. For instance, $\mathbb{E}_{s,i}[R_{s,j}]$ represents the mean residual time of the i -th service process observed at the completion times of the j -th service process, for some j different from i ; (7.53) implies that $\mathbb{E}_{s,i}[R_{s,j}]$ matches the steady-state mean $\mu\mathbb{E}[S^2]/2$ under the assumption that $\mathbb{E}_{s,i}[R_{s,j}] = \mathbb{E}_{s,j}[R_{s,i}]$, although our analysis requires neither the proof nor the assumption of such symmetry.

Equation (7.55) identifies the fraction of service completions that occur when $Q = 0$; these completions are “wasted” in the sense that a virtual job is immediately added upon the completion of the previous job. Since the ratio of the arrival rate to the total service rate is ρ , exactly a $(1 - \rho)$ fraction of the completions must be wasted in the long run to maintain stability. A similar intuition also applies to (7.56).

Finally, Lemma 7.17 establishes that the queue length Q has a finite mean under the Palm distributions $\mathbb{P}_a$ and $\mathbb{P}_{s,i}$. This result follows from the finiteness of $\mathbb{E}_\pi[Q]$ (Assumption 7.3). This lemma is proved in Appendix 7.D.5. **Lemma 7.17.** *Consider the modified GI/GI/n system satisfying Assumptions 7.1 and 7.3, with the load $\rho < 1$. Then we have $\mathbb{E}_a[Q] < \infty$ and $\mathbb{E}_{s,i}[Q] < \infty$ for all $i \in [n]$.*

7.7 Proof of main theorem

This section is dedicated to proving Theorem 7.5, following the two-step approach outlined in Section 7.5:

Step 1 (Section 7.7.1): We derive an exact expression for the mean queue length, $\mathbb{E}_\pi[Q]$. This expression is in terms of the load ρ , the variances $\text{Var}[\mu S]$ and $\text{Var}[\Lambda A]$, and some “covariance” terms, $\Gamma_{s,j}$ and Γ_a .

Step 2 (Section 7.7.2): We bound $\Gamma_{s,j}$ and Γ_a .

The first step follows from the established framework of BAR, while the second step contains the novel argument based on our leave-one-out technique. Substituting the bound in Step 2 into the equality in Step 1 proves Theorem 7.5.

7.7.1 Step 1: exact expression of $\mathbb{E}[Q]$

Lemma 7.18. *Consider the modified GI/GI/n system satisfying Assumptions 7.1 and 7.3, with the load $\rho < 1$. We have*

$$(1 - \rho)\mathbb{E}_\pi[Q] = \frac{\rho \text{Var}[\Lambda A] + \text{Var}[\mu S] + 1 - \rho}{2} + \sum_{j=1}^n \Gamma_{s,j} - \Gamma_a, \quad (7.57)$$

where $\Gamma_{s,j}$ and Γ_a are defined as:

$$\Gamma_{s,j} = \mathbb{E}_s[\mathbf{1}\{Q = 0\}\mu R_{s,j}] - (1 - \rho)\mathbb{E}_\pi[\mu R_{s,j}] \quad \forall j \in [n] \quad (7.58)$$

$$\Gamma_a = \mathbb{E}_s[\mathbf{1}\{Q = 0\}\Lambda R_a] - (1 - \rho)\mathbb{E}_\pi[\Lambda R_a], \quad (7.59)$$

and $\mathbb{E}_s[\cdot] \triangleq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{s,i}[\cdot]$ is the Palm expectation with respect to all completion events.

Remark 7.19. When n is large, the term $\Gamma_{s,j}$ can be roughly viewed as the covariance between $\mathbf{1}\{Q = 0\}$ and $\mu R_{s,j}$ under the measure $\mathbb{P}_s$, because $\mathbb{P}_s[Q = 0] = 1 - \rho$ and $\mathbb{E}_s[R_{s,j}] = \frac{1}{n} \sum_{i: i \neq j} \mathbb{E}_{s,i}[R_{s,j}] = \frac{n-1}{n} \mathbb{E}_\pi[R_{s,j}]$. Here we use the fact that $\mathbb{E}_{s,i}[R_{s,j}] = \mathbb{E}_\pi[R_{s,j}]$, which is slightly stronger than (7.53) in Lemma 7.16, but can be shown by replacing $\mathbf{1}\{Q_{-j} = 0\}$ with 1 in the proof of Lemma 7.15. We skip the detailed argument because it is not needed to prove our main result.

The term Γ_a can be interpreted similarly if the interarrival-time distribution is non-lattice.

Next, we prove Lemma 7.18, focusing on presenting the case where $\rho < 1 - 1/n$. The case where $\rho \geq 1 - 1/n$ follows from identical algebraic steps, with the only modification being that the BAR (Lemma 7.12) must be stated for the original state, $X(t)$, rather than the augmented state, $\bar{X}(t)$. This distinction arises because the augmented state involves $Q_{-j}(t)$, which is

unstable when $\rho \geq 1 - 1/n$; nevertheless, $Q_{-j}(t)$ is not required for the proof of Lemma 7.18.

Proof of Lemma 7.18. We define the random variable

$$V \triangleq Q + \mu \sum_{j=1}^n R_{s,j} - \Lambda R_a.$$

As discussed in Section 7.5.3, V serves as a smoothed version of the queue length Q . Because V possesses more tractable dynamics, we focus the majority of our proof on calculating $\mathbb{E}_\pi[V]$, from which the original goal of calculating $\mathbb{E}_\pi[Q]$ can be straightforwardly recovered.

To apply the Basic Adjoint Relationship (BAR, Lemma 7.12) to the quadratic term V^2 , we must first address the requirement that the test function should be bounded. We thus consider the truncated version of V ,

$$V_\kappa \triangleq (Q \wedge \kappa) + \mu \sum_{j=1}^n (R_{s,j} \wedge \kappa) - \Lambda (R_a \wedge \kappa)$$

for some integer $\kappa > 0$.

Let $f(\bar{X}) = V_\kappa^2$. Since f is bounded, locally Lipschitz continuous, and has right partial derivatives with respect to R_a and $R_{s,i}$'s, BAR implies that

$$\begin{aligned} 0 &= \mathbb{E}_\pi \left[2V_\kappa \left(-\mu \sum_{j=1}^n \mathbb{1}\{R_{s,j} \leq \kappa\} + \Lambda \mathbb{1}\{R_a \leq \kappa\} \right) \right] \\ &\quad + \Lambda \mathbb{E}_a \left[\left(V_\kappa + \mathbb{1}\{Q \leq \kappa - 1\} - \Lambda A \wedge \kappa \right)^2 - V_\kappa^2 \right] \\ &\quad + n\mu \mathbb{E}_s \left[\left(V_\kappa - \mathbb{1}\{0 < Q \leq \kappa\} + \mu S \wedge \kappa \right)^2 - V_\kappa^2 \right]. \end{aligned}$$

Expanding the squares and rearranging the terms, the above equation simplifies into

$$\begin{aligned} &2\mathbb{E}_\pi \left[V_\kappa \left(\mu \sum_{j=1}^n \mathbb{1}\{R_{s,j} \leq \kappa\} - \Lambda \mathbb{1}\{R_a \leq \kappa\} \right) \right] \\ &= \Lambda \mathbb{E}_a \left[2V_\kappa \left(\mathbb{1}\{Q \leq \kappa - 1\} - \Lambda A \wedge \kappa \right) \right] \end{aligned}$$

$$\begin{aligned}
& + \left(\mathbb{1}\{Q \leq \kappa - 1\} - \Lambda A \wedge \kappa \right)^2 \Big] \\
& + n\mu \mathbb{E}_s \left[2V \left(-\mathbb{1}\{0 < Q \leq \kappa\} + \mu S \wedge \kappa \right) \right. \\
& \quad \left. + \left(-\mathbb{1}\{0 < Q \leq \kappa\} + \mu S \wedge \kappa \right)^2 \right].
\end{aligned}$$

Now we want to let $\kappa \rightarrow \infty$ in the previous equation and apply the dominated convergence theorem (DCT). To apply DCT, recall the following finiteness results: $\mathbb{E}_\pi[Q] < \infty$ (Assumption 7.3), $\mathbb{E}_a[Q], \mathbb{E}_s[Q] < \infty$ (Lemma 7.17), $\mathbb{E}_\pi[R_a], \mathbb{E}_{s,i}[R_a], \mathbb{E}_\pi[R_{s,i}], \mathbb{E}_a[R_{s,i}], \mathbb{E}_{s,i}[R_{s,j}] < \infty$ (Lemma 7.16), $\mathbb{E}_a[R_a] = \mathbb{E}_{s,i}[R_{s,i}] = 0$, and $\mathbb{E}[A^2], \mathbb{E}[S^2] < \infty$ (Assumption 7.1). Then it is not hard to see that, in the previous displayed equation, the expression inside $\mathbb{E}_\pi[\cdot]$ is dominated by $(Q + \sum_{j=1}^n \mu R_{s,j} + \Lambda R_a)(n\mu + \Lambda)$, which is independent of κ and has a finite mean under the distribution π . For similar reasons, the expressions inside $\mathbb{E}_a[\cdot]$ and $\mathbb{E}_s[\cdot]$ are dominated by quantities which are independent of κ and have finite means under $\mathbb{P}_a$ and $\mathbb{P}_s$, respectively. Therefore, DCT implies that

$$2(n\mu - \Lambda)\mathbb{E}_\pi[V] \tag{7.60}$$

$$= \Lambda \mathbb{E}_a \left[2V(1 - \Lambda A) + (1 - \Lambda A)^2 \right] \tag{7.61}$$

$$+ n\mu \mathbb{E}_s \left[2V(-\mathbb{1}\{Q > 0\} + \mu S) + (-\mathbb{1}\{Q > 0\} + \mu S)^2 \right]. \tag{7.62}$$

Next, we simplify the terms in (7.61), which correspond to the arrival events:

$$\begin{aligned}
& \Lambda \mathbb{E}_a \left[2V(1 - \Lambda A) + (1 - \Lambda A)^2 \right] \\
& = 2\Lambda \mathbb{E}_a[V] \mathbb{E}[1 - \Lambda A] + \Lambda \mathbb{E}[(1 - \Lambda A)^2]
\end{aligned} \tag{7.63}$$

$$= \Lambda \text{Var}[\Lambda A], \tag{7.64}$$

where the first equality is due to the fact that A is independent of Q and $R_{s,j}$; the second equality is due to $\Lambda \mathbb{E}[A] = 1$.

We then simplify the terms in (7.62), which correspond to the completion events:

$$n\mu \mathbb{E}_s \left[2V(-\mathbb{1}\{Q > 0\} + \mu S) + (-\mathbb{1}\{Q > 0\} + \mu S)^2 \right]$$

$$\begin{aligned}
&= 2n\mu\mathbb{E}_s[V] \mathbb{E}[-1 + \mu S] + 2n\mu\mathbb{E}_s[V\mathbb{1}\{Q = 0\}] \\
&\quad + n\mu\mathbb{E}[(-1 + \mu S)^2] + 2n\mu\mathbb{P}_s[Q = 0] \mathbb{E}[-1 + \mu S] + n\mu\mathbb{P}_s[Q = 0] \\
&= 2n\mu\mathbb{E}_s[V\mathbb{1}\{Q = 0\}] + n\mu\text{Var}[\mu S] + n\mu\mathbb{P}_s[Q = 0] \\
&= 2n\mu\mathbb{E}_s[V\mathbb{1}\{Q = 0\}] + n\mu\text{Var}[\mu S] + n\mu(1 - \rho), \tag{7.65}
\end{aligned}$$

where the first equality uses $\mathbb{1}\{Q > 0\} = 1 - \mathbb{1}\{Q = 0\}$ and the fact that S is independent of Q , $R_{s,j}$, and R_a ; the second equality is due to $\mathbb{E}[\mu S] = 1$; the third equality is due to $\mathbb{P}_s[Q = 0] = 1 - \rho$ (Lemma 7.16).

Substituting (7.64) and (7.65) back to (7.61) and (7.62), we get

$$2(n\mu - \Lambda)\mathbb{E}_\pi[V] = 2n\mu\mathbb{E}_s[V\mathbb{1}\{Q = 0\}] + \Lambda\text{Var}[\Lambda A] + n\mu\text{Var}[\mu S] + n\mu(1 - \rho).$$

Dividing both sides by $2n\mu$ and using the fact that $\Lambda/(n\mu) = \rho$, the previous equality becomes

$$(1 - \rho)\mathbb{E}_\pi[V] = \mathbb{E}_s[V\mathbb{1}\{Q = 0\}] + \frac{\rho\text{Var}[\Lambda A] + \text{Var}[\mu S] + 1 - \rho}{2}.$$

Because $Q = V - \mu \sum_{j=1}^n R_{s,j} + \Lambda R_a$, rearranging the terms, we get

$$\begin{aligned}
&(1 - \rho)\mathbb{E}_\pi[Q] \\
&= \frac{\rho\text{Var}[\Lambda A] + \text{Var}[\mu S] + 1 - \rho}{2} \\
&\quad + \mathbb{E}_s\left[\left(\mu \sum_{j=1}^n R_{s,j} - \Lambda R_a\right)\mathbb{1}\{Q = 0\}\right] - (1 - \rho)\mathbb{E}_\pi\left[\mu \sum_{j=1}^n R_{s,j} - \Lambda R_a\right],
\end{aligned}$$

which implies (7.57). $\square$

7.7.2 Step 2: analyzing the covariance terms

In this section, we bound the covariance terms, $\Gamma_{s,j}$ and Γ_a , for $j \in [n]$.

Lemma 7.20. *Consider the modified GI/GI/n system satisfying Assumptions 7.1 and 7.3, with the load $\rho < 1$. Let $\Gamma_{s,j}$ and Γ_a be the terms defined in (7.58) and (7.59), respectively. Then*

$$\Gamma_{s,j} \leq \min\left(1 - \rho, \frac{1}{n}\right)\left(R_s^{\max} - \frac{1}{2}\mathbb{E}[(\mu S)^2]\right) \tag{7.66}$$

$$-\Gamma_a \leq (1 - \rho)\left(\frac{1}{2}\mathbb{E}[(\Lambda A)^2] - R_a^{\min}\right) \tag{7.67}$$

When proving Lemma 7.20, we focus on presenting the case of $\rho < 1 - 1/n$, since the case of $\rho \geq 1 - 1/n$ is relatively well understood, with the main argument sketched in Section 7.5.3.

Proof. Recall that

$$\Gamma_{s,j} = \mathbb{E}_s[\mathbb{1}\{Q = 0\}\mu R_{s,j}] - (1 - \rho)\mathbb{E}_\pi[\mu R_{s,j}] \quad \forall j \in [n] \quad (7.58)$$

$$\Gamma_a = \mathbb{E}_s[\mathbb{1}\{Q = 0\}\Lambda R_a] - (1 - \rho)\mathbb{E}_\pi[\Lambda R_a], \quad (7.59)$$

We first bound $\Gamma_{s,j}$ and then bound $-\Gamma_a$.

Bounding $\Gamma_{s,j}$ Observe that when $\rho < 1 - 1/n$, the load of the j -th leave-one-out system is less than 1, so it has a well-defined stationary distribution (Assumption 7.3). We consider the indicator $\mathbb{1}\{Q_{-j} = 0\}$ as a “baseline” and subtract it from $\mathbb{1}\{Q = 0\}$. This leads to the following *key decomposition* of $\mathbb{E}_s[\mathbb{1}\{Q = 0\}\mu R_{s,j}]$:

$$\begin{aligned} & \mathbb{E}_s[\mathbb{1}\{Q = 0\}\mu R_{s,j}] \\ &= \mathbb{E}_s\left[\left(\mathbb{1}\{Q = 0\} - \mathbb{1}\{Q_{-j} = 0\}\right)\mu R_{s,j}\right] + \mathbb{E}_s[\mathbb{1}\{Q_{-j} = 0\}\mu R_{s,j}]. \end{aligned} \quad (7.68)$$

We then bound the two terms on the right-hand side of (7.68) separately.

We bound the first term on the right-hand side of (7.68) using similar arguments as in Case 1, i.e., by applying the conditional expectation bound of $R_{s,j}$ proved in Lemma 7.14:

$$\begin{aligned} & \mathbb{E}_s\left[\left(\mathbb{1}\{Q = 0\} - \mathbb{1}\{Q_{-j} = 0\}\right)\mu R_{s,j}\right] \\ &= \frac{1}{n} \sum_{i: i \neq j} \mathbb{E}_{s,i}\left[\left(\mathbb{1}\{Q = 0\} - \mathbb{1}\{Q_{-j} = 0\}\right)\mu R_{s,j}\right] \end{aligned} \quad (7.69)$$

$$\leq \frac{1}{n} \sum_{i: i \neq j} \mathbb{E}_{s,i}[\mathbb{1}\{Q = 0\} - \mathbb{1}\{Q_{-j} = 0\}] R_s^{\max} \quad (7.70)$$

$$\leq \left(\mathbb{P}_s[Q = 0] - \frac{1}{n} \sum_{i: i \neq j} \mathbb{P}_{s,i}[Q_{-j} = 0]\right) R_s^{\max} \quad (7.71)$$

$$\begin{aligned} &= \left(1 - \rho - \left(1 - \rho - \frac{1}{n}\right)\right) R_s^{\max} \\ &= \frac{1}{n} R_s^{\max}. \end{aligned} \quad (7.72)$$

Specifically, the equality (7.69) holds because $R_{s,i} = 0$ almost surely under $\mathbb{P}_{s,i}$ (Lemma 7.12). Here, we have suppressed the superscript $-$ for variables inside Palm expectations. The next inequality (7.70) follows from $\mathbb{E}_{s,i}[\mu R_{s,j} \mid Q, Q_{-j}] \leq R_s^{\max}$ (Lemma 7.14) and $\mathbb{1}\{Q = 0\} - \mathbb{1}\{Q_{-j} = 0\} \geq 0$ (Lemma 7.11). Furthermore, (7.71) follows from the definition $\mathbb{P}_s[Q = 0] \triangleq \sum_{i=1}^n \mathbb{P}_{s,i}[Q = 0]/n$, while the equality (7.72) is obtained by substituting the values $\mathbb{P}_s[Q = 0] = 1 - \rho$ and $\sum_{i \neq j} \mathbb{P}_{s,i}[Q_{-j} = 0]/n = 1 - \rho - 1/n$ (Lemma 7.16).

To bound the second term on the right-hand side of (7.68), recall that $\mathbb{1}\{Q_{-j} = 0\}$ and $R_{s,j}$ are uncorrelated under the distribution $\mathbb{P}_{s,i}$ for $i \neq j$ (Lemma 7.15), so we can exactly calculate this term as

$$\begin{aligned} \mathbb{E}_s[\mathbb{1}\{Q_{-j} = 0\} \mu R_{s,j}] &= \frac{1}{n} \sum_{i: i \neq j} \mathbb{E}_{s,i}[\mathbb{1}\{Q_{-j} = 0\} \mu R_{s,j}] \\ &= \frac{1}{n} \sum_{i: i \neq j} \mathbb{P}_{s,i}[Q_{-j} = 0] \mathbb{E}_\pi[\mu R_{s,j}] \end{aligned} \quad (7.73)$$

$$= \left(1 - \rho - \frac{1}{n}\right) \mathbb{E}_\pi[\mu R_{s,j}], \quad (7.74)$$

where (7.73) is due to Lemma 7.15, and (7.74) is due to Lemma 7.16.

Substituting the above calculations back to (7.68), we get

$$\mathbb{E}_s[\mathbb{1}\{Q = 0\} \mu R_{s,j}] \leq \frac{1}{n} R_s^{\max} + \left(1 - \rho - \frac{1}{n}\right) \mathbb{E}_\pi[\mu R_{s,j}].$$

Therefore,

$$\Gamma_{s,j} = \mathbb{E}_s[\mathbb{1}\{Q = 0\} \mu R_{s,j}] - (1 - \rho) \mathbb{E}_\pi[\mu R_{s,j}] \leq \frac{1}{n} \left(R_s^{\max} - \mathbb{E}_\pi[\mu R_{s,j}]\right).$$

Because $\mathbb{E}_\pi[\mu R_{s,j}] = \mathbb{E}[(\mu S)^2]/2$, this finishes the proof of (7.66).

Bounding $-\Gamma_a$ Finally, we prove (7.67).

$$\begin{aligned} -\Gamma_a &= (1 - \rho) \mathbb{E}_\pi[\Lambda R_a] - \mathbb{E}_s[\mathbb{1}\{Q = 0\} \Lambda R_a] \\ &= (1 - \rho) \mathbb{E}_\pi[\Lambda R_a] - \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{s,i}[\mathbb{1}\{Q = 0\} \Lambda R_a] \\ &= (1 - \rho) \mathbb{E}_\pi[\Lambda R_a] - \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{s,i}[\Lambda R_a \mid Q = 0] \mathbb{P}_{s,i}[Q = 0] \end{aligned}$$

$$\leq (1 - \rho) \mathbb{E}_\pi[\Lambda R_a] - R_a^{\min} \mathbb{P}_s[Q = 0] \quad (7.75)$$

$$= (1 - \rho) \left(\frac{1}{2} \mathbb{E}[(\Lambda A)^2] - R_a^{\min} \right). \quad (7.76)$$

where (7.75) is due to the fact that $\mathbb{E}_{s,i}[\Lambda R_a \mid Q = 0] \geq R_a^{\min}$ (Lemma 7.14), and (7.76) follows from substituting the values $\mathbb{E}_\pi[\Lambda R_a] = \mathbb{E}[(\Lambda A)^2]/2$ and $\mathbb{P}_s[Q = 0] = 1 - \rho$ (Lemma 7.16). $\square$

7.8 Generalization to heterogeneous servers

In the previous sections, we have proved an $O(1/(1 - \rho))$ bound for the steady-state mean queue length of the $GI/GI/n$ queue with homogeneous servers. In this section, we generalize our bound to the $GI/GI/n$ queue with fully heterogeneous servers, where the servers could have mutually distinct service-time distributions.

The organization of this section roughly mirrors the high-level structure of Sections 7.3, 7.4, 7.6, and 7.7. Specifically, we first set up the $GI/GI/n$ queue and modified $GI/GI/n$ queue with heterogeneous servers in Section 7.8.1. Then in Section 7.8.2, we state the assumptions and the main theorem, Theorem 7.25. In Section 7.8.3, we establish the technical framework and state the supporting lemmas, which then allows us to prove Theorem 7.25 in Section 7.8.4.

7.8.1 $GI/GI/n$ queue and modified $GI/GI/n$ queue with heterogeneous servers

We consider the $GI/GI/n$ queue with heterogeneous servers under the FCFS policy. This queueing system consists of a central queue and n servers, where the service-time distributions could be *server-dependent*. Jobs arrive according to a renewal process. A newly-arrived job joins the end of the queue if all servers are busy, or start being served by one of the idle servers. When a job is in service, it occupies the server for a random amount of time following the service-time distribution of the server; the job leaves the system after completing the service. When a server completes a job, it immediately starts serving a job from the head of the queue if the queue is non-empty. We allow an *arbitrary routing policy* for assigning jobs to idle servers when a new job arrives or when multiple servers free up simultaneously.

The definitions of initial state and sample paths of the $GI/GI/n$ queue can be straightforwardly generalized to the heterogeneous setting, which we do not repeat here. We use A to denote a generic interarrival time, and use S_i to denote a generic service time of the i -th server. We denote the service rate of the i -th server as $\mu_i \triangleq 1/\mathbb{E}[S_i]$, and define the total service rate as $\mu_\Sigma \triangleq \sum_{i=1}^n \mu_i$. Let the load $\rho = \Lambda/\mu_\Sigma$. Let $R_a^{\min} \triangleq \inf_{t \geq 0} \mathbb{E}[\Lambda A - t \mid \Lambda A \geq t]$ and $R_{s,i}^{\max} \triangleq \sup_{t \geq 0} \mathbb{E}[\mu_i S_i - t \mid \mu_i S_i \geq t]$ for $i \in [n]$.

The $GI/GI/n$ queue with heterogeneous servers has the same state representation as the homogeneous setting: at any time $t \geq 0$, its state $X(t)$ is given by

$$X(t) \triangleq (R_{s,1}(t), R_{s,2}(t), \dots, R_{s,n}(t), R_a(t), Q(t)),$$

where $R_{s,i}(t)$ is the residual service time of the i -th server, $R_a(t)$ is the residual arrival time, and $Q(t)$ is the queue length. The state space of $\{X(t)\}_{t \geq 0}$ is $\mathcal{X} \triangleq \mathbb{R}^{n+1} \times \mathbb{Z}_{\geq 0}$.

As in the homogeneous setting, here we also consider the modified $GI/GI/n$ queue, which introduces a virtual job whenever a server frees up and the queue is empty. The modified system provides a queue-length upper bound for the original system, as shown in Lemma 7.21 below. We prove Lemma 7.21 in Appendix 7.C.

Lemma 7.21. *With the same initial states, at any time t , the queue length of the $GI/GI/n$ queue with heterogeneous servers is stochastically dominated by the queue length of the corresponding modified $GI/GI/n$ queue.*

With a slight abuse of notation, we also use $X(t)$ to denote the state of the modified $GI/GI/n$ system. Under the dynamics of the modified $GI/GI/n$ system, $R_{s,i}(t)$ is the forward recurrence time of a renewal process with the inter-event distribution S_i , for each $i \in [n]$; $R_a(t)$ is the forward recurrence time of a renewal process with the inter-event distribution A ; the queue length $Q(t)$ is determined by these $(n+1)$ independent renewal processes. The detailed update rule of $X(t)$ is identical to those summarized in Section 7.3.2.

7.8.2 Assumptions and results

Our results hold under the following three assumptions, which are analogous to Assumptions 7.1, 7.2, and 7.3 in the homogeneous setting. Note that in Assumption 7.24, the load of a modified $GI/GI/n$ queue refers to the ratio of its arrival rate and the total service rate of its servers.

Assumption 7.22. We assume that $\mathbb{E}[A^2] < \infty$ and $R_{s,i}^{\max} < \infty$ for all $i \in [n]$.

Assumption 7.23. For each $i \in [n]$, the service-time distribution S_i is non-lattice.

Assumption 7.24. Each modified $GI/GI/n$ queue with heterogeneous servers considered in this section is positive Harris recurrent if its load is less than 1. In addition, let ν be its unique stationary distribution, then $\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \mathbb{E}[Q(t)] dt = \mathbb{E}_\nu[Q] < \infty$.

Theorem 7.25. Consider the $GI/GI/n$ queue with heterogeneous servers under the FCFS policy, satisfying Assumptions 7.22, 7.23, and 7.24. When the load $\rho < 1$, let π be the unique stationary distribution of the system. Then the mean queue length under π satisfies

$$\begin{aligned} \mathbb{E}_\pi[Q] \leq & \frac{\rho \text{Var}[\Lambda A] + \sum_{j=1}^n (\mu_j / \mu_\Sigma) \text{Var}[\mu_j S_j] + 1 - \rho}{2(1 - \rho)} \\ & + \sum_{j=1}^n \min\left(1, \frac{\mu_j}{(1 - \rho)\mu_\Sigma}\right) \left(R_{s,j}^{\max} - \frac{1}{2} \mathbb{E}[(\mu_j S_j)^2]\right) \\ & + \frac{1}{2} \mathbb{E}[(\Lambda A)^2] - R_a^{\min}. \end{aligned} \quad (7.77)$$

The bound in (7.77) implies the following simpler bound after replacing each minimum with $\mu_j / ((1 - \rho)\mu_\Sigma)$:

$$\mathbb{E}_\pi[Q] \leq \frac{\rho \text{Var}[\Lambda A] - \rho + 2 \sum_{j=1}^n (\mu_j / \mu_\Sigma) R_{s,j}^{\max}}{2(1 - \rho)} + \frac{1}{2} \mathbb{E}[(\Lambda A)^2] - R_a^{\min}. \quad (7.78)$$

When the arrival process is Poisson, the bound in (7.78) further simplifies into

$$\mathbb{E}_\pi[Q] \leq \frac{\sum_{j=1}^n (\mu_j / \mu_\Sigma) R_{s,j}^{\max}}{1 - \rho}. \quad (7.79)$$

Notably, the bounds in (7.78) and (7.79) only depend on the service-time distributions through the *weighted average* of $R_{s,j}^{\max}$ across the servers. This makes these bounds relatively robust, in the sense that having a small fraction of servers with bad tail distributions does not significantly worsen the bounds. In particular, when $\text{Var}[\Lambda A]$ and $\sum_{j=1}^n (\mu_j / \mu_\Sigma) R_{s,j}^{\max}$ are bounded by constants, these bounds have the order $O(1/(1 - \rho))$ for all joint scalings of n and ρ .

7.8.3 Technical framework and supporting lemmas

This subsection has a similar structure as Section 7.6: we define leave-one-out systems, state the Basic Adjoint Relationship (BAR), and then state four additional supporting lemmas. We also adopt the following no-simultaneous-event assumption to simplify the presentation. The argument for removing this assumption is identical to that provided in Remark 7.13, so we do not repeat it here.

Assumption 7.26. In the modified $GI/GI/n$ queue with heterogeneous servers, no two events (i.e., an arrival and a completion, two arrivals, or two completions) happen at the same time.

We first define the leave-one-out systems in the same way as in Section 7.6.1: recall that for each $j \in [n]$, the j -th leave-one-out system is a modified $GI/GI/n$ queue whose queue length process, $Q_{-j}(t)$, is driven by the arrival process $R_a(t)$ and the completion processes $\{R_{s,i}(t)\}_{i \in [n] \setminus \{j\}}$ of the main modified $GI/GI/n$ queue under study.

The following lemma states that the queue length of each leave-one-out system pathwise dominates that of the main system. We omit the proof as it is identical to the one provided for the homogeneous case in Lemma 7.11.

Lemma 7.27. *Consider the modified $GI/GI/n$ queue with heterogeneous servers. For each $j = 1, 2, \dots, n$, the queue length of the j -th leave-one-out system pathwise dominates the queue length of the main system under study, i.e.,*

$$Q_{-j}(t) \geq Q(t) \quad \forall t \geq 0.$$

Our analysis focuses on the main modified $GI/GI/n$ system coupled with that subset of leave-one-out systems that have well-defined stationary distributions. Specifically, we denote the load of the j -th leave-one-out system as $\rho_{-j} \triangleq \Lambda/(\mu_\Sigma - \mu_j)$. We let $J^s \triangleq \{j \in [n] : \rho_{-j} < 1\}$ be the indices of leave-one-out systems with subcritical loads and let $\tilde{Q}^s(t) \triangleq (Q_{-j}(t))_{j \in J^s}$ denote the vector of their queue lengths at the time t .

We then define the state-augmented Markov process $\{\bar{X}(t)\}_{t \geq 0}$ as

$$\bar{X}(t) \triangleq (R_{s,1}(t), R_{s,2}(t), \dots, R_{s,n}(t), R_a(t), Q(t), (Q_{-j}(t))_{j \in J^s}). \quad (7.80)$$

The state space of $\{\bar{X}(t)\}_{t \geq 0}$ is thus given by $\bar{\mathcal{X}} \triangleq \mathbb{R}_{\geq 0}^{n+1} \times \mathbb{Z}_{\geq 0}^{1+|J^s|}$.

With the differential operator $\mathcal{A}$ and the Palm distributions defined in the same way as in Section 7.6.2, we can prove BAR for the heterogeneous setting:

Lemma 7.28 (Basic Adjoint Relationship for $GI/GI/n$ with heterogeneous servers). *Consider the modified $GI/GI/n$ system with heterogeneous servers. Suppose Assumptions 7.22, 7.23 and 7.24 hold, and the load $\rho < 1$. For any function $f: \bar{\mathcal{X}} \rightarrow \mathbb{R}$ that is bounded, locally Lipschitz continuous, and partially differentiable from the right with respect to its first $(n+1)$ coordinates, we have*

$$\mathbb{E}_\pi[\mathcal{A}f(\bar{X})] + \Lambda \mathbb{E}_a[f(\bar{X}^+) - f(\bar{X}^-)] + \sum_{i=1}^n \mu_i \mathbb{E}_{s,i}[f(\bar{X}^+) - f(\bar{X}^-)] = 0, \quad (7.81)$$

where $\bar{X}, \bar{X}^+, \bar{X}^- \in \bar{\mathcal{X}}$ are random elements following the distributions specified in the expectation subscripts. Moreover, the Palm distributions $\mathbb{P}_a$ and $\mathbb{P}_{s,i}$ satisfy the following with probability 1:

- (i) If $(\bar{X}^+, \bar{X}^-) \sim \mathbb{P}_a$, then $R_a^- = 0$ and the coordinates of $\bar{X}^+$ and $\bar{X}^-$ satisfy the relations:

$$R_a^+ = R_a^- + A \quad (7.82)$$

$$Q^+ = Q^- + 1 \quad (7.83)$$

$$Q_{-j}^+ = Q_{-j}^- + 1 \quad \forall j \in J^s, \quad (7.84)$$

where A represents a generic interarrival time and is independent of $\bar{X}^-$; $J^s \triangleq \{j \in [n]: \rho_{-j} < 1\}$; other coordinates of $\bar{X}^+$ are identical to $\bar{X}^-$.

- (ii) For each $i \in [n]$, if $(\bar{X}^+, \bar{X}^-) \sim \mathbb{P}_{s,i}$, then $R_{s,i}^- = 0$ and the coordinates of $\bar{X}^+$ and $\bar{X}^-$ satisfy the relations:

$$R_{s,i}^+ = R_{s,i}^- + S_i \quad (7.85)$$

$$Q^+ = Q^- - \mathbf{1}\{Q^- > 0\} \quad (7.86)$$

$$Q_{-j}^+ = Q_{-j}^- - \mathbf{1}\{Q_{-j}^- > 0\} \quad \forall j \in J^s \setminus \{i\}, \quad (7.87)$$

where S_i represents a generic service time of the i -th server and is independent of $\bar{X}^-$; other coordinates of $\bar{X}^+$ are identical to $\bar{X}^-$.

Lemma 7.28 is proved in Appendix 7.D.2. To simplify the notation, sometimes we would equivalently write the main equation of BAR (7.81) as

$$\mathbb{E}_\pi[\mathcal{A}f(\bar{X})] + \Lambda \mathbb{E}_a[f(\bar{X}^+) - f(\bar{X}^-)] + \mu_\Sigma \mathbb{E}_s[f(\bar{X}^+) - f(\bar{X}^-)] = 0, \quad (7.88)$$

where

$$\mathbb{E}_s[\cdot] \triangleq \sum_{i=1}^n \frac{\mu_i}{\mu_\Sigma} \mathbb{E}_{s,i}[\cdot].$$

We omit the superscript $-$ from $\bar{X}^-$ when it does not appear together with $\bar{X}^+$.

Next, we state the four additional supporting lemmas that generalize Lemmas 7.14, 7.15, 7.16, and 7.17 to the heterogeneous case. Their proofs are provided in Appendix 7.D.

Lemma 7.29 (Bounded mean residual times, heterogeneous). *Consider the modified GI/GI/n queue with heterogeneous servers. Suppose Assumptions 7.22 and 7.24 hold and the load $\rho < 1$. Then for any $j \in [n]$,*

$$\mu_j \mathbb{E}_\nu [R_{s,j} \mid Q, \tilde{Q}^s] \leq R_{s,j}^{\max} \quad (7.48)$$

$$\Lambda \mathbb{E}_\nu [R_a \mid Q, \tilde{Q}^s] \geq R_a^{\min}. \quad (7.49)$$

where $\mathbb{E}_\nu[\cdot]$ can denote either $\mathbb{E}_\pi[\cdot]$ or $\mathbb{E}_{s,i}[\cdot]$ for $i \neq j$.

Lemma 7.30 (Uncorrelation lemma, heterogeneous). *Consider the modified GI/GI/n queue with heterogeneous servers. Suppose Assumptions 7.22, 7.23, and 7.24 hold. Then for any $i, j \in [n]$ such that $i \neq j$ and $\rho_{-j} < 1$,*

$$\mathbb{E}_{s,i}[\mathbf{1}\{Q_{-j} = 0\} R_{s,j}] = \mathbb{P}_{s,i}[Q_{-j} = 0] \mathbb{E}_\pi[R_{s,j}]. \quad (7.89)$$

Lemma 7.31. *Consider the modified GI/GI/n queue with heterogeneous servers. Suppose Assumptions 7.22 and 7.24 hold, and the load $\rho < 1$. For any $i, j \in [n]$ such that $i \neq j$, we have*

$$\mathbb{E}_\pi[R_{s,i}] = \frac{1}{2} \mu_i \mathbb{E}[S_i^2] \quad (7.90)$$

$$\mathbb{E}_\pi[R_a] = \frac{1}{2} \Lambda \mathbb{E}[A^2] \quad (7.91)$$

$$\mathbb{E}_{s,i}[R_{s,j}] + \mathbb{E}_{s,j}[R_{s,i}] = \frac{1}{2} \mu_i \mathbb{E}[S_i^2] + \frac{1}{2} \mu_j \mathbb{E}[S_j^2] \quad (7.92)$$

$$\mathbb{E}_{s,i}[R_a] + \mathbb{E}_a[R_{s,i}] = \frac{1}{2} \mu_i \mathbb{E}[S_i^2] + \frac{1}{2} \Lambda \mathbb{E}[A^2] \quad (7.93)$$

$$\mathbb{P}_s[Q = 0] = 1 - \rho \quad (7.94)$$

In addition, for each $j \in [n]$, if $\rho_{-j} < 1$, the j -th leave-one-out system satisfies

$$\frac{1}{\mu_\Sigma} \sum_{i: i \neq j} \mu_i \mathbb{P}_{s,i}[Q_{-j} = 0] = 1 - \rho - \frac{\mu_j}{\mu_\Sigma}. \quad (7.95)$$

Lemma 7.32. *Consider the modified GI/GI/n system with heterogeneous servers. Suppose Assumptions 7.22 and 7.24 hold, and the load $\rho < 1$. Then we have $\mathbb{E}_a[Q] < \infty$ and $\mathbb{E}_{s,i}[Q] < \infty$ for all $i \in [n]$.*

7.8.4 Proof of Theorem 7.25

We are now ready to prove Theorem 7.25. The proof follows the same two-step structure used for the homogeneous case (Section 7.7), with the changes being primarily notational. We first derive an expression of $\mathbb{E}_\pi[Q]$ involving covariance terms (Lemma 7.33), and then bound these terms (Lemma 7.34) to establish the theorem. To avoid repeating the detailed arguments from the previous section, we present only the key steps of the proof and omit explicit citations to some of the corresponding lemmas.

Lemma 7.33. *Consider the modified GI/GI/n system satisfying Assumptions 7.22 and 7.24, with the load $\rho < 1$. We have*

$$(1 - \rho)\mathbb{E}_\pi[Q] = \frac{\rho \text{Var}[\Lambda A] + \sum_{i=1}^n (\mu_i / \mu_\Sigma) \text{Var}[\mu_i S_i] + 1 - \rho}{2} + \sum_{j=1}^n \Gamma_{s,j} - \Gamma_a, \quad (7.96)$$

where $\Gamma_{s,j}$ and Γ_a are defined as

$$\Gamma_{s,j} = \mathbb{E}_s[\mathbf{1}\{Q = 0\} \mu_j R_{s,j}] - (1 - \rho) \mathbb{E}_\pi[\mu_j R_{s,j}] \quad \forall j \in [n] \quad (7.97)$$

$$\Gamma_a = \mathbb{E}_s[\mathbf{1}\{Q = 0\} \Lambda R_a] - (1 - \rho) \mathbb{E}_\pi[\Lambda R_a]. \quad (7.98)$$

Proof. For any integer $\kappa > 0$, we apply (7.88) to the test function

$$f(\bar{X}) = \left(Q \wedge \kappa + \sum_{j=1}^n \mu_j R_{s,j} \wedge \kappa - \Lambda R_a \wedge \kappa \right)^2,$$

rearrange the terms. After taking $\kappa \rightarrow \infty$, we get

$$2(\mu_\Sigma - \Lambda) \mathbb{E}_\pi \left[Q + \sum_{j=1}^n \mu_j R_{s,j} - \Lambda R_a \right] \quad (7.99)$$

$$= \Lambda \mathbb{E}_a \left[2 \left(Q + \sum_{j=1}^n \mu_j R_{s,j} \right) (1 - \Lambda A) + (1 - \Lambda A)^2 \right] \quad (7.100)$$

$$+ \sum_{i=1}^n \mu_i \mathbb{E}_{s,i} \left[2 \left(Q + \sum_{j=1}^n \mu_j R_{s,j} - \Lambda R_a \right) (\mu_i S_i - \mathbf{1}\{Q > 0\}) + (\mu_i S_i - \mathbf{1}\{Q > 0\})^2 \right]. \quad (7.101)$$

The interchange of the limit $\kappa \rightarrow \infty$ and the expectation is justified by an argument analogous to that in the proof of Lemma 7.18 (see Section 7.7.1). The key non-trivial condition for this step is the finiteness of the moments $\mathbb{E}_\pi[Q]$, $\mathbb{E}_a[Q]$, and $\mathbb{E}_{s,i}[Q]$, which is guaranteed by Assumption 7.24 and Lemma 7.32.

We now simplify the expressions in (7.100) and (7.101). First, due to the independence between A and $(Q, (R_{s,j})_{j \in [n]})$, the terms in (7.100) simplifies into $\Lambda \text{Var}[\Lambda A]$. Second, due to the independence between S_i and $(Q, R_a, (R_{s,j})_{j \in [n]})$, the summation in (7.101) simplifies into:

$$2\mu_i \mathbb{E}_{s,i} \left[\left(\sum_{j=1}^n \mu_j R_{s,j} - \Lambda R_a \right) \mathbb{1}\{Q = 0\} \right] + \mu_i \text{Var}[\mu_i S_i] + \mu_i(1 - \rho).$$

Substituting these expressions back into (7.100) and (7.101) yields the following:

$$\begin{aligned} & 2(\mu_\Sigma - \Lambda) \mathbb{E}_\pi \left[Q + \sum_{j=1}^n \mu_j R_{s,j} - \Lambda R_a \right] \\ &= 2\mu_\Sigma \mathbb{E}_s \left[\left(\sum_{j=1}^n \mu_j R_{s,j} - \Lambda R_a \right) \mathbb{1}\{Q = 0\} \right] + \Lambda \text{Var}[\Lambda A] \\ & \quad + \sum_{i=1}^n \mu_i \text{Var}[\mu_i S_i] + \mu_\Sigma(1 - \rho). \end{aligned}$$

Dividing both sides by $2\mu_\Sigma$ and rearranging the terms complete the proof of Lemma 7.33. $\square$

Lemma 7.34. *Consider the modified GI/GI/n system satisfying Assumptions 7.22, 7.23 and 7.24, with the load $\rho < 1$. Let $\Gamma_{s,j}$ and Γ_a be the terms defined in (7.97) and (7.98), respectively. Then*

$$\Gamma_{s,j} \leq \min \left(1 - \rho, \frac{\mu_j}{\mu_\Sigma} \right) \left(R_{s,j}^{\max} - \frac{1}{2} \mathbb{E}[(\mu_j S_j)^2] \right) \quad (7.102)$$

$$-\Gamma_a \leq (1 - \rho) \left(\frac{1}{2} \mathbb{E}[(\Lambda A)^2] - R_a^{\min} \right) \quad (7.103)$$

Proof. We first bound $\Gamma_{s,j}$ based on two cases: $\rho_{-j} \geq 1$ or $\rho_{-j} < 1$; then we bound $-\Gamma_a$.

Bounding $\Gamma_{s,j}$ when $\rho_{-j} \geq 1$ In this case, $\Lambda/(\mu_\Sigma - \mu_j) \triangleq \rho_{-j} \geq 1$, so we have $\Lambda \geq \mu_\Sigma - \mu_j$. Consequently, $1 - \rho = 1 - \Lambda/\mu_\Sigma \leq \mu_j/\mu_\Sigma$. It is thus sufficient to show that

$$\Gamma_{s,j} \leq (1 - \rho) \left(R_{s,j}^{\max} - \frac{1}{2} \mathbb{E}[(\mu_j S_j)^2] \right).$$

We bound $\Gamma_{s,j}$ as follows:

$$\begin{aligned} \Gamma_{s,j} &= \mathbb{E}_s[\mathbb{1}\{Q = 0\} \mu_j R_{s,j}] - (1 - \rho) \mathbb{E}_\pi[\mu_j R_{s,j}] \\ &= \frac{1}{\mu_\Sigma} \sum_{i=1}^n \mu_i \mathbb{E}_{s,i}[\mathbb{1}\{Q = 0\} \mu_j R_{s,j}] - (1 - \rho) \mathbb{E}_\pi[\mu_j R_{s,j}] \\ &= \frac{1}{\mu_\Sigma} \sum_{i: i \neq j} \mu_i \mathbb{E}_{s,i}[\mathbb{1}\{Q = 0\} \mu_j R_{s,j}] - (1 - \rho) \mathbb{E}_\pi[\mu_j R_{s,j}] \end{aligned} \quad (7.104)$$

$$\leq \frac{1}{\mu_\Sigma} \sum_{i: i \neq j} \mu_i \mathbb{P}_{s,i}[Q = 0] R_{s,j}^{\max} - (1 - \rho) \mathbb{E}_\pi[\mu_j R_{s,j}] \quad (7.105)$$

$$\begin{aligned} &\leq \mathbb{P}_s[Q = 0] R_{s,j}^{\max} - (1 - \rho) \mathbb{E}_\pi[\mu_j R_{s,j}] \\ &= (1 - \rho) R_{s,j}^{\max} - (1 - \rho) \mathbb{E}_\pi[\mu_j R_{s,j}] \end{aligned} \quad (7.106)$$

$$= (1 - \rho) \left(R_{s,j}^{\max} - \frac{1}{2} \mathbb{E}[(\mu_j S_j)^2] \right), \quad (7.107)$$

where (7.104) is because $R_{s,i} = 0$ almost surely under the distribution $\mathbb{P}_{s,i}$, as proved in Lemma 7.28 (we have skipped all superscripts $-$ for variables inside the Palm expectations in the above derivations). Furthermore, the inequality (7.105) uses the fact that $\mathbb{E}_{s,i}[\mu_j R_{s,j} | Q] \leq R_{s,j}^{\max}$ for $i \neq j$ (Lemma 7.29). Finally, the equalities (7.106) and (7.107) follow from substituting the values $\mathbb{P}_s[Q = 0] = 1 - \rho$ and $\mathbb{E}_\pi[\mu_j R_{s,j}] = \mathbb{E}[(\mu_j S_j)^2]/2$. This implies (7.102) when $\rho_{-j} \geq 1$.

Bounding $\Gamma_{s,j}$ when $\rho_{-j} < 1$ In this case, $1 - \rho > \mu_j/\mu_\Sigma$, so it suffices to show that

$$\Gamma_{s,j} \leq \frac{\mu_j}{\mu_\Sigma} \left(R_{s,j}^{\max} - \frac{1}{2} \mathbb{E}[(\mu_j S_j)^2] \right).$$

Because $\rho_{-j} < 1$, by Assumption 7.24, the j -th leave-one-out system has a well-defined stationary distribution. We can thus perform the following decomposition:

$$\begin{aligned} &\mathbb{E}_s[\mathbb{1}\{Q = 0\} \mu_j R_{s,j}] \\ &= \mathbb{E}_s \left[\left(\mathbb{1}\{Q = 0\} - \mathbb{1}\{Q_{-j} = 0\} \right) \mu_j R_{s,j} \right] + \mathbb{E}_s[\mathbb{1}\{Q_{-j} = 0\} \mu_j R_{s,j}]. \end{aligned} \quad (7.108)$$

We then bound the two terms on the right-hand side of (7.108) separately.

We bound the first term on the right-hand side of (7.108) by applying the conditional expectation bound of $R_{s,j}$ in Lemma 7.29, and note that $\mathbb{1}\{Q = 0\} - \mathbb{1}\{Q_{-j} = 0\} \geq 0$ (Lemma 7.27):

$$\begin{aligned}
& \mathbb{E}_s \left[\left(\mathbb{1}\{Q = 0\} - \mathbb{1}\{Q_{-j} = 0\} \right) \mu_j R_{s,j} \right] \\
&= \sum_{i=1}^n \frac{\mu_i}{\mu_\Sigma} \mathbb{E}_{s,i} \left[\left(\mathbb{1}\{Q = 0\} - \mathbb{1}\{Q_{-j} = 0\} \right) \mu_j R_{s,j} \right] \\
&\leq \sum_{i: i \neq j} \frac{\mu_i}{\mu_\Sigma} \mathbb{E}_{s,i} \left[\left(\mathbb{1}\{Q = 0\} - \mathbb{1}\{Q_{-j} = 0\} \right) \right] R_{s,j}^{\max} \quad (7.109) \\
&\leq \left(\mathbb{P}_s [Q = 0] - \frac{1}{\mu_\Sigma} \sum_{i: i \neq j} \mu_i \mathbb{P}_{s,i} [Q_{-j} = 0] \right) R_{s,j}^{\max} \\
&= \left(1 - \rho - \left(1 - \rho - \frac{\mu_j}{\mu_\Sigma} \right) \right) R_{s,j}^{\max} \\
&= \frac{\mu_j}{\mu_\Sigma} R_{s,j}^{\max}.
\end{aligned}$$

where the inequality (7.109) follows from three facts: $R_{s,i} = 0$ almost surely under $\mathbb{P}_{s,i}$ (Lemma 7.28), $\mathbb{E}_{s,i} [\mu_j R_{s,j} \mid Q, \tilde{Q}^s] \leq R_{s,j}^{\max}$ for $i \neq j$ (Lemma 7.29), and $\mathbb{1}\{Q = 0\} \geq \mathbb{1}\{Q_{-j} = 0\}$.

For the second term on the right-hand side of (7.108), recall that $\mathbb{1}\{Q_{-j} = 0\}$ and $R_{s,j}$ are uncorrelated under the distribution $\mathbb{P}_{s,i}$ for $i \neq j$ (Lemma 7.30). Therefore,

$$\begin{aligned}
\mathbb{E}_s [\mathbb{1}\{Q_{-j} = 0\} \mu_j R_{s,j}] &= \sum_{i: i \neq j} \frac{\mu_i}{\mu_\Sigma} \mathbb{E}_{s,i} [\mathbb{1}\{Q_{-j} = 0\} \mu_j R_{s,j}] \\
&= \sum_{i: i \neq j} \frac{\mu_i}{\mu_\Sigma} \mathbb{P}_{s,i} [Q_{-j} = 0] \mathbb{E}_\pi [\mu_j R_{s,j}] \\
&= \left(1 - \rho - \frac{\mu_j}{\mu_\Sigma} \right) \mathbb{E}_\pi [\mu_j R_{s,j}].
\end{aligned}$$

Substituting the above calculation into (7.108), we get

$$\mathbb{E}_s [\mathbb{1}\{Q = 0\} \mu_j R_{s,j}] \leq \frac{\mu_j}{\mu_\Sigma} R_{s,j}^{\max} + \left(1 - \rho - \frac{\mu_j}{\mu_\Sigma} \right) \mathbb{E}_\pi [\mu_j R_{s,j}].$$

Therefore,

$$\Gamma_{s,j} = \mathbb{E}_s [\mathbb{1}\{Q = 0\} \mu_j R_{s,j}] - (1 - \rho) \mathbb{E}_\pi [\mu_j R_{s,j}] \leq \frac{\mu_j}{\mu_\Sigma} \left(R_{s,j}^{\max} - \mathbb{E}_\pi [\mu_j R_{s,j}] \right).$$

Because $\mathbb{E}_\pi[\mu_j R_{s,j}] = \mathbb{E}[(\mu_j S_j)^2]/2$, this finishes the proof of (7.102) when $\rho_{-j} < 1$.

Bounding $-\Gamma_a$

$$\begin{aligned} -\Gamma_a &= (1 - \rho)\mathbb{E}_\pi[\Lambda R_a] - \mathbb{E}_s[\mathbb{1}\{Q = 0\}\Lambda R_a] \\ &= (1 - \rho)\mathbb{E}_\pi[\Lambda R_a] - \sum_{i=1}^n \frac{\mu_i}{\mu_\Sigma} \mathbb{E}_{s,i}[\mathbb{1}\{Q = 0\}\Lambda R_a] \\ &\leq (1 - \rho)\mathbb{E}_\pi[\Lambda R_a] - (1 - \rho)R_a^{\min} \\ &= (1 - \rho)\left(\frac{1}{2}\mathbb{E}[(\Lambda A)^2] - R_a^{\min}\right). \end{aligned}$$

where the inequality is due to the conditional expectation lower bound of R_a in Lemma 7.29 and the fact that $\mathbb{P}_s[Q = 0] = 1 - \rho$. $\square$

7.9 Conclusion

In this chapter, we prove a new universal bound on the steady-state mean queue length for the $GI/GI/n$ queue. Our bound achieves the order $O(1/(1 - \rho))$ and, while requiring slightly stronger assumptions than those in [180], provides significantly smaller constants. Our proof synthesizes three key elements: a modified $GI/GI/n$ queue, the Basic Adjoint Relationship (BAR), and our new leave-one-out technique. We also extend our result to the $GI/GI/n$ queue with fully heterogeneous servers.

A natural direction for future work is to investigate the tightness of our bounds. While it may be unrealistic to expect any simple closed-form bound to be arbitrarily tight for the $GI/GI/n$ queue, there is potential for improvement. An ideal bound might resemble Kingman's bound, depending only on the low-order moments of the service-time distribution. Such a bound would yield smaller constants for heavy-tailed systems. Furthermore, in sub-Halfin-Whitt regimes, research in special cases suggests that better orders than $O(1/(1 - \rho))$ may be possible [e.g., 48, 231].

Another promising direction is to apply our proof techniques to other large-scale queueing systems, such as load-balancing models with general service-time distributions.

Finally, it would be interesting to explore the theoretical connections between the leave-one-out technique and other existing techniques for proving

asymptotic independence [e.g. 270]. Clarifying their similarities and respective scopes would enrich the analytical toolbox for queueing systems.

7.A Bounded mean residual time implies exponential tail

In this section, we show that if a distribution has a bounded mean residual time given any age, it must have an exponentially decaying tail, so all of its moments must be finite. This fact has also been proved in [85]. Given this fact, Assumption 7.1 and Assumption 7.22 automatically imply that the service-time distributions have exponentially decaying tails and finite moments.

Lemma 7.35. *For any non-negative random variable V , suppose there exists a constant $C > 0$ such that*

$$\mathbb{E}[V - t \mid V \geq t] \leq C \quad \forall t \geq 0, \quad (7.110)$$

Then we have

$$\mathbb{P}[V \geq t] \leq C \exp(-(t-1)/C) \quad \forall t \geq 1, \quad (7.111)$$

and consequently, $\mathbb{E}[V^m] < \infty$ for all $m \in \mathbb{Z}_{\geq 0}$.

Proof. Let $\bar{F}(t)$ be the complementary CDF of V . The condition (7.110) can be rewritten as

$$\int_t^\infty \bar{F}(u) du \leq C \bar{F}(t) \quad \forall t \geq 0. \quad (7.112)$$

We will mimic the proof of Gronwall's inequality to derive an exponential tail bound for $\bar{F}(t)$. We define the function $G(t)$ as

$$G(t) = \exp(t/C) \int_t^\infty \bar{F}(u) du.$$

Because $\bar{F}(t)$ is bounded on $[0, \infty)$, it is not hard to see that $\int_t^\infty \bar{F}(u) du$ is locally Lipschitz continuous, and thus almost everywhere differentiable. For any t at which $\int_t^\infty \bar{F}(u) du$ is differentiable, $G(t)$ is also differentiable, and $G'(t)$ is given by

$$G'(t) = \exp(t/C) \left(\frac{1}{C} \int_t^\infty \bar{F}(u) du - \bar{F}(t) \right),$$

which is non-positive by (7.112). Moreover, $G(t)$ is also locally Lipschitz continuous. Therefore, on any compact interval, $G(t)$ is absolutely continuous and satisfies the fundamental theorem of calculus. In particular,

$$G(t) = G(0) + \int_0^t G'(u)du \leq G(0) = \int_0^\infty \bar{F}(u)du \leq C\bar{F}(0) = C,$$

where the second inequality is due to (7.112). By the definition of $G(t)$, we have

$$\int_t^\infty \bar{F}(u)du \leq C \exp(-t/C).$$

Because $\bar{F}(t)$ is non-increasing and non-negative, we can bound $\bar{F}(t+1)$ as

$$\bar{F}(t+1) \leq \int_t^{t+1} \bar{F}(u)du \leq C \exp(-t/C).$$

This proves (7.111). The finiteness of $\mathbb{E}[V^m]$ follows straightforwardly from the exponential tail bound. $\square$

7.B Stability of modified $GI/GI/n$

In this section, we discuss the conditions under which the modified $GI/GI/n$ queue is positive Harris recurrent and has a finite mean queue length. As mentioned in Remark 7.4, in addition to $\rho < 1$, some regularity conditions are required on the distributions of the interarrival time and service times. Our results have been proved by directly assuming the positive Harris recurrence and the finiteness of mean queue length (Assumptions 7.3 and 7.22). In this section, we will provide a more detailed discussion on these conditions, and sketch why they are likely to imply Assumptions 7.3 and 7.22, without giving formal proofs. Throughout this section, we focus on modified $GI/GI/n$ queues with heterogeneous servers and simply refer to them as “modified $GI/GI/n$ queue”.

We consider adapting the results of Dai and Meyn [74], which considers queueing networks consisting of single-server stations. In particular, their Theorem 6.2 shows that if such a queueing network is fluid stable and satisfies certain conditions on the interarrival-time and service-time distributions [74, Conditions A1)–A3)], it is positive Harris recurrent, with certain finite steady-state moments. We conjecture that a similar result can be shown for

the modified $GI/GI/n$ queue. Specifically, in addition to Assumption 7.22, we further assume that the interarrival-time distribution A is *spread out and unbounded*, and the service-time distribution S_i is *spread out* for each server $i \in [n]$ (a distribution ν is spread out if there exists n such that the distribution of the sum of n independent variables following ν dominates a non-negative absolutely-continuous measure; see 16, Section VII.1). These assumptions are analogous to Conditions A1)–A3) of [74]. We conjecture that, under these assumptions, a modified $GI/GI/n$ queue is positive Harris recurrent with finite mean queue length whenever $\rho < 1$.

Instead of providing a rigorous full proof, we highlight an intermediate step to provide intuition on why our conjectured result is likely to be true. In the proof of [74, Proposition 5.1], an important fact is that the queue lengths of the queueing network converge to zero in probability under fluid scaling, i.e.,

$$\lim_{|x| \rightarrow \infty} \frac{1}{|x|} |Q^x(t_0 | x)| = 0 \quad \text{in probability,} \quad (7.113)$$

where x is a state of the queueing network, $|x|$ represents the sum of all queue lengths and residual times in state x , and $Q^x(t)$ represents the queue length process with the initial state x . This fact follows from the assumed stability of the queueing network's fluid model, utilizing the results in [69, Section IV].

To show (7.113) for the modified $GI/GI/n$ queue, we invoke the following formula of queue length:

$$Q(t) = \max \left\{ Q(0) + N_a(t) - \sum_{i=1}^n N_i(t), \sup_{0 \leq t' \leq t} \left(N_a(t', t) - \sum_{i=1}^n N_i(t', t) \right) \right\}, \quad (7.114)$$

where $N_a(t)$ denotes the number of real-job arrivals by time t , $N_i(t)$ denotes the number of completions at server i by time t , and $N_a(t', t) \triangleq N_a(t) - N_a(t')$, $N_i(t', t) \triangleq N_i(t) - N_i(t')$. The special case of (7.114) is proved in [97, Corollary 1], under the additional assumptions that $Q(0) = 0$, the servers are homogeneous and no events occur simultaneously; dropping these assumptions is straightforward from their proof.

We intuitively argue that (7.114) implies (7.113): Let $x = (r_1, \dots, r_n, r_a, q)$ be the initial state of the modified $GI/GI/n$ queue. Then (7.114) implies that

$$\begin{aligned} \frac{1}{|x|} Q^x(t_0 |x|) = \max & \left\{ \frac{1}{|x|} \left(q + N_a(t_0 |x|) - \sum_{i=1}^n N_i(t_0 |x|) \right), \right. \\ & \left. \frac{1}{|x|} \sup_{0 \leq t' \leq t_0 |x|} \left(N_a(t', t_0 |x|) - \sum_{i=1}^n N_i(t', t_0 |x|) \right) \right\}. \end{aligned} \quad (7.115)$$

Due to the properties of renewal processes, as $|x|$ gets large, $N_a(t_0 |x|) - \sum_{i=1}^n N_i(t_0 |x|)$ is concentrated around its mean $\Lambda t_0 |x| - \mu_\Sigma t_0 |x|$, if the residual times are initialized at equilibrium. For other initial residual times, with high probability, $N_a(t_0 |x|) - \sum_{i=1}^n N_i(t_0 |x|)$ should not be significantly greater than

$$\Lambda t_0 |x| + 1 - \mu_\Sigma \left(t_0 |x| - \sum_{i=1}^n r_i \right) + o(|x|) = -(\mu_\Sigma - \Lambda) t_0 |x| + 1 + \sum_{i=1}^n \mu_i r_i + o(|x|).$$

Therefore, intuitively,

$$\begin{aligned} & \frac{1}{|x|} \left(q + N_a(t_0 |x|) - \sum_{i=1}^n N_i(t_0 |x|) \right) \\ & \leq \frac{q - (\mu_\Sigma - \Lambda) t_0 |x| + 1 + \sum_{i=1}^n \mu_i r_i + o(|x|)}{|x|} \\ & \leq \max_i \mu_i - (\mu_\Sigma - \Lambda) t_0 + o(1) \quad \text{with high probability,} \end{aligned}$$

So, taking t_0 to be a sufficiently large constant, the first term inside the maximum of (7.115) is negative with high probability. Moreover, $\sup_{0 \leq t' \leq t_0 |x|} (N_a(t', t_0 |x|) - \sum_{i=1}^n N_i(t', t_0 |x|))$ should also scale sublinearly in $|x|$, so the second term inside the maximum of (7.115) converges to zero in probability as $|x| \rightarrow \infty$. We have thus informally justified (7.113) for the modified $GI/GI/n$ queue.

Intuitively, (7.113) implies that no matter how large the initial state x is, the queue length becomes significantly shorter after a time proportional to $|x|$. After some technical work that mimics those in Proposition 5.1, 5.3, and 5.4 of [74], one can show that the state of the modified $GI/GI/n$ queue would return to a compact subset of the state space reasonably fast and thus

satisfy a certain drift condition (Proposition 6.1 of [74]). Furthermore, the spread out and unbounded conditions of A and S_i 's can be used to show that compact subsets under the system's dynamics are petite. Theorem 5.3 of [207] then implies that the modified $GI/GI/n$ queue is positive Harris recurrent, with a finite mean queue length.

7.C Proof of modified $GI/GI/n$ queue's dominance

In this section, we prove that the queue length distribution of the modified $GI/GI/n$ queue stochastically dominates that of the $GI/GI/n$ queue. We focus on the general case of heterogeneous servers. The proof for homogeneous $GI/GI/n$ can be also found in [97, Proposition 1]; our proof follows essentially the same logic, but we include it for completeness.

Lemma 7.21. *With the same initial states, at any time t , the queue length of the $GI/GI/n$ queue with heterogeneous servers is stochastically dominated by the queue length of the corresponding modified $GI/GI/n$ queue.*

Proof. We prove the lemma by constructing a coupling for the sample paths of the $GI/GI/n$ queue and the modified $GI/GI/n$ queue. First, by the assumption of the lemma, two systems have the same initial queue length $Q(0)$, the same set of initially busy servers that are serving real jobs $\mathcal{I}(0)$, and the same initial residual service times for the real jobs in service $(R_{s,i}(0))_{i \in \mathcal{I}(0)}$. For each $k \in \mathbb{Z}_{\geq 1}$, consider the k -th real job that begins service after time 0. We let this job arrive at time τ_a^k in both systems. This job may begin service at different times in the original system and in the modified system, denoted as τ_s^k and $\hat{\tau}_s^k$, respectively. However, when this job begins service, we route it to the same server and let it have the same service time S^k in both systems. Note that here, since we allow the service time distributions to be heterogeneous across servers, the sequence $\{S^k\}_{k=1}^\infty$ is not necessarily i.i.d., but this does not complicate our proof.

Let $Q(t)$ and $\hat{Q}(t)$ denote the queue lengths in the original and modified systems, respectively. They can be represented as

$$Q(t) = |\{k \in \mathbb{Z}_{\geq 1} : \tau_a^k \leq t, \tau_s^k > t\}| \quad (7.116)$$

$$\hat{Q}(t) = |\{k \in \mathbb{Z}_{\geq 1} : \tau_a^k \leq t, \hat{\tau}_s^k > t\}|. \quad (7.117)$$

Our goal is to show that $Q(t) \leq \hat{Q}(t)$ for all $t \geq 0$. By (7.116) and (7.117), it suffices to show that $\tau_s^k \leq \hat{\tau}_s^k$ for all $k \in \mathbb{Z}_{\geq 1}$.

We define some additional notation. For $k \in \mathbb{Z}_{\geq 1}$, we let $Z^k(t)$ be the number of jobs that begin service no later than the k -th job, but have not left the system by time t in the original system. We define $\hat{Z}^k(t)$ similarly for the real jobs in the modified system (excluding the virtual jobs). We let $Z^0(t)$ and $\hat{Z}^0(t)$ be the number of real jobs initially in the two systems and have not left by time t . Formally, for $k \in \mathbb{Z}_{\geq 0}$,

$$\begin{aligned} Z^k(t) &\triangleq \left| \{k' \in \mathbb{Z}_{\geq 1} : k' \leq k, \tau_s^{k'} + S^{k'} > t\} \right| + |\{i \in \mathcal{I} : R_{s,i}(0) > t\}| \\ \hat{Z}^k(t) &\triangleq \left| \{k' \in \mathbb{Z}_{\geq 1} : k' \leq k, \hat{\tau}_s^{k'} + S^{k'} > t\} \right| + |\{i \in \mathcal{I} : R_{s,i}(0) > t\}|. \end{aligned}$$

We let $\hat{V}(t)$ be the number of virtual jobs in service at time t .

Now we prove by induction that $\tau_s^k \leq \hat{\tau}_s^k$ for all $k \in \mathbb{Z}_{\geq 1}$, i.e., each job in the original GI/GI/n queue begins service no later than in the modified GI/GI/n queue.

- **Base step:** $k = 1$. Because job 1 is the first job that begins service after time 0, we have

$$\begin{aligned} \tau_s^1 &= \inf\{t > \tau_a^1 : Z^0(t) \leq n - 1\} \\ \hat{\tau}_s^1 &= \inf\{t > \tau_a^1 : \hat{Z}^0(t) + \hat{V}(t) \leq n - 1\}. \end{aligned}$$

By definitions, $Z^0(t) = \hat{Z}^0(t)$ for all $t \geq 0$, and $\hat{V}(t) \geq 0$, so we have $\tau_s^1 \leq \hat{\tau}_s^1$.

- **Induction step:** Suppose we have proved $\tau_s^{k'} \leq \hat{\tau}_s^{k'}$ for $1 \leq k' \leq k$, and we want to show that $\tau_s^{k+1} \leq \hat{\tau}_s^{k+1}$. Observe that

$$\begin{aligned} \tau_s^{k+1} &= \inf\{t > \tau_a^{k+1} : Z^k(t) \leq n - 1\} \\ \hat{\tau}_s^{k+1} &= \inf\{t > \tau_a^{k+1} : \hat{Z}^k(t) + \hat{V}(t) \leq n - 1\}. \end{aligned}$$

Because $\tau_s^{k'} + S^{k'} \leq \hat{\tau}_s^{k'} + S^{k'}$ for all $1 \leq k' \leq k$, we have $Z^k(t) \leq \hat{Z}^k(t)$ for all $t \geq 0$. Moreover, $\hat{V}(t) \geq 0$, so we get $\tau_s^{k+1} \leq \hat{\tau}_s^{k+1}$.

By induction, we have shown that $\tau_s^k \leq \hat{\tau}_s^k$ for all $k \in \mathbb{Z}_{\geq 1}$. Combining the induction results and the formula of $Q(t)$ and $\hat{Q}(t)$ in (7.116) and (7.117), we have finished the proof. $\square$

7.D Proofs of supporting lemmas

We prove the supporting lemmas for homogeneous $GI/GI/n$ (Section 7.6) and heterogeneous $GI/GI/n$ (Section 7.8) together. This section is organized as follows. We first introduce some additional notation and definitions in Appendix 7.D.1. Then we prove BAR (Lemmas 7.12 and 7.28) in Appendix 7.D.2. Next, in Appendix 7.D.3, we prove the lemmas of bounded mean residual times (Lemmas 7.14 and 7.29). In Appendix 7.D.4, we prove the uncorrelation lemmas (Lemmas 7.15 and 7.30). Finally, in Appendix 7.D.5, we prove the preliminary steady-state equalities (Lemmas 7.16 and 7.31) and finiteness results (Lemmas 7.17 and 7.32).

Throughout Appendix 7.D, we focus on the modified $GI/GI/n$, whose service-time distributions are allowed to be heterogeneous across servers. We focus on the case where $\rho < 1$, and use the convention that the system is in steady state at time 0, consistent with Section 7.6.2. Without loss of generality, we assume that no events (arrivals or completions) happen at the same time point (Assumption 7.10); see Remark 7.13 for the necessary changes to proof without this assumption.

7.D.1 Additional notation and definitions

In this subsection, we introduce some additional notation and definitions regarding the modified $GI/GI/n$ queue and the leave-one-out systems. Since we do not consider the original $GI/GI/n$ queue throughout this section, we will reuse some notation in Section 7.3 and Appendix 7.C that was reserved for the original $GI/GI/n$ queue.

Recall that the modified $GI/GI/n$ queue is driven by a set of independent renewal processes corresponding to the arrivals and completions. In the main body, we describe these renewal processes mainly using the residual time processes (i.e., forward recurrence times), $R_a(t)$ and $\{R_{s,i}(t)\}_{i \in [n]}$; here, we define the corresponding ages processes (i.e., backward recurrence times): at each time t ,

- Let $\alpha_a(t)$ be the time elapsed since the last arrival.
- Let $\alpha_j(t)$ be the time elapsed since the last completion of the j -th server, for each $j \in [n]$.

Next, we define the alternative state representation based on the ages: we let $\bar{Y}(t) = (Y(t), \tilde{Q}^s(t))$, where $Y(t) \triangleq (\alpha_1(t), \dots, \alpha_n(t), \alpha_a(t), Q(t))$, and

recall that the vector $\tilde{Q}^s$ represents the queue lengths of the set of leave-one-out systems whose loads are below 1.

We let $\bar{Y}(0)$ follow the stationary distribution. Each sample path of $\{\bar{Y}(t)\}_{t \geq 0}$ is determined by the initial state $\bar{Y}(0)$, and the sequences $\{A^k\}_{k \in \mathbb{Z}_{\geq 0}}$ and $\{S_i^k\}_{k \in \mathbb{Z}_{\geq 0}}$ for $i \in [n]$, defined as follows:

- $A^0 - \alpha_a(0)$ is the time of the first arrival, and A^k is the interarrival time between the k -th and $(k+1)$ -th arrival events for $k \in \mathbb{Z}_{\geq 1}$. Then $A^k \sim \mathbb{P}[A \in \cdot]$ for $k \in \mathbb{Z}_{\geq 1}$, and $A^0 \sim \mathbb{P}[A \in \cdot \mid A \geq \alpha_a(0)]$, where A denotes a generic interarrival time independent of anything else.
- $S_i^0 - \alpha_i(0)$ is the completion time of the initial job on the i -th server, and S_i^k is the time between the k -th and $(k+1)$ -th completion events on the i -th server, for each $i \in [n]$ and $k \in \mathbb{Z}_{\geq 1}$. Then $S_i^k \sim \mathbb{P}[S_i \in \cdot]$ for $k \in \mathbb{Z}_{\geq 1}$, and $S_i^0 \sim \mathbb{P}[S_i \in \cdot \mid S_i \geq \alpha_i(0)]$, where S_i denotes a generic service time of the i -th server and is independent of anything else.

The above-defined random variables and $\bar{Y}(0)$ are “mutually independent” in the following sense: A^0 is independent of $\bar{Y}(0)$ conditioned on $\alpha_a(0)$; S_i^0 is independent of $\bar{Y}(0)$ conditioned on $\alpha_i(0)$; for each $k \in \mathbb{Z}_{\geq 1}$, A^k is independent of everything else, i.e., $\bar{Y}(0)$, $\{S_i^\ell\}_{i \in [n], \ell \in \mathbb{Z}_{\geq 0}}$ and $\{A^\ell\}_{\ell \neq k}$; for each $k \in \mathbb{Z}_{\geq 1}$ and $i \in [n]$, S_i^k is independent of everything else, i.e., $\bar{Y}(0)$, $\{S_j^\ell\}_{(j,\ell) \neq (i,k)}$, and $\{A^\ell\}_{\ell \in \mathbb{Z}_{\geq 0}}$.

Now we can write some basic quantities using the stochastic primitives defined above:

For each $k \in \mathbb{Z}_{\geq 1}$, the k -th arrival time is $\tau_a^k = \sum_{\ell=0}^{k-1} A^\ell - \alpha_a(0)$, and we let $\tau_a^0 = -\alpha_a(0)$. For each time $t \geq 0$, the number of arrivals by time t is given by $N_a(t) \triangleq \min\{k \in \mathbb{Z}_{\geq 0} : \tau_a^k + A^k > t\}$; the age at time t is $\alpha_a(t) = t - \tau_a^{N_a(t)}$; the residual arrival time at time t is $R_a(t) = A^{N_a(t)} - \alpha_a(t)$.

Similarly, on the i -th server, for each $k \in \mathbb{Z}_{\geq 1}$, the k -th completion time is $\tau_i^k = \sum_{\ell=0}^{k-1} S_i^\ell - \alpha_i(0)$, and we let $\tau_i^0 = -\alpha_i(0)$. For each time $t \geq 0$, the number of completions by time t is $N_i(t) = \min\{k \in \mathbb{Z}_{\geq 0} : \tau_i^k + S_i^k > t\}$; the age at time t is $\alpha_i(t) = t - \tau_i^{N_i(t)}$; the residual completion time at time t is $R_{s,i}(t) = S_i^{N_i(t)} - \alpha_i(t)$.

We then define a set of filtrations, corresponding to the arrival process, completion processes, and the whole system:

$$\begin{aligned} \mathcal{F}_{a,t} &= \sigma\left(\alpha_a(0), A^0, A^1, \dots, A^{N_a(t)-1}, \mathbb{1}\{A^{N_a(t)} > t - \tau_a^{N_a(t)}\}\right) \quad \forall t \geq 0 \\ \mathcal{F}_{i,t} &= \sigma\left(\alpha_i(0), S_i^0, S_i^1, \dots, S_i^{N_i(t)-1}, \mathbb{1}\{S_i^{N_i(t)} > t - \tau_i^{N_i(t)}\}\right) \quad \forall i \in [n], t \geq 0 \end{aligned}$$

$$\mathcal{F}_t = \sigma\left(\bar{Y}(0), \mathcal{F}_{a,t}, \cup_{i \in [n]} \mathcal{F}_{i,t}\right) \quad \forall t \geq 0.$$

We denote $\mathcal{F}_{a,\infty} = \cup_{t \geq 0} \mathcal{F}_{a,t}$, $\mathcal{F}_{i,\infty} = \cup_{t \geq 0} \mathcal{F}_{i,t}$ for $i \in [n]$, and $\mathcal{F}_\infty = \cup_{t \geq 0} \mathcal{F}_t$.

It is not hard to see that $N_a(t), \alpha_a(t) \in \mathcal{F}_{a,t}$ for $t \geq 0$, but $R_a(t) \notin \mathcal{F}_{a,t}$. Similarly, $N_i(t), \alpha_i(t) \in \mathcal{F}_{i,t}$ for each $i \in [n]$ and $t \geq 0$, but $R_{s,i}(t) \notin \mathcal{F}_{i,t}$. Furthermore, since $Q(t)$ is determined by the initial state and the arrival and completion events by time t , we have $Q(t) \in \mathcal{F}_t$. Similarly, we have $Q_{-j}(t) \in \sigma(Q_{-j}(0), \mathcal{F}_{a,t}, \cup_{i \neq j} \mathcal{F}_{i,t})$ for $j \in [n]$ and $t \geq 0$.

7.D.2 Proof of BAR

In this subsection, we prove BAR for the heterogeneous case Lemma 7.28, which will imply Lemma 7.12 for the homogeneous case as a special case. The arguments we use are adapted from Section 6 of [54].

Lemma 7.28 (Basic Adjoint Relationship for $GI/GI/n$ with heterogeneous servers). *Consider the modified $GI/GI/n$ system with heterogeneous servers. Suppose Assumptions 7.22, 7.23 and 7.24 hold, and the load $\rho < 1$. For any function $f: \bar{\mathcal{X}} \rightarrow \mathbb{R}$ that is bounded, locally Lipschitz continuous, and partially differentiable from the right with respect to its first $(n+1)$ coordinates, we have*

$$\mathbb{E}_\pi[\mathcal{A}f(\bar{X})] + \Lambda \mathbb{E}_a[f(\bar{X}^+) - f(\bar{X}^-)] + \sum_{i=1}^n \mu_i \mathbb{E}_{s,i}[f(\bar{X}^+) - f(\bar{X}^-)] = 0, \quad (7.81)$$

where $\bar{X}, \bar{X}^+, \bar{X}^- \in \bar{\mathcal{X}}$ are random elements following the distributions specified in the expectation subscripts. Moreover, the Palm distributions $\mathbb{P}_a$ and $\mathbb{P}_{s,i}$ satisfy the following with probability 1:

- (i) If $(\bar{X}^+, \bar{X}^-) \sim \mathbb{P}_a$, then $R_a^- = 0$ and the coordinates of $\bar{X}^+$ and $\bar{X}^-$ satisfy the relations:

$$R_a^+ = R_a^- + A \quad (7.82)$$

$$Q^+ = Q^- + 1 \quad (7.83)$$

$$Q_{-j}^+ = Q_{-j}^- + 1 \quad \forall j \in J^s, \quad (7.84)$$

where A represents a generic interarrival time and is independent of $\bar{X}^-$; $J^s \triangleq \{j \in [n]: \rho_{-j} < 1\}$; other coordinates of $\bar{X}^+$ are identical to $\bar{X}^-$.

(ii) For each $i \in [n]$, if $(\bar{X}^+, \bar{X}^-) \sim \mathbb{P}_{s,i}$, then $R_{s,i}^- = 0$ and the coordinates of $\bar{X}^+$ and $\bar{X}^-$ satisfy the relations:

$$R_{s,i}^+ = R_{s,i}^- + S_i \quad (7.85)$$

$$Q^+ = Q^- - \mathbf{1}\{Q^- > 0\} \quad (7.86)$$

$$Q_{-j}^+ = Q_{-j}^- - \mathbf{1}\{Q_{-j}^- > 0\} \quad \forall j \in J^s \setminus \{i\}, \quad (7.87)$$

where S_i represents a generic service time of the i -th server and is independent of $\bar{X}^-$; other coordinates of $\bar{X}^+$ are identical to $\bar{X}^-$.

Proof. We let $\bar{X}(0)$ follow the stationary distribution π at time 0. Let $\tau_1, \tau_2, \dots$ be the times of all arrivals or completion events after time 0, indexed in temporal order. As assumed throughout our proofs, no events occur simultaneously, and we have sketched what should be changed for the general case in Remark 7.13.

Proving BAR In the first part of the proof, we prove the main BAR equation (7.81) utilizing the fundamental theorem of calculus. We start by examining the differentiability of $f(\bar{X}(t))$ between two consecutive events, τ^k and τ^{k+1} , for each $k \in \mathbb{Z}_{\geq 1}$. Observe that when $t \in (\tau^k, \tau^{k+1})$, $d\bar{X}_\ell(t)/dt = -1$ for $\ell = 1, 2, \dots, n+1$, and the other coordinates of $\bar{X}(t)$ remain constant. This implies that $\bar{X}(t)$ is Lipschitz continuous for $t \in [\tau^k, \tau^{k+1}]$. Given the local Lipschitz continuity of $f(\cdot)$ and the compactness of $[\tau^k, \tau^{k+1}]$, we deduce that $f(\bar{X}(t))$ is Lipschitz continuous for $t \in [\tau^k, \tau^{k+1}]$. Since Lipschitz continuity implies absolute continuity, $f(\bar{X}(t))$ is almost everywhere differentiable on $[\tau^k, \tau^{k+1}]$. Then for any t where $f(\bar{X}(t))$ is differentiable, the derivative of $f(\bar{X}(t))$ agrees with the right derivative:

$$\frac{df(\bar{X}(t))}{dt} = - \sum_{\ell=1}^{n+1} \partial_\ell^+ f(\bar{X}(t)) \triangleq \mathcal{A}f(\bar{X}(t)).$$

Moreover, because the fundamental theorem of calculus holds for absolutely continuous functions, we have

$$\begin{aligned} f(\bar{X}(\tau^{k+1} \wedge t)) - f(\bar{X}(\tau^k)) &= \int_{\tau^k}^{\tau^{k+1} \wedge t} \frac{df(\bar{X}(u))}{du} du \\ &= \int_{\tau^k}^{\tau^{k+1} \wedge t} \mathcal{A}f(\bar{X}(u)) du, \end{aligned} \quad (7.118)$$

for $1 \leq k < N(t) \triangleq \min\{k' : \tau^{k'} > t\}$. Similarly, we can apply the fundamental theorem of calculus to $f(\bar{X}(\tau^1 \wedge t)) - f(\bar{X}(0))$. Summing up all the resulting equations, we get

$$f(\bar{X}(t)) - f(\bar{X}(0)) = \int_0^t \mathcal{A}f(\bar{X}(u))du + \sum_{k=1}^{N(t)-1} (f(\bar{X}(\tau^k)) - f(\bar{X}(\tau^k-))). \quad (7.119)$$

Rearranging the terms based on different types of events, we get

$$\begin{aligned} & f(\bar{X}(t)) - f(\bar{X}(0)) \\ &= \int_0^t \mathcal{A}f(\bar{X}(u))du + \sum_{k=1}^{\infty} (f(\bar{X}(\tau_a^k)) - f(\bar{X}(\tau_a^k-))) \mathbf{1}\{\tau_a^k \leq t\} \\ &+ \sum_{i=1}^n \sum_{k=1}^{\infty} (f(\bar{X}(\tau_i^k)) - f(\bar{X}(\tau_i^k-))) \mathbf{1}\{\tau_i^k \leq t\}. \end{aligned} \quad (7.120)$$

Now we take the expectations on both sides of (7.120) to get

$$\begin{aligned} & \mathbb{E} [f(\bar{X}(t)) - f(\bar{X}(0))] \\ &= \mathbb{E} \left[\int_0^t \mathcal{A}f(\bar{X}(u))du \right] \\ &+ \mathbb{E} \left[\sum_{k=1}^{\infty} (f(\bar{X}(\tau_a^k)) - f(\bar{X}(\tau_a^k-))) \mathbf{1}\{\tau_a^k \leq t\} \right] \\ &+ \sum_{i=1}^n \mathbb{E} \left[\sum_{k=1}^{\infty} (f(\bar{X}(\tau_i^k)) - f(\bar{X}(\tau_i^k-))) \mathbf{1}\{\tau_i^k \leq t\} \right]. \end{aligned} \quad (7.121)$$

Because $\bar{X}(0)$ follows the stationary distribution π , $\bar{X}(0) \stackrel{d}{=} \bar{X}(u) \sim \pi$ for all $u \geq 0$. This implies that

$$\begin{aligned} \mathbb{E} [f(\bar{X}(t)) - f(\bar{X}(0))] &= 0 \\ \mathbb{E} \left[\int_0^t \mathcal{A}f(\bar{X}(u))du \right] &= \int_0^t \mathbb{E} [\mathcal{A}f(\bar{X}(u))] du \\ &= t \mathbb{E}_{\pi} [\mathcal{A}f(\bar{X})], \end{aligned}$$

where the interchange of $\mathbb{E}[\cdot]$ with $\int_0^t \cdot$ is valid because $\mathcal{A}f(\bar{X}(u)) = -\sum_{\ell=1}^{n+1} \partial_{\ell}^+ f(\bar{X}(u))$ is bounded for $u \in [0, t]$, due to the local Lipschitz

continuity of $f(\cdot)$. Therefore, substituting the above two equations back to (7.121) and taking $t = 1$, we get

$$\begin{aligned} 0 = \mathbb{E}_\pi[\mathcal{A}f(\bar{X})] + \mathbb{E} \left[\sum_{k=1}^{\infty} (f(\bar{X}(\tau_a^k)) - f(\bar{X}(\tau_a^k-))) \mathbb{1}\{\tau_a^k \leq 1\} \right] \\ + \sum_{i=1}^n \mathbb{E} \left[\sum_{k=1}^{\infty} (f(\bar{X}(\tau_i^k)) - f(\bar{X}(\tau_i^k-))) \mathbb{1}\{\tau_i^k \leq 1\} \right]. \end{aligned} \quad (7.122)$$

Note that the second and third terms of (7.122) have similar forms as the Palm expectations: as defined in Section 7.6.2, we have for each $i \in [n]$,

$$\begin{aligned} \mathbb{E}_a[f(\bar{X}^+) - f(\bar{X}^-)] &= \frac{1}{\Lambda} \mathbb{E} \left[\sum_{k=1}^{\infty} (f(\bar{X}(\tau_a^k)) - f(\bar{X}(\tau_a^k-))) \mathbb{1}\{\tau_a^k \leq 1\} \right] \\ \mathbb{E}_{s,i}[f(\bar{X}^+) - f(\bar{X}^-)] &= \frac{1}{\mu_i} \mathbb{E} \left[\sum_{k=1}^{\infty} (f(\bar{X}(\tau_i^k)) - f(\bar{X}(\tau_i^k-))) \mathbb{1}\{\tau_i^k \leq 1\} \right]. \end{aligned}$$

Substituting the definitions into (7.122), we get the main equation of BAR:

$$\mathbb{E}_\pi[\mathcal{A}f(\bar{X})] + \Lambda \mathbb{E}_a[f(\bar{X}^+) - f(\bar{X}^-)] + \sum_{i=1}^n \mu_i \mathbb{E}_{s,i}[f(\bar{X}^+) - f(\bar{X}^-)] = 0. \quad (7.81)$$

Characterizing Palm expectations In the remainder of the proof, we prove Lemma 7.28 (i) and Lemma 7.28 (ii), which characterize the joint distribution of $\bar{X}^+$ and $\bar{X}^-$ under the Palm expectations $\mathbb{P}_a$ and $\mathbb{P}_{s,i}$ for $i \in [n]$.

We start by proving Lemma 7.28 (i). First, because $R_a(\tau_a^k-) = 0$ for all $k \in \mathbb{Z}_{\geq 1}$, by the definition of $\mathbb{P}_a$, we have

$$\mathbb{P}_a[R_a = 0] = \frac{1}{\Lambda} \mathbb{E} \left[\sum_{k=1}^{\infty} \mathbb{1}\{R_a(\tau_a^k-) = 0, \tau_a^k \leq 1\} \right] = 1.$$

Next, to understand the relations between the coordinates of $\bar{X}^+$ and $\bar{X}^-$ under $\mathbb{P}_a$, we examine the expectation $\mathbb{E}_a[f(\bar{X}^-)g(\bar{X}^+ - \bar{X}^-)]$ for any bounded measurable functions $f, g: \mathbb{R}^{n+1} \times \mathbb{Z}^{1+|J^s|} \rightarrow \mathbb{R}$:

$$\mathbb{E}_a[f(\bar{X}^-)g(\bar{X}^+ - \bar{X}^-)]$$

$$\begin{aligned}
&= \frac{1}{\Lambda} \mathbb{E} \left[\sum_{k=1}^{\infty} f(\bar{X}(\tau_a^k -)) g(\bar{X}(\tau_a^k) - \bar{X}(\tau_a^k -)) \mathbf{1}\{\tau_a^k \leq 1\} \right] \\
&= \frac{1}{\Lambda} \sum_{k=1}^{\infty} \mathbb{E} [f(\bar{X}(\tau_a^k -)) g(\bar{X}(\tau_a^k) - \bar{X}(\tau_a^k -)) \mathbf{1}\{\tau_a^k \leq 1\}].
\end{aligned}$$

By the definitions of the sample paths, for each $k \in \mathbb{Z}_{\geq 1}$, at the time of arrival τ_a^k ,

$$\bar{X}(\tau_a^k) = \bar{X}(\tau_a^k -) + A^k \mathbf{e}_a + \mathbf{e}_q + \sum_{j \in J^s} \mathbf{e}_{q,-j}, \quad (7.123)$$

where $\mathbf{e}_a, \mathbf{e}_q, \mathbf{e}_{q,-j} \in \bar{\mathcal{X}}$ are one-hot vectors with the non-zero entries at the coordinates of R_a , Q , and Q_{-j} , respectively. Note that A^k is independent of $(\tau_a^k, \bar{X}(\tau_a^k -))$ and has the same distribution as A , so

$$\begin{aligned}
&\mathbb{E}_a [f(\bar{X}^-) g(\bar{X}^+ - \bar{X}^-)] \\
&= \frac{1}{\Lambda} \sum_{k=1}^{\infty} \mathbb{E} [f(\bar{X}(\tau_a^k -)) \mathbf{1}\{\tau_a^k \leq 1\}] \mathbb{E} \left[g \left(A \mathbf{e}_a + \mathbf{e}_q + \sum_{j \in J^s} \mathbf{e}_{q,-j} \right) \right] \\
&= \mathbb{E}_a [f(\bar{X}^-)] \mathbb{E} \left[g \left(A \mathbf{e}_a + \mathbf{e}_q + \sum_{j \in J^s} \mathbf{e}_{q,-j} \right) \right]. \quad (7.124)
\end{aligned}$$

Because f and g can be arbitrary bounded measurable functions, (7.124) implies that under $\mathbb{P}_a$, $\bar{X}^+ - \bar{X}^-$ is independent of $\bar{X}^-$, and has the same distribution as $A \mathbf{e}_a + \mathbf{e}_q + \sum_{j \in J^s} \mathbf{e}_{q,-j}$. This proves Lemma 7.28 (i).

To show Lemma 7.28 (ii), note that for each $i \in [n]$ and $k \in \mathbb{Z}_{\geq 1}$, we have $R_{s,i}(\tau_i^k -) = 0$, which implies that $R_{s,i}^- = 0$ with probability 1 under $\mathbb{P}_{s,i}$. Moreover, by definitions,

$$\bar{X}(\tau_i^k) = \bar{X}(\tau_i^k -) + S_i^k \mathbf{e}_i - \mathbf{1}\{Q(\tau_i^k -) > 0\} \mathbf{e}_q - \sum_{j \in J^s \setminus \{i\}} \mathbf{1}\{Q_{-j}(\tau_i^k -) > 0\} \mathbf{e}_{q,-j}, \quad (7.125)$$

where $\mathbf{e}_i \in \bar{\mathcal{X}}$ denotes the one-hot vector with the non-zero entry at the coordinate of $R_{s,i}$. Because S_i^k is independent of $(\tau_i^k, \bar{X}(\tau_i^k -))$ and has the same distribution as S_i , for any bounded measurable functions $f: \mathbb{R}^{n+1} \times \mathbb{Z}^{1+|J^s|} \rightarrow \mathbb{R}$ and $g: \mathbb{R} \rightarrow \mathbb{R}$, we have

$$\mathbb{E}_{s,i} [f(\bar{X}^-) g(R_{s,i}^+ - R_{s,i}^-)]$$

$$\begin{aligned}
&= \frac{1}{\Lambda} \sum_{k=1}^{\infty} \mathbb{E} [f(\bar{X}(\tau_i^k -)) g(R_{s,i}(\tau_i^k) - R_{s,i}(\tau_i^k -)) \mathbf{1}\{\tau_i^k \leq 1\}] \\
&= \frac{1}{\Lambda} \sum_{k=1}^{\infty} \mathbb{E} [f(\bar{X}(\tau_i^k -)) \mathbf{1}\{\tau_i^k \leq 1\}] \mathbb{E}[g(S_i)] \\
&= \mathbb{E}_{s,i} [f(\bar{X}^-)] \mathbb{E}[g(S_i)].
\end{aligned} \tag{7.126}$$

This implies that under $\mathbb{P}_{s,i}$, $R_{s,i}^+ - R_{s,i}^-$ is independent of $\bar{X}^-$ and has the same distribution as S_i . In addition, for any bounded measurable function $f: \mathbb{R}^{n+1} \times \mathbb{Z}^{1+|J^s|} \times \mathbb{Z} \rightarrow \mathbb{R}$,

$$\begin{aligned}
&\mathbb{E}_{s,i} [f(\bar{X}^-, Q^+)] \\
&= \frac{1}{\Lambda} \sum_{k=1}^{\infty} \mathbb{E} [f(\bar{X}(\tau_i^k -), Q(\tau_i^k)) \mathbf{1}\{\tau_i^k \leq 1\}] \\
&= \frac{1}{\Lambda} \sum_{k=1}^{\infty} \mathbb{E} \left[f\left(\bar{X}(\tau_i^k -), Q(\tau_i^k -) - \mathbf{1}\{Q(\tau_i^k -) > 0\}\right) \mathbf{1}\{\tau_i^k \leq 1\} \right] \\
&= \mathbb{E}_{s,i} [f(\bar{X}^-, Q^- - \mathbf{1}\{Q^- > 0\})],
\end{aligned}$$

which implies that $Q^+ = Q^- - \mathbf{1}\{Q^- > 0\}$ under the distribution $\mathbb{P}_{s,i}$ with probability 1. Similarly, we can show that $Q_{-j}^+ = Q_{-j}^- - \mathbf{1}\{Q_{-j}^- > 0\}$ for $j \in J^s \setminus \{i\}$, $R_a^+ = R_a^-$, and $R_{s,j}^+ = R_{s,j}^-$ for $j \in [n] \setminus \{i\}$ under $\mathbb{P}_{s,i}$ with probability 1. This proves Lemma 7.28 (ii). $\square$

7.D.3 Proof of bounded mean residual times (Lemmas 7.14 and 7.29)

In this section, we prove Lemma 7.14 and Lemma 7.29, restated as follows:

Lemma 7.14 (Bounded mean residual times). *Consider the modified GI/GI/n queue satisfying Assumptions 7.1 and 7.3, with the load $\rho < 1 - 1/n$. For any $j \in [n]$,*

$$\mu \mathbb{E}_{\nu} [R_{s,j} \mid Q, \tilde{Q}] \leq R_s^{\max} \tag{7.48}$$

$$\Lambda \mathbb{E}_{\nu} [R_a \mid Q, \tilde{Q}] \geq R_a^{\min} \tag{7.49}$$

where $\mathbb{E}_{\nu}[\cdot]$ can denote either $\mathbb{E}_{\pi}[\cdot]$ or $\mathbb{E}_{s,i}[\cdot]$ for $i \in [n] \setminus \{j\}$.

Lemma 7.29 (Bounded mean residual times, heterogeneous). *Consider the modified GI/GI/n queue with heterogeneous servers. Suppose Assumptions 7.22 and 7.24 hold and the load $\rho < 1$. Then for any $j \in [n]$,*

$$\mu_j \mathbb{E}_\nu [R_{s,j} \mid Q, \tilde{Q}^s] \leq R_{s,j}^{\max} \quad (7.48)$$

$$\Lambda \mathbb{E}_\nu [R_a \mid Q, \tilde{Q}^s] \geq R_a^{\min}. \quad (7.49)$$

where $\mathbb{E}_\nu[\cdot]$ can denote either $\mathbb{E}_\pi[\cdot]$ or $\mathbb{E}_{s,i}[\cdot]$ for $i \neq j$.

Below, we focus on proving Lemma 7.29 since it is a strict generalization of Lemma 7.14.

Proof of Lemma 7.29. We first show (7.48) for $\nu = \pi$. By the construction of the sample paths in Appendix 7.D.1 and the tower property of conditional expectations,

$$\begin{aligned} \mathbb{E}_\pi [R_{s,j} \mid Q, \tilde{Q}^s] &= \mathbb{E} [R_{s,j}(0) \mid Q(0), \tilde{Q}^s(0)] \\ &= \mathbb{E} \left[\mathbb{E} [R_{s,j}(0) \mid Q(0), \tilde{Q}^s(0), \alpha_j(0)] \mid Q(0), \tilde{Q}^s(0) \right] \\ &= \mathbb{E} \left[\mathbb{E} [S_j^0 - \alpha_j(0) \mid Q(0), \tilde{Q}^s(0), \alpha_j(0)] \mid Q(0), \tilde{Q}^s(0) \right]. \end{aligned} \quad (7.127)$$

The inner conditional expectation in the last expression can be bounded as follows:

$$\mathbb{E} [S_j^0 - \alpha_j(0) \mid Q(0), \tilde{Q}^s(0), \alpha_j(0)] = \mathbb{E} [S_j^0 - \alpha_j(0) \mid \alpha_j(0)] \quad (7.128)$$

$$= \mathbb{E} [S_j - \alpha_j(0) \mid S_j \geq \alpha_j(0)] \quad (7.129)$$

$$\leq R_{s,j}^{\max} / \mu_j. \quad (7.130)$$

where (7.128) is because S_j^0 is by definition independent of $Q(0)$ and $\tilde{Q}^s(0)$ given $\alpha_j(0)$; (7.129) is by the distribution of S_j^0 ; (7.130) is due to the definition $R_{s,j}^{\max} \triangleq \sup_{t' \geq 0} \mathbb{E} [\mu_j S_j - t' \mid \mu_j S_j \geq t']$. Substituting the bound in (7.130) back to (7.127) yields $\mu_j \mathbb{E}_\pi [R_{s,j} \mid Q, \tilde{Q}^s] \leq R_{s,j}^{\max}$.

Next, we prove (7.48) for $\nu = \mathbb{P}_{s,i}$ with $i \neq j$. Let $(q, \tilde{q}^s)$ be any realization of $(Q, \tilde{Q}^s)$. By the definition of $\mathbb{E}_{s,i}[\cdot]$, we have

$$\mathbb{E}_{s,i} [R_{s,j} \mathbb{1} \{ (Q, \tilde{Q}^s) = (q, \tilde{q}^s) \}]$$

$$\begin{aligned}
&= \frac{1}{\mu_i} \mathbb{E} \left[\sum_{k=1}^{\infty} \mathbb{1} \left\{ \tau_i^k \leq 1, (Q(\tau_i^k -), \tilde{Q}^s(\tau_i^k -)) = (q, \tilde{q}^s) \right\} R_{s,j}(\tau_i^k -) \right] \\
&= \frac{1}{\mu_i} \sum_{k=1}^{\infty} \mathbb{E} \left[\mathbb{E} \left[\mathbb{1} \left\{ \tau_i^k \leq 1, (Q(\tau_i^k -), \tilde{Q}^s(\tau_i^k -)) = (q, \tilde{q}^s) \right\} \right. \right. \\
&\quad \left. \left. \cdot R_{s,j}(\tau_i^k -) \mid \overline{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell \neq j}, \mathcal{F}_{j,\tau_i^k} \right] \right] \\
&= \frac{1}{\mu_i} \sum_{k=1}^{\infty} \mathbb{E} \left[\mathbb{1} \left\{ \tau_i^k \leq 1 \right\} \mathbb{1} \left\{ (Q(\tau_i^k -), \tilde{Q}^s(\tau_i^k -)) = (q, \tilde{q}^s) \right\} \right. \\
&\quad \left. \cdot \mathbb{E} \left[R_{s,j}(\tau_i^k -) \mid \overline{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell \neq j}, \mathcal{F}_{j,\tau_i^k} \right] \right], \tag{7.131}
\end{aligned}$$

where (7.131) is because τ_i^k , $Q(\tau_i^k -)$, and $\tilde{Q}^s(\tau_i^k -)$ are measurable with respect to the sigma algebra

$$\sigma(\overline{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell \neq j}, \mathcal{F}_{j,\tau_i^k}).$$

To bound the conditional expectation of $R_{s,j}(\tau_i^k -)$ in (7.131), we rewrite it as:

$$\begin{aligned}
&\mathbb{E} \left[R_{s,j}(\tau_i^k -) \mid \overline{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell \neq j}, \mathcal{F}_{j,\tau_i^k} \right] \\
&= \mathbb{E} \left[S_j^{N_j(\tau_i^k -)} - \alpha_j(\tau_i^k -) \mid \overline{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell \neq j}, \mathcal{F}_{j,\tau_i^k} \right] \\
&= \sum_{m=0}^{\infty} \mathbb{E} \left[\left(S_j^m - \alpha_j(\tau_i^k -) \right) \mathbb{1} \{ N_j(\tau_i^k -) = m \} \right. \\
&\quad \left. \mid \overline{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell \neq j}, \mathcal{F}_{j,\tau_i^k} \right]. \tag{7.132}
\end{aligned}$$

Because τ_i^k is measurable with respect to $\sigma(\overline{Y}(0), \mathcal{F}_{i,\infty})$, we can rewrite (7.132) as $\sum_{m=0}^{\infty} U_m(\tau_i^k -)$, where for each $t \geq 0$, $U_m(t)$ is the random variable define as

$$U_m(t) \triangleq \mathbb{E} \left[\left(S_j^m - \alpha_j(t) \right) \mathbb{1} \{ N_j(t) = m \} \mid \overline{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell \neq j}, \mathcal{F}_{j,t} \right].$$

For each $t \geq 0$ and $m \in \mathbb{Z}_{\geq 0}$, $U_m(t)$ can be bounded as

$$U_m(t) = \mathbb{E} \left[\left(S_j^m - \alpha_j(t) \right) \mathbb{1} \{ N_j(t) = m \} \mid \mathcal{F}_{j,t} \right] \tag{7.133}$$

$$= \mathbb{E} \left[\left(S_j^m - (t - \tau_j^m) \right) \mathbb{1} \{ \tau_j^m \leq t < \tau_j^m + S_j^m \} \right] \quad (7.134)$$

$$\begin{aligned} & \left| \alpha_j(0), \{S_j^\ell\}_{\ell=0}^{m-1}, \mathbb{1} \{ S_j^m > (t - \tau_j^m) \} \right] \\ &= (\mathbb{E} [S_j^m \mid \alpha_j(0), \{S_j^\ell\}_{\ell=0}^{m-1}, \mathbb{1} \{ S_j^m > (t - \tau_j^m) \}]) - (t - \tau_j^m) \\ & \quad \cdot \mathbb{1} \{ \tau_j^m \leq t < \tau_j^m + S_j^m \} \end{aligned}$$

$$= (\mathbb{E} [S_j \mid S_j > t - \tau_j^m] - (t - \tau_j^m)) \cdot \mathbb{1} \{ \tau_j^m \leq t < \tau_j^m + S_j^m \} \quad (7.135)$$

$$\leq (R_{s,j}^{\max} / \mu_j) \cdot \mathbb{1} \{ \tau_j^m \leq t < \tau_j^m + S_j^m \}. \quad (7.136)$$

Here, (7.133) follows from the conditional independence of $(S_j^m, \alpha_j(t), N_j(t))$ and $\sigma(\bar{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell \neq j})$ given $\alpha_j(0)$, and the fact that $\alpha_j(0) \in \mathcal{F}_{j,t}$. Furthermore, (7.134) is due to the definitions of $N_j(t)$ and $\mathcal{F}_{j,t}$; (7.135) is by the definition of S_j^m and its independence with $(\alpha_j(0), \{S_j^\ell\}_{\ell=0}^{m-1}, \tau_j^m)$ if $m \geq 1$; (7.136) follows from the definition $R_{s,j}^{\max} \triangleq \sup_{t' \geq 0} \mathbb{E}[\mu_j S_j - t' \mid \mu_j S_j \geq t']$. Substituting (7.136) back to (7.132) implies that for $k \in \mathbb{Z}_{\geq 1}$,

$$\begin{aligned} & \mathbb{E} \left[R_{s,j}(\tau_i^k -) \mid \bar{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell \neq j}, \mathcal{F}_{j,\tau_i^k} \right] \\ & \leq \sum_{m=0}^{\infty} U_m(\tau_i^k -) \\ & \leq (R_{s,j}^{\max} / \mu_j) \cdot \sum_{m=0}^{\infty} \mathbb{1} \{ \tau_j^m < \tau_i^k \leq \tau_j^m + S_j^m \} \\ & = (R_{s,j}^{\max} / \mu_j) \cdot \mathbb{1} \{ \tau_j^0 < \tau_i^k < \infty \} \\ & \leq R_{s,j}^{\max} / \mu_j \quad a.s. \end{aligned}$$

Further substituting the above inequality back to (7.131) yields

$$\begin{aligned} & \mathbb{E}_{s,i} \left[R_{s,j} \mathbb{1} \{ (Q, \tilde{Q}^s) = (q, \tilde{q}^s) \} \right] \\ & \leq (R_{s,j}^{\max} / \mu_j) \cdot \frac{1}{\mu_i} \sum_{k=1}^{\infty} \mathbb{E} \left[\mathbb{1} \{ \tau_i^k \leq 1 \} \mathbb{1} \{ (Q(\tau_i^k -), \tilde{Q}^s(\tau_i^k -)) = (q, \tilde{q}^s) \} \right] \\ & = (R_{s,j}^{\max} / \mu_j) \cdot \mathbb{P}_{s,i} \left[(Q, \tilde{Q}^s) = (q, \tilde{q}^s) \right], \end{aligned}$$

where $(q, \tilde{q}^s)$ is an arbitrary realization of $(Q, \tilde{Q}^s)$. This proves

$$\mathbb{E}_{s,i} \left[R_{s,j} \mid Q, \tilde{Q}^s \right] \leq R_{s,j}^{\max} / \mu_j.$$

The proof for (7.49) mirrors the proof for (7.48) with the following two modifications:

1. All quantities associated with the j -th server's completion process are swapped with the corresponding quantities associated with the arrival process.
2. The definition of $R_a^{\min}$ is used in place of $R_{s,j}^{\max}$.

Given the similarity, the detailed proof for (7.49) is omitted. $\square$

7.D.4 Proof of the uncorrelation lemmas (Lemmas 7.15 and 7.30)

In this subsection, we prove the uncorrelation lemmas, Lemma 7.15 and Lemma 7.30, restated as follows:

Lemma 7.15 (Uncorrelation lemma). *Consider the modified GI/GI/ n queue satisfying Assumptions 7.1, 7.2, and 7.3, with the load $\rho < 1 - 1/n$. For any $i, j \in [n]$ such that $i \neq j$, we have*

$$\mathbb{E}_{s,i}[\mathbb{1}\{Q_{-j} = 0\}R_{s,j}] = \mathbb{P}_{s,i}[Q_{-j} = 0] \mathbb{E}_{\pi}[R_{s,j}]. \quad (7.50)$$

Lemma 7.30 (Uncorrelation lemma, heterogeneous). *Consider the modified GI/GI/ n queue with heterogeneous servers. Suppose Assumptions 7.22, 7.23, and 7.24 hold. Then for any $i, j \in [n]$ such that $i \neq j$ and $\rho_{-j} < 1$,*

$$\mathbb{E}_{s,i}[\mathbb{1}\{Q_{-j} = 0\}R_{s,j}] = \mathbb{P}_{s,i}[Q_{-j} = 0] \mathbb{E}_{\pi}[R_{s,j}]. \quad (7.89)$$

We focus on the uncorrelation lemma for the heterogeneous setting (Lemma 7.30), since it is a strict generalization of Lemma 7.15. The proof of Lemma 7.30 is based on the following lemma:

Lemma 7.36. *Consider the modified GI/GI/ n queue with heterogeneous servers. For any $j \in [n]$, consider the residual service time process of the j -th server, $\{R_{s,j}(t)\}_{t \geq 0}$, and assume that it has non-lattice service times. Then for any $\epsilon > 0$, there exists a deterministic $T_0 > 0$ such that for any $t \geq T_0 + R_{s,j}(0)$, we have*

$$|\mathbb{E}[R_{s,j}(t) \mid R_{s,j}(0)] - \mathbb{E}_{\pi}[R_{s,j}]| \leq \epsilon \quad w.p. \ 1.$$

Proof of Lemma 7.36. We first consider the case when $R_{s,j}(0) = 0$, so $R_{s,j}(t)$ is a pure renewal process. Because the inter-event time S_j of this renewal

process is non-lattice (Assumption 7.23), we can invoke the Key Renewal Theorem [16, Theorem V.4.3]. Restated in our setting, this theorem implies that

$$\lim_{t \rightarrow \infty} \mathbb{E}[R_{s,j}(t) \mid R_{s,j}(0) = 0] = \mathbb{E}_\pi[R_{s,j}],$$

if we can show that $z(t) \triangleq \mathbb{E}[\max(S_j - t, 0)]$ is *directly Riemann integrable* and satisfies

$$\mathbb{E}[R_{s,j}(t) \mid R_{s,j}(0) = 0] = z(t) + \int_0^t \mathbb{E}[R_{s,j}(t-u) \mid R_{s,j}(0) = 0] F_j(du), \quad (7.137)$$

where $\int \cdot F_j(du)$ denotes the integral with respect to u when u follows the distribution of S_j .

We first show that $z(t)$ is directly Riemann integrable. By Proposition V.4.1 of [16], $z(t)$ is directly Riemann integrable if it is non-increasing and Lebesgue integrable on $[0, \infty)$. Obviously, $z(t)$ is non-increasing. Moreover, $z(t)$ is non-negative and its Lebesgue integral can be bounded as

$$\begin{aligned} \int_0^\infty z(t) dt &= \int_0^\infty \mathbb{E}[\max(S_j - t, 0)] dt \\ &= \mathbb{E} \left[\int_0^\infty \max(S_j - t, 0) dt \right] \\ &= \frac{1}{2} \mathbb{E}[S_j^2] \\ &\leq \frac{1}{2} (R_{s,j}^{\max})^2, \end{aligned}$$

which is finite by Assumption 7.22 (or Assumption 7.1 in the homogeneous case).

Next, we show that $z(t)$ satisfies (7.137). Recall that S_j^1 is the service time of the job that starts service at time $R_{s,j}(0)$. We decompose $\mathbb{E}[R_{s,j}(t) \mid R_{s,j}(0) = 0]$ based on whether $S_j^1 > t$ to get the following:

$$\begin{aligned} &\mathbb{E}[R_{s,j}(t) \mid R_{s,j}(0) = 0] \\ &= \mathbb{E}[R_{s,j}(t) \mathbb{1}\{S_j^1 > t\} \mid R_{s,j}(0) = 0] \\ &\quad + \mathbb{E}[R_{s,j}(t) \mathbb{1}\{S_j^1 \leq t\} \mid R_{s,j}(0) = 0] \\ &= \mathbb{E}[\max(S_j^1 - t, 0)] + \int_0^t \mathbb{E}[R_{s,j}(t) \mid R_{s,j}(0) = 0, S_j^1 = u] F_j(du) \end{aligned}$$

$$= z(t) + \int_0^t \mathbb{E}[R_{s,j}(t-u) \mid R_{s,j}(0) = 0] F_j(du).$$

The above arguments verify the assumptions of the Key Renewal Theorem. Therefore, given $R_{s,j}(0) = 0$, for any $\epsilon > 0$, there exists $T_0 > 0$ such that for any $t \geq T_0$,

$$|\mathbb{E}[R_{s,j}(t) \mid R_{s,j}(0) = 0] - \mathbb{E}_\pi[R_{s,j}]| \leq \epsilon. \quad (7.138)$$

In general, when $R_{s,j}(0) \neq 0$, $R_{s,j}(t)$ is a delayed renewal process and we have

$$\mathbb{E}[R_{s,j}(t + R_{s,j}(0)) \mid R_{s,j}(0)] = \mathbb{E}[R_{s,j}(t) \mid R_{s,j}(0) = 0] \quad \forall t > 0.$$

Substituting the last equation into (7.138) finishes the proof. $\square$

Now we are ready to prove Lemma 7.30.

Proof of Lemma 7.30. By the definition of the Palm expectation $\mathbb{E}_{s,i}[\cdot]$, for arbitrary $T > 0$, we have

$$\begin{aligned} & \mathbb{E}_{s,i}[\mathbb{1}\{Q_{-j} = 0\} R_{s,j}] \\ &= \frac{1}{\mu_i T} \mathbb{E} \left[\sum_{k=1}^{\infty} \mathbb{1}\{\tau_i^k \leq T, Q_{-j}(\tau_i^k -) = 0\} R_{s,j}(\tau_i^k -) \right] \\ &= \frac{1}{\mu_i T} \sum_{k=1}^{\infty} \mathbb{E} \left[\mathbb{E} \left[\mathbb{1}\{\tau_i^k \leq T, Q_{-j}(\tau_i^k -) = 0\} R_{s,j}(\tau_i^k -) \right. \right. \\ & \quad \left. \left. \mid \bar{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell \neq j} \right] \right] \\ &= \frac{1}{\mu_i T} \sum_{k=1}^{\infty} \mathbb{E} \left[\mathbb{1}\{\tau_i^k \leq T, Q_{-j}(\tau_i^k -) = 0\} \right. \\ & \quad \left. \cdot \mathbb{E} \left[R_{s,j}(\tau_i^k -) \mid \bar{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell \neq j} \right] \right] \end{aligned} \quad (7.139)$$

$$= \frac{1}{\mu_i T} \sum_{k=1}^{\infty} \mathbb{E} \left[\mathbb{1}\{\tau_i^k \leq T, Q_{-j}(\tau_i^k -) = 0\} \cdot \bar{R}(\tau_i^k -) \right], \quad (7.140)$$

where (7.139) is because $(\tau_i^k, Q_{-j}(\tau_i^k -)) \in \sigma(\bar{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell \neq j})$; $\bar{R}(\cdot)$ is defined as

$$\bar{R}(t) \triangleq \mathbb{E}[R_{s,j}(t) \mid \bar{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell \neq j}] \quad \forall t \geq 0,$$

and (7.140) is due to the fact that $\tau_i^k \in \mathcal{F}_{i,\infty}$. Note that because the process $R_{s,j}(t)$ is conditionally independent of $\bar{Y}(0)$ given $\alpha_j(0)$, we can simplify $\bar{R}(t)$ into

$$\bar{R}(t) = \mathbb{E}[R_{s,j}(t) \mid \alpha_j(0)]. \quad (7.141)$$

In the subsequent argument, we will use the fact that

$$\bar{R}(t) = \mathbb{E}[\mathbb{E}[R_{s,j}(t) \mid R_{s,j}(0)] \mid \alpha_j(0)], \quad (7.142)$$

which is true because $R_{s,j}(t) = R_{s,j}(0) + \sum_{\ell=1}^{N_j(t)} S_j^\ell - t$ and $\alpha_j(0) = S_j^0 - R_{s,j}(0)$ are conditionally independent given $R_{s,j}(0)$.

Next, we bound the difference between $\bar{R}(\tau_i^k -)$ and $\mathbb{E}_\pi[R_{s,j}]$ for each $k \in \mathbb{Z}_{\geq 1}$. Recall that by Lemma 7.36, for any $\epsilon > 0$, there exists a deterministic $T_0 > 0$ such that for any $t \geq T_0 + R_{s,j}(0)$, we have

$$|\mathbb{E}[R_{s,j}(t) \mid R_{s,j}(0)] - \mathbb{E}_\pi[R_{s,j}]| \leq \epsilon \quad \text{w.p. 1.} \quad (7.143)$$

Let the random variable $K^\epsilon \triangleq \min\{k' \in \mathbb{Z}_{\geq 1} : \tau_i^{k'+1} > T_0 + R_{s,j}(0)\}$. Then by (7.142) and (7.143), we have

$$|\bar{R}(\tau_i^k -) - \mathbb{E}_\pi[R_{s,j}]| \leq \epsilon \quad \text{if } k \geq K^\epsilon + 1. \quad (7.144)$$

Moreover, we argue that

$$\mathbb{E}[K^\epsilon] \leq \mu_i T_0 + \mu_i \mathbb{E}_\pi[R_{s,j}] + \mathbb{E}[(\mu_i S_i)^2] < \infty. \quad (7.145)$$

To see this, recall that $\tau_i^{k'+1} = \sum_{\ell=0}^{k'} S_i^\ell - \alpha_i(0) = \sum_{\ell=1}^{k'} S_i^\ell + R_{s,i}(0)$, where S_i^ℓ are i.i.d. for $\ell \geq 1$. We can thus rewrite $K^\epsilon = \inf\{k' \in \mathbb{Z}_{\geq 1} : \sum_{\ell=1}^{k'} S_i^\ell > T_0 + R_{s,j}(0) - R_{s,i}(0)\}$, which shows that K^ϵ is a stopping time with respect to the discrete-time random walk $(\sum_{\ell=1}^{k'} S_i^\ell)_{k'=1}^\infty$. By Wald's equation and Lorden's inequality [191], we have

$$\begin{aligned} \mathbb{E}[K^\epsilon \mid R_{s,j}(0), R_{s,i}(0)] \mathbb{E}[S_i] &= \mathbb{E}\left[\sum_{\ell=1}^{K^\epsilon} S_i^\ell \mid R_{s,j}(0), R_{s,i}(0)\right] \\ &\leq \max(T_0 + R_{s,j}(0) - R_{s,i}(0), 0) + \mu_i \mathbb{E}[S_i^2], \end{aligned}$$

which implies (7.145) after rearranging the terms.

Now we can finish the proof utilizing (7.140), (7.144), and (7.145). First, by the definition of $\mathbb{P}_{s,i}[Q_{-j} = 0]$, we have that for any $T > 0$,

$$\mathbb{P}_{s,i}[Q_{-j} = 0] = \frac{1}{\mu_i T} \mathbb{E}\left[\sum_{k=1}^{\infty} \mathbf{1}\{\tau_i^k \leq T, Q_{-j}(\tau_i^k -) = 0\}\right]$$

Combining the last equality with the equality in (7.140), we get

$$\begin{aligned}
& |\mathbb{E}_{s,i}[\mathbb{1}\{Q_{-j} = 0\}R_{s,j}] - \mathbb{P}_{s,i}[Q_{-j} = 0] \mathbb{E}_\pi[R_{s,j}]| \\
&= \frac{1}{\mu_i T} \left| \mathbb{E} \left[\sum_{k=1}^{\infty} \mathbb{1}\{\tau_i^k \leq T, Q_{-j}(\tau_i^k -) = 0\} \cdot \bar{R}(\tau_i^k -) \right] \right. \\
&\quad \left. - \mathbb{E} \left[\sum_{k=1}^{\infty} \mathbb{1}\{\tau_i^k \leq T, Q_{-j}(\tau_i^k -) = 0\} \right] \mathbb{E}_\pi[R_{s,j}] \right| \\
&= \frac{1}{\mu_i T} \left| \mathbb{E} \left[\sum_{k=1}^{\infty} \mathbb{1}\{\tau_i^k \leq T, Q_{-j}(\tau_i^k -) = 0\} (\bar{R}(\tau_i^k -) - \mathbb{E}_\pi[R_{s,j}]) \right] \right| \\
&\leq \frac{1}{\mu_i T} \mathbb{E} \left[\sum_{k=1}^{\infty} \mathbb{1}\{\tau_i^k \leq T\} \cdot |\bar{R}(\tau_i^k -) - \mathbb{E}_\pi[R_{s,j}]| \right].
\end{aligned}$$

Fixing any $\epsilon > 0$, we break the infinite sum above into two parts based on K^ϵ :

$$\begin{aligned}
& |\mathbb{E}_{s,i}[\mathbb{1}\{Q_{-j} = 0\}R_{s,j}] - \mathbb{P}_{s,i}[Q_{-j} = 0] \mathbb{E}_\pi[R_{s,j}]| \\
&\leq \frac{1}{\mu_i T} \mathbb{E} \left[\sum_{k=1}^{\infty} \mathbb{1}\{\tau_i^k \leq T\} \cdot |\bar{R}(\tau_i^k -) - \mathbb{E}_\pi[R_{s,j}]| \right] \\
&= \frac{1}{\mu_i T} \mathbb{E} \left[\sum_{k=1}^{K^\epsilon} \mathbb{1}\{\tau_i^k \leq T\} |\bar{R}(\tau_i^k -) - \mathbb{E}_\pi[R_{s,j}]| \right] \\
&\quad + \frac{1}{\mu_i T} \mathbb{E} \left[\sum_{k=K^\epsilon+1}^{\infty} \mathbb{1}\{\tau_i^k \leq T\} |\bar{R}(\tau_i^k -) - \mathbb{E}_\pi[R_{s,j}]| \right] \tag{7.146}
\end{aligned}$$

$$\leq \frac{2R_{s,j}^{\max} \mathbb{E}[K^\epsilon]}{\mu_i T} + \frac{\epsilon}{\mu_i T} \mathbb{E} \left[\sum_{k=K^\epsilon+1}^{\infty} \mathbb{1}\{\tau_i^k \leq T\} \right] \tag{7.147}$$

$$\leq \frac{2R_{s,j}^{\max} \mathbb{E}[K^\epsilon]}{\mu_i T} + \epsilon. \tag{7.148}$$

Here, to get (7.147), we bound the expectation in (7.146) using the argument that

$$|\bar{R}(\tau_i^k -) - \mathbb{E}_\pi[R_{s,j}]| \leq \bar{R}(\tau_i^k -) + \mathbb{E}_\pi[R_{s,j}] \leq 2R_{s,j}^{\max} \quad a.s.$$

To get (7.148), we use the fact that $\mathbb{E}[\sum_{k=1}^{\infty} \mathbb{1}\{\tau_i^k \leq T\}] = \mu_i T$. Now, because $\mathbb{E}[K^\epsilon] < \infty$, and T and ϵ are arbitrary, first taking $T \rightarrow \infty$ and then taking $\epsilon \rightarrow 0$ in (7.148) finish the proof. $\square$

7.D.5 Proofs of some steady-state equalities and finiteness results

In this subsection, we will prove several basic equalities about the residual times of the completion processes and the idle probability of the queue (Lemmas 7.16 and 7.31). We will also prove that modified $GI/GI/n$ queues have finite mean queue lengths under the Palm distributions $\mathbb{P}_a$ and $\mathbb{P}_{s,i}$ (Lemmas 7.17 and 7.32). All the results in this subsection are derived from BAR (Lemma 7.28), which we restate below for ease of reference:

Lemma 7.28 (Basic Adjoint Relationship for $GI/GI/n$ with heterogeneous servers). *Consider the modified $GI/GI/n$ system with heterogeneous servers. Suppose Assumptions 7.22, 7.23 and 7.24 hold, and the load $\rho < 1$. For any function $f: \bar{\mathcal{X}} \rightarrow \mathbb{R}$ that is bounded, locally Lipschitz continuous, and partially differentiable from the right with respect to its first $(n+1)$ coordinates, we have*

$$\mathbb{E}_\pi[\mathcal{A}f(\bar{X})] + \Lambda \mathbb{E}_a[f(\bar{X}^+) - f(\bar{X}^-)] + \sum_{i=1}^n \mu_i \mathbb{E}_{s,i}[f(\bar{X}^+) - f(\bar{X}^-)] = 0, \quad (7.81)$$

where $\bar{X}, \bar{X}^+, \bar{X}^- \in \bar{\mathcal{X}}$ are random elements following the distributions specified in the expectation subscripts. Moreover, the Palm distributions $\mathbb{P}_a$ and $\mathbb{P}_{s,i}$ satisfy the following with probability 1:

(i) If $(\bar{X}^+, \bar{X}^-) \sim \mathbb{P}_a$, then $R_a^- = 0$ and the coordinates of $\bar{X}^+$ and $\bar{X}^-$ satisfy the relations:

$$R_a^+ = R_a^- + A \quad (7.82)$$

$$Q^+ = Q^- + 1 \quad (7.83)$$

$$Q_{-j}^+ = Q_{-j}^- + 1 \quad \forall j \in J^s, \quad (7.84)$$

where A represents a generic interarrival time and is independent of $\bar{X}^-$; $J^s \triangleq \{j \in [n]: \rho_{-j} < 1\}$; other coordinates of $\bar{X}^+$ are identical to $\bar{X}^-$.

(ii) For each $i \in [n]$, if $(\bar{X}^+, \bar{X}^-) \sim \mathbb{P}_{s,i}$, then $R_{s,i}^- = 0$ and the coordinates of $\bar{X}^+$ and $\bar{X}^-$ satisfy the relations:

$$R_{s,i}^+ = R_{s,i}^- + S_i \quad (7.85)$$

$$Q^+ = Q^- - \mathbb{1}\{Q^- > 0\} \quad (7.86)$$

$$Q_{-j}^+ = Q_{-j}^- - \mathbb{1}\{Q_{-j}^- > 0\} \quad \forall j \in J^s \setminus \{i\}, \quad (7.87)$$

where S_i represents a generic service time of the i -th server and is independent of $\bar{X}^-$; other coordinates of $\bar{X}^+$ are identical to $\bar{X}^-$.

Now, we restate Lemma 7.16 and its generalization in the heterogeneous setting, Lemma 7.31. We will directly prove Lemma 7.31.

Lemma 7.16. *Consider the modified GI/GI/ n queue satisfying Assumptions 7.1 and 7.3. If $\rho < 1$, for $i, j \in [n]$ such that $i \neq j$, we have*

$$\mathbb{E}_\pi[R_{s,i}] = \frac{1}{2}\mu\mathbb{E}[S^2] \quad (7.51)$$

$$\mathbb{E}_\pi[R_a] = \frac{1}{2}\Lambda\mathbb{E}[A^2] \quad (7.52)$$

$$\mathbb{E}_{s,i}[R_{s,j}] + \mathbb{E}_{s,j}[R_{s,i}] = \mu\mathbb{E}[S^2] \quad (7.53)$$

$$\mathbb{E}_{s,i}[R_a] + \mathbb{E}_a[R_{s,i}] = \frac{1}{2}\mu\mathbb{E}[S^2] + \frac{1}{2}\Lambda\mathbb{E}[A^2] \quad (7.54)$$

$$\mathbb{P}_s[Q = 0] = 1 - \rho \quad (7.55)$$

In addition, if $\rho < 1 - 1/n$, for each $j \in [n]$, the j -th leave-one-out system satisfies

$$\frac{1}{n} \sum_{i: i \neq j} \mathbb{P}_{s,i}[Q_{-j} = 0] = 1 - \rho - \frac{1}{n}. \quad (7.56)$$

Lemma 7.31. *Consider the modified GI/GI/ n queue with heterogeneous servers. Suppose Assumptions 7.22 and 7.24 hold, and the load $\rho < 1$. For any $i, j \in [n]$ such that $i \neq j$, we have*

$$\mathbb{E}_\pi[R_{s,i}] = \frac{1}{2}\mu_i\mathbb{E}[S_i^2] \quad (7.90)$$

$$\mathbb{E}_\pi[R_a] = \frac{1}{2}\Lambda\mathbb{E}[A^2] \quad (7.91)$$

$$\mathbb{E}_{s,i}[R_{s,j}] + \mathbb{E}_{s,j}[R_{s,i}] = \frac{1}{2}\mu_i\mathbb{E}[S_i^2] + \frac{1}{2}\mu_j\mathbb{E}[S_j^2] \quad (7.92)$$

$$\mathbb{E}_{s,i}[R_a] + \mathbb{E}_a[R_{s,i}] = \frac{1}{2}\mu_i\mathbb{E}[S_i^2] + \frac{1}{2}\Lambda\mathbb{E}[A^2] \quad (7.93)$$

$$\mathbb{P}_s[Q = 0] = 1 - \rho \quad (7.94)$$

In addition, for each $j \in [n]$, if $\rho_{-j} < 1$, the j -th leave-one-out system satisfies

$$\frac{1}{\mu_\Sigma} \sum_{i: i \neq j} \mu_i \mathbb{P}_{s,i}[Q_{-j} = 0] = 1 - \rho - \frac{\mu_j}{\mu_\Sigma}. \quad (7.95)$$

Proof of Lemma 7.31. Proof of (7.90). For any truncation threshold $\kappa > 0$, consider the test function $f(\bar{X}) = (R_{s,i} \wedge \kappa)^2$, which is bounded, locally Lipschitz continuous, and has right partial derivatives. Consequently, Basic Adjoint Relationship (7.81) implies that

$$-2\mathbb{E}_\pi[R_{s,i}\mathbb{1}\{R_{s,i} < \kappa\}] + \Lambda\mathbb{E}_{s,i}\left[\left((R_{s,i} + S_i) \wedge \kappa\right)^2 - (R_{s,i} \wedge \kappa)^2\right] = 0.$$

Because $R_{s,i} = 0$ with probability 1 under the measure $\mathbb{P}_{s,i}$, rearranging the terms, we get

$$\mathbb{E}_\pi[R_{s,i}\mathbb{1}\{R_{s,i} < \kappa\}] = \frac{1}{2}\Lambda\mathbb{E}\left[(S_i \wedge \kappa)^2\right].$$

Observe that $R_{s,i}\mathbb{1}\{R_{s,i} < \kappa\}$ and $(S_i \wedge \kappa)^2$ are both non-negative and non-decreasing in κ , by the monotone convergence theorem,

$$\mathbb{E}_\pi[R_{s,i}] = \lim_{\kappa \rightarrow \infty} \mathbb{E}_\pi[R_{s,i}\mathbb{1}\{R_{s,i} < \kappa\}] = \lim_{\kappa \rightarrow \infty} \frac{1}{2}\Lambda\mathbb{E}\left[(S_i \wedge \kappa)^2\right] = \frac{1}{2}\Lambda\mathbb{E}\left[S_i^2\right], \quad (7.149)$$

which implies (7.90).

Proof of (7.91). This proof is identical to the proof of (7.90) except that $R_{s,i}$ is replaced by R_a .

Proof of (7.92). For any $\kappa > 0$, we take the test function $f(\bar{X}) = (R_{s,i} \wedge \kappa)(R_{s,j} \wedge \kappa)$ in (7.81):

$$\begin{aligned} & -\mathbb{E}_\pi[\mathbb{1}\{R_{s,i} < \kappa\}(R_{s,j} \wedge \kappa) + \mathbb{1}\{R_{s,j} < \kappa\}(R_{s,i} \wedge \kappa)] \\ & + \mu_i\mathbb{E}_{s,i}[(S_i \wedge \kappa)(R_{s,j} \wedge \kappa) - 0] + \mu_j\mathbb{E}_{s,j}[(S_j \wedge \kappa)(R_{s,i} \wedge \kappa) - 0] = 0, \end{aligned}$$

where in the second and third expectations we have used the facts that $R_{s,\ell} = 0$ with probability 1 under $\mathbb{P}_{s,\ell}$ for any $\ell \in [n]$. Because S_i is independent of $R_{s,j}$ under the measures $\mathbb{P}_{s,i}$, and S_j is independent of $R_{s,i}$ under the measure $\mathbb{P}_{s,j}$, we get

$$\begin{aligned} & \mathbb{E}_\pi[\mathbb{1}\{R_{s,i} < \kappa\}(R_{s,j} \wedge \kappa) + \mathbb{1}\{R_{s,j} < \kappa\}(R_{s,i} \wedge \kappa)] \\ & = \mu_i\mathbb{E}[S_i \wedge \kappa]\mathbb{E}_{s,i}[R_{s,j} \wedge \kappa] + \mu_j\mathbb{E}[S_j \wedge \kappa]\mathbb{E}_{s,j}[R_{s,i} \wedge \kappa]. \end{aligned}$$

Because $\mathbb{1}\{R_{s,i} < \kappa\}(R_{s,j} \wedge \kappa)$, $\mathbb{1}\{R_{s,j} < \kappa\}(R_{s,i} \wedge \kappa)$, $S_i \wedge \kappa$, $S_j \wedge \kappa$, $R_{s,j} \wedge \kappa$, and $R_{s,i} \wedge \kappa$ are all non-negative and non-decreasing in κ , by the monotone convergence theorem, we have the following after letting $\kappa \rightarrow \infty$:

$$\mathbb{E}_\pi[R_{s,j} + R_{s,i}] = \mu_i\mathbb{E}[S_i]\mathbb{E}_{s,i}[R_{s,j}] + \mu_j\mathbb{E}[S_j]\mathbb{E}_{s,j}[R_{s,i}].$$

Note that $\mathbb{E}[S_i] = 1/\mu_i$, $\mathbb{E}[S_j] = 1/\mu_j$, so the last equality simplifies into

$$\mathbb{E}_\pi[R_{s,j} + R_{s,i}] = \mathbb{E}_{s,i}[R_{s,j}] + \mathbb{E}_{s,j}[R_{s,i}]. \quad (7.150)$$

By (7.90), the left-hand side of (7.150) equals $\mu_i \mathbb{E}[S_i^2]/2 + \mu_j \mathbb{E}[S_j^2]/2$. This proves (7.92).

Proof of (7.93). For any $\kappa > 0$, we take the test function $f(\bar{X}) = (R_{s,i} \wedge \kappa)(R_a \wedge \kappa)$ in (7.81). The rest of the proof is identical to the proof of (7.92) except that $R_{s,j}$ is replaced by R_a .

Proof of (7.94). For any integer $\kappa > 0$, taking the test function $f(\bar{X}) = Q \wedge \kappa$ in (7.81) yields

$$\begin{aligned} \Lambda \mathbb{E}_a[(Q + 1) \wedge \kappa - Q \wedge \kappa] + \mu_\Sigma \mathbb{E}_s[(Q - 1)^+ \wedge \kappa - Q \wedge \kappa] &= 0, \\ \implies \Lambda \mathbb{E}_a[\mathbb{1}\{Q \leq \kappa - 1\}] - \mu_\Sigma \mathbb{E}_s[\mathbb{1}\{0 < Q \leq \kappa\}] &= 0. \end{aligned}$$

Because each term in the expectations is bounded, letting $\kappa \rightarrow \infty$ and invoking the bounded convergence theorem, we get

$$\Lambda - \mu_\Sigma \mathbb{P}_s[Q > 0] = 0. \quad (7.151)$$

Rearranging the terms yields (7.55).

Proof of (7.95). This proof is identical to the proof of (7.94) except that Q is replaced by Q_{-j} . $\square$

Next, we restate Lemma 7.17 and its generalization in the heterogeneous setting Lemma 7.32. We will directly prove Lemma 7.32.

Lemma 7.17. *Consider the modified GI/GI/n system satisfying Assumptions 7.1 and 7.3, with the load $\rho < 1$. Then we have $\mathbb{E}_a[Q] < \infty$ and $\mathbb{E}_{s,i}[Q] < \infty$ for all $i \in [n]$.*

Lemma 7.32. *Consider the modified GI/GI/n system with heterogeneous servers. Suppose Assumptions 7.22 and 7.24 hold, and the load $\rho < 1$. Then we have $\mathbb{E}_a[Q] < \infty$ and $\mathbb{E}_{s,i}[Q] < \infty$ for all $i \in [n]$.*

Proof of Lemma 7.32. For any positive integer κ , we consider the test function $f(\bar{X}) = (Q \wedge \kappa - 2\Lambda R_a \wedge \kappa)^2$, which is bounded, locally Lipschitz continuous, and has right partial derivatives. Consequently, Basic Adjoint Relationship (7.81) implies that

$$0 = \mathbb{E}_\pi[2(Q \wedge \kappa - 2\Lambda R_a \wedge \kappa) \cdot (2\Lambda \mathbb{1}\{R_a \leq \kappa\})] \quad (7.152)$$

$$+ \Lambda \mathbb{E}_a \left[((Q+1) \wedge \kappa - 2\Lambda A \wedge \kappa)^2 - (Q \wedge \kappa)^2 \right] \quad (7.153)$$

$$+ \mu_\Sigma \mathbb{E}_s \left[((Q - \mathbf{1}\{Q > 0\}) \wedge \kappa - 2\Lambda R_a \wedge \kappa)^2 - (Q \wedge \kappa - 2\Lambda R_a \wedge \kappa)^2 \right], \quad (7.154)$$

where to get the term in (7.153), we have used the fact that $R_a = 0$ with probability 1 under $\mathbb{P}_a$.

We first simplify the term in (7.153): substituting $(Q+1) \wedge \kappa = Q \wedge \kappa + \mathbf{1}\{Q \leq \kappa - 1\}$ and rearranging the terms, we get

$$\begin{aligned} & ((Q+1) \wedge \kappa - 2\Lambda A \wedge \kappa)^2 - (Q \wedge \kappa)^2 \\ &= 2(Q \wedge \kappa) \cdot (\mathbf{1}\{Q \leq \kappa - 1\} - 2\Lambda A \wedge \kappa) \\ & \quad + (\mathbf{1}\{Q \leq \kappa - 1\} - 2\Lambda A \wedge \kappa)^2. \end{aligned}$$

Because A is independent of Q under the measure $\mathbb{P}_a$, we have

$$\begin{aligned} & \mathbb{E}_a \left[((Q+1) \wedge \kappa - 2\Lambda A \wedge \kappa)^2 - (Q \wedge \kappa)^2 \right] \\ &= 2\mathbb{E}_a[(Q \wedge \kappa) \mathbf{1}\{Q \leq \kappa - 1\}] - 4\mathbb{E}_a[Q \wedge \kappa] \mathbb{E}[\Lambda A \wedge \kappa] \\ & \quad + \mathbb{E}_a[(\mathbf{1}\{Q \leq \kappa - 1\} - 2\Lambda A \wedge \kappa)^2]. \end{aligned}$$

We then simplify the term in (7.154). Substituting $(Q - \mathbf{1}\{Q > 0\}) \wedge \kappa = Q \wedge \kappa - \mathbf{1}\{0 < Q \leq \kappa\}$ and rearranging the terms, we get

$$\begin{aligned} & \mathbb{E}_s \left[((Q - \mathbf{1}\{Q > 0\}) \wedge \kappa - 2\Lambda R_a \wedge \kappa)^2 - (Q \wedge \kappa - 2\Lambda R_a \wedge \kappa)^2 \right] \\ &= \mathbb{E}_s[-2(Q \wedge \kappa - 2\Lambda R_a \wedge \kappa) \mathbf{1}\{0 < Q \leq \kappa\} + \mathbf{1}\{0 < Q \leq \kappa\}] \\ &= -2\mathbb{E}_s[(Q \wedge \kappa) \mathbf{1}\{0 < Q \leq \kappa\}] + 4\mathbb{E}_s[(\Lambda R_a \wedge \kappa) \mathbf{1}\{0 < Q \leq \kappa\}] \\ & \quad + \mathbb{P}_s[0 < Q \leq \kappa]. \end{aligned}$$

Substituting the above calculations back into (7.153) and (7.154), we get

$$\begin{aligned} 0 &= \mathbb{E}_\pi[2(Q \wedge \kappa - 2\Lambda R_a \wedge \kappa) \cdot (2\Lambda \mathbf{1}\{R_a \leq \kappa\})] \\ & \quad + 2\Lambda \mathbb{E}_a[(Q \wedge \kappa) \mathbf{1}\{Q \leq \kappa - 1\}] \\ & \quad - 4\Lambda \mathbb{E}_a[Q \wedge \kappa] \mathbb{E}[\Lambda A \wedge \kappa] + \Lambda \mathbb{E}_a[(\mathbf{1}\{Q \leq \kappa - 1\} - 2\Lambda A \wedge \kappa)^2] \\ & \quad - 2\mu_\Sigma \mathbb{E}_s[(Q \wedge \kappa) \mathbf{1}\{0 < Q \leq \kappa\}] + 4\mu_\Sigma \mathbb{E}_s[(\Lambda R_a \wedge \kappa) \mathbf{1}\{0 < Q \leq \kappa\}] \\ & \quad + \mu_\Sigma \mathbb{P}_s[0 < Q \leq \kappa]. \end{aligned}$$

Moving the terms involving $\mathbb{E}_a[Q \wedge \kappa]$ and $\mathbb{E}_s[Q \wedge \kappa]$ to the left-hand side, we get

$$\begin{aligned}
& -2\Lambda\mathbb{E}_a[(Q \wedge \kappa)\mathbb{1}\{Q \leq \kappa - 1\}] + 4\Lambda\mathbb{E}_a[Q \wedge \kappa]\mathbb{E}[\Lambda A \wedge \kappa] \\
& + 2\mu_\Sigma\mathbb{E}_s[(Q \wedge \kappa)\mathbb{1}\{0 < Q \leq \kappa\}] \\
& = \mathbb{E}_\pi[2(Q \wedge \kappa - 2\Lambda R_a \wedge \kappa) \cdot (2\Lambda\mathbb{1}\{R_a \leq \kappa\})] \\
& + \Lambda\mathbb{E}_a[(\mathbb{1}\{Q \leq \kappa - 1\} - 2\Lambda A \wedge \kappa)^2] \\
& + 4\mu_\Sigma\mathbb{E}_s[(\Lambda R_a \wedge \kappa)\mathbb{1}\{0 < Q \leq \kappa\}] + \mu_\Sigma\mathbb{P}_s[0 < Q \leq \kappa] \\
& \leq 4\Lambda\mathbb{E}_\pi[Q] + \Lambda\mathbb{E}[(2\Lambda A)^2] + \Lambda + 4\mu_\Sigma\mathbb{E}_s[\Lambda R_a] + \mu_\Sigma.
\end{aligned} \tag{7.155}$$

Note that each term in the last expression is finite. In particular, $\mathbb{E}_\pi[Q] < \infty$ due to Assumption 7.24, and $\mathbb{E}_s[\Lambda R_a] < \infty$ due to the following argument:

$$\begin{aligned}
\mathbb{E}_s[\Lambda R_a] &= \frac{1}{\mu_\Sigma} \sum_{i=1}^n \mu_i \mathbb{E}_{s,i}[\Lambda R_a] \\
&\leq \frac{\Lambda}{\mu_\Sigma} \sum_{i=1}^n \mu_i (\mathbb{E}_{s,i}[R_a] + \mathbb{E}_a[R_{s,i}]) \\
&= \frac{\Lambda}{\mu_\Sigma} \sum_{i=1}^n \frac{\mu_i}{2} \left(\mu_i \mathbb{E}[S_i^2] + \Lambda \mathbb{E}[A^2] \right) \\
&< \infty,
\end{aligned}$$

where the last equality follows from (7.93) of Lemma 7.31. Rearranging the terms in (7.155) and using the fact that $\mathbb{1}\{Q \leq \kappa - 1\} \leq 1$, we get

$$\begin{aligned}
& \Lambda\mathbb{E}_a[Q \wedge \kappa] \left(4\mathbb{E}[\Lambda A \wedge \kappa] - 2 \right) + 2\mu_\Sigma\mathbb{E}_s[(Q \wedge \kappa)\mathbb{1}\{0 < Q \leq \kappa\}] \\
& \leq 4\Lambda\mathbb{E}_\pi[Q] + \Lambda\mathbb{E}[(2\Lambda A)^2] + \Lambda + 4\mu_\Sigma\mathbb{E}_s[\Lambda R_a] + \mu_\Sigma \\
& < \infty.
\end{aligned} \tag{7.156}$$

Now we try to take $\kappa \rightarrow \infty$ in (7.156). Because $\lim_{\kappa \rightarrow \infty} \mathbb{E}[\Lambda A \wedge \kappa] = \Lambda\mathbb{E}[A] = 1$, there exists κ_0 such that for all $\kappa > \kappa_0$, we have $\mathbb{E}[\Lambda A \wedge \kappa] > 1/2$ and $4\mathbb{E}[\Lambda A \wedge \kappa] - 2 > 0$. Because $Q \wedge \kappa$ and $(Q \wedge \kappa)\mathbb{1}\{0 < Q \leq \kappa\}$ are both non-negative and non-decreasing in κ , by the monotone convergence theorem and (7.156), we obtain

$$2\Lambda\mathbb{E}_a[Q] + 2\mu_\Sigma\mathbb{E}_s[Q]$$

$$\begin{aligned}
&= \lim_{\kappa \rightarrow \infty} \left(\Lambda \mathbb{E}_a[Q \wedge \kappa] \left(4\mathbb{E}[\Lambda A \wedge \kappa] - 2 \right) \right. \\
&\quad \left. + 2\mu_\Sigma \mathbb{E}_s[(Q \wedge \kappa) \mathbf{1}\{0 < Q \leq \kappa\}] \right) \\
&\leq 4\Lambda \mathbb{E}_\pi[Q] + \Lambda \mathbb{E}[(2\Lambda A)^2] + \Lambda + 4\mu_\Sigma \mathbb{E}_s[\Lambda R_a] + \mu_\Sigma \quad (7.157) \\
&< \infty. \quad (7.158)
\end{aligned}$$

Recall that $\mathbb{E}_s[\cdot] = \sum_{i=1}^n \mu_i \mathbb{E}_{s,i}[\cdot] / \mu_\Sigma$, so (7.157) implies that $\mathbb{E}_a[Q] < \infty$ and $\mathbb{E}_{s,i}[Q] < \infty$ for all $i \in [n]$. This completes the proof. $\square$

Chapter 8

Performance of the Gittins Policy in the G/G/1 and G/G/k, With and Without Setup Times

8.1 Introduction

We consider the classic problem of preemptively scheduling jobs in a queue to minimize mean number-in-system, or equivalently mean response time (a.k.a. sojourn time). Even in single-server queueing models, this can be a nontrivial problem whose answer depends on the information available to the scheduler. The simplest case is when the scheduler knows each job's size (a.k.a. service time), for which the optimal policy is Shortest Remaining Processing Time (SRPT) [241]: always serve the job of least remaining work.

In the more realistic case of scheduling with unknown or partially known job sizes, the optimal policy is only known for the M/G/1. It is called the *Gittins* policy (a.k.a. Gittins index policy) [1, 2, 114, 243]. Based on whatever service time information is available for each job, Gittins assigns each job a scalar *rank* (i.e. priority), then serves the job of least rank. For example, SRPT is the special case of Gittins where job sizes are known exactly, and a job's rank is its remaining work. More generally, a job's rank is, roughly speaking, an estimate of its remaining work based on whatever information is available.

The Gittins policy is known to be optimal in the M/G/1 [114, 243]. But plenty of systems and models have more complex features, including:

- (a) *Multiple servers*, such as the M/G/k.
- (b) *Non-Poisson arrival processes*, such as the G/G/1 (more specifically, the GI/GI/1).
- (c) *Periods of server unavailability*, such as models with setup times.

Either (a) or (b) alone makes optimal scheduling intractable. Combining all three, as in the G/G/k with setup times (G/G/k/setup), only adds to the challenge.

With optimality out of reach, we are left to find a tractable near-optimal policy. We thus ask:

How well does Gittins perform in systems with features (a), (b), and (c) like the G/G/k/setup?

Gittins is a natural candidate because its definition naturally generalizes beyond the M/G/1, even if its optimality proof does not [114]. For instance, in a G/G/k, Gittins simply serves the k jobs of k least ranks, or all jobs if there are fewer than k .

Only feature (a) has been addressed in full generality in prior work [129, 242, 245]. Specifically, it is known that in the M/G/k, the additive suboptimality gap of Gittins (abbreviated in equations to “Gtn”) is bounded by [242]¹

$$\mathbb{E}[N]_{M/G/k}^{\text{Gtn}} - \inf_{\text{policies } \pi} \mathbb{E}[N]_{M/G/k}^{\pi} \leq C(k-1) \log \frac{1}{1-\rho}. \quad (8.1)$$

Let us briefly explain the notation used in (8.1):

- $\mathbb{E}[N]_{M/G/k}^{\pi}$ is the mean number-in-system under policy π in M/G/k.
- k is the number of servers.
- $\rho \in [0, 1)$ is the *load* (a.k.a. utilization), namely the average fraction of servers that are busy.
- $C \approx 3.775$ is a constant.

A notable feature of (8.1) is that under mild conditions [245], the right-hand side is dominated by $\inf_{\pi} \mathbb{E}[N]_{M/G/k}^{\pi}$, the performance of the optimal policy, in the *heavy-traffic limit*, meaning as $\rho \rightarrow 1$. That is, as the M/G/k gets busier and busier, the difference between Gittins’s performance and that of

¹Throughout this chapter, $\log$ is the natural logarithm.

the optimal policy becomes negligible. Gittins is thus considered *heavy-traffic optimal* in the $M/G/k$.

The above progress on analyzing Gittins in the multiserver $M/G/k$ is certainly promising for handling (a). But, as we explain in more detail in Section 8.1.2, key steps of the prior $M/G/k$ analysis rely on Poisson arrivals and uninterrupted server availability, so they cannot handle (b) and (c).

8.1.1 Results: Performance Bounds and Heavy-Traffic Optimality

We give the first analysis of the Gittins policy for systems with any combination of (a) multiple servers, (b) G/G arrivals, and (c) setup times. We frame our results in terms of the $G/G/k$ with setup times ($G/G/k/setup$), which can have all three features. But they also apply to systems with a subset of the features, such as the $G/G/1$ or $M/G/k/setup$, because we can view them as special cases of the $G/G/k/setup$.

Our main results, presented in full in Section 8.4, can be summarized as

$$\mathbb{E}[N]_{G/G/k/setup}^{\text{Gtn}} - \inf_{\text{policies } \pi} \mathbb{E}[N]_{G/G/k/setup}^{\pi} \leq \ell_{(a)} + \ell_{(b)} + \ell_{(a)\&(c)}, \quad (8.2)$$

where each term on the right-hand side is a “suboptimality loss” that is non-zero when the model has the features in the subscript. For example, the $M/G/k/setup$ has only features (a) and (c), so $\ell_{(b)} = 0$. A particularly notable case is the $M/G/1/setup$, which has only feature (c), so all three loss terms in (8.2) are zero. Indeed, we show that Gittins is optimal among non-idling policies in the $M/G/1/setup$, a previously unknown result.

Our result generalizes prior work on Gittins in the $M/G/k$ in the sense that $\ell_{(a)}$ turns out to be the right-hand side of (8.1). Remarkably, the other loss terms, $\ell_{(b)}$ and $\ell_{(a)\&(c)}$, are uniformly bounded at all loads. This implies that, under mild conditions, Gittins is heavy-traffic optimal in the $G/G/k/setup$.

Stability of the $G/G/k/setup$ under Complex Scheduling Policies

Perhaps a more basic question than optimizing mean response time is stability: under which scheduling policies is the $G/G/k/setup$ stable? This is an open question which is outside the scope of our work. As such, our results hold under an assumption on the stability region of the $G/G/k/setup$ (Assumption 8.1). We believe the assumption always holds, giving a partial proof sketch in Appendix 8.D.

Beyond the $G/G/k$ /setup

The techniques underlying our results are very general, applying even beyond the $G/G/k$ /setup. In Section 8.10, we sketch how our results could be extended to other systems.

- Building on the theme of multiple servers, we consider systems with *multiserver jobs*, which must be simultaneously served by multiple servers [55]. Due to the prevalence of multiserver jobs in cloud computing, these models have received lots of recent attention [128, 129, 138, 148, 238, 239, 271].
- Building on the theme of non-Poisson arrivals, we consider *batch arrivals* of jobs [61].
- Building on the theme of setup times, we consider *generalized vacations*, which model a variety of scenarios where servers are temporarily unavailable [95, 96].

8.1.2 Main Obstacles and Key Ideas

While there is a substantial literature on scheduling in the $M/G/1$ [137, Part VII], much less is known as soon as we introduce features (a), (b), and (c). Any two of these, let alone all three, yields a system where optimal scheduling has never been studied. This is perhaps unsurprising in light of the fact analyzing these systems under First-Come First-Served (FCFS) is already very difficult. See Li and Goldberg [180] and references therein for a review of the $G/G/k$, and similarly for Williams et al. [286] for the $M/G/k$ /setup. Even in the $G/G/1$, we only know the optimal scheduling policy for known job sizes, when it is SRPT [241].

Fortunately, recent advances analyzing Gittins in the $M/G/k$ [129, 242, 245] give us hope in the form of a new avenue for analyzing performance. Scully et al. [245] introduce a new queueing identity, now known as *WINE* [242] (Section 8.6),² which relates the number of jobs N in the system to, roughly speaking, the amount of *work* in the system. This is helpful because bounding the amount of work in an $M/G/k$, which WINE turns into a bound on $\mathbb{E}[N]$, turns out to be significantly easier than directly bounding $\mathbb{E}[N]$.

²Scully [242, Section 2.2.3] notes that WINE builds upon several similar identities that precede it [27, 116, 117, 236].

WINE holds in any queueing system, including the $G/G/k/\text{setup}$, so we can and do use the same overall strategy of bounding work, then using WINE to turn the work bound into an $\mathbb{E}[N]$ bound. However, there are significant obstacles to carrying out this strategy in the $G/G/k/\text{setup}$.

Non-Poisson Arrivals

The first step of our strategy is to analyze the amount of work in the system. The approach that prior work takes to analyze the $M/G/k$ is to use a *work decomposition law*. This is a result which, in its most general form, relates the amount of work in a generic system with M/G arrivals to the amount of work in a “resource-pooled” $M/G/1$ experiencing the same arrivals. The prior $M/G/k$ bound in (8.1) comes from the fact that the $M/G/k$ and resource-pooled $M/G/1$ turn out to have similar amounts of work. We would like to take a similar approach with the $G/G/k/\text{setup}$. Unfortunately, the combination of G/G arrivals and multiple servers rules out using existing work decomposition laws (Section 8.2.3).

To overcome this, we prove a *new work decomposition law for G/G arrivals* (Section 8.7). We view this as the main technical contribution that makes our results possible. Indeed, by combining WINE and our new work decomposition law, the $\ell_{(a)}$ and $\ell_{(b)}$ terms of (8.2) follows relatively easily. But the $\ell_{(c)}$ term and heavy-traffic analysis present additional obstacles, as discussed below.

Setup Times

One of the key observations behind the prior $M/G/k$ analysis is that whenever there are k jobs in the system, all servers are occupied. This implies that in terms of work, an $M/G/k$ never falls too far behind an $M/G/1$, where the $M/G/1$ experiences the same arrivals and has the same total service capacity. But in an $M/G/k/\text{setup}$ or $G/G/k/\text{setup}$, there is no analogous limit to how far behind an $M/G/1/\text{setup}$ or $G/G/k/\text{setup}$ system can be, because there is no limit on the number of jobs that might arrive during a setup time.

To overcome this, we perform a novel analysis of setup times to bound the number of arrivals during one setup time *in expectation*. This analysis is the basis of the $\ell_{(a)\&(c)}$ term of (8.2).

Heavy Traffic Analysis

The above ideas are enough to prove the bound in (8.2). But the question remains: is the right-hand side of (8.2) small or large relative to $\inf_{\pi} \mathbb{E}[N]_{G/G/k/setup}^{\pi}$, the performance of the optimal policy? If the latter dominates the former in the $\rho \rightarrow 1$ limit, then Gittins is optimal in heavy traffic. The right-hand side grows as $O(\log \frac{1}{1-\rho})$, so the main challenge is to give a lower bound on the performance of the optimal policy. In prior work on the $M/G/k$, one can use SRPT in a resource-pooled $M/G/1$ as a lower bound on the optimal policy, which is helpful because the SRPT has been analyzed in heavy traffic [183]. We would like to use the same approach with the $G/G/1$ as the lower bound, but SRPT has never been analyzed in the heavy-traffic $G/G/1$.

To overcome this, we give the *first heavy-traffic analysis of SRPT in the $G/G/1$* . This provides a lower bound on $\inf_{\pi} \mathbb{E}[N]_{G/G/k/setup}^{\pi}$, which turns out to be enough for our purposes. The key ingredient of our heavy-traffic analysis is, once again, our new work decomposition law for G/G arrivals, underscoring its importance as our key technical contribution.

8.1.3 Outline

The rest of the chapter is organized as follows:

- Section 8.2 reviews related work.
- Section 8.3 describes our $G/G/k/setup$ model, and in particular details of the setup times.
- Section 8.4 presents our main results on Gittins: suboptimality gap bounds (Theorems 8.7 and 8.8) and heavy-traffic optimality (Theorem 8.9).
- Section 8.5 gives a high-level overview of how we prove our main results.
- Section 8.6 reviews necessary background on Gittins and WINE.
- Section 8.7 proves a *new work decomposition law for systems with G/G arrivals*. This is the key technical contribution that underlies all of our other results in this chapter.
- Section 8.8 proves the suboptimality gap bounds (Theorems 8.7 and 8.8).

- Section 8.9 proves heavy-traffic optimality (Theorem 8.9). The key step involves giving *first heavy-traffic analysis of SRPT in the G/G/1*, a result of independent interest.

8.2 Related Work

8.2.1 Optimal Scheduling in Queues

Gittins in Single-Server Systems

The Gittins policy was originally conceived to solve the Markovian multi-armed bandit problem [113, 114], but it was soon adapted to also solve the problem of scheduling in an M/G/1 to minimize mean number of jobs and similar metrics. See Scully and Harchol-Balter [243] and the references therein for a review of Gittins in the M/G/1. However, aside from some particular cases [235, 241], the degree to which Gittins performs well in the G/G/1 or G/G/1/setup was previously unknown.

The “SOAP” technique of Scully et al. [244] can be used to analyze the performance of the Gittins policy in the M/G/1. However, while SOAP is convenient for analyzing any fixed size distribution (e.g. numerically), using it to prove theorems that hold for all size distributions is cumbersome [246, Section 1.1]. Moreover, SOAP is limited to the M/G/1 and, thanks to an extension by van Vreumingen [263], the M/G/1/setup. Analyzing Gittins with G/G arrivals or multiple servers seems to be beyond SOAP [248, Appendix B].

Gittins in Multi-Server Systems

Gittins is known to be suboptimal with multiple servers [114], but researchers have studied the extent to which the suboptimality gap is large or small. The earliest results of this type analyzed an M/M/ k with Bernoulli feedback [117] and *nonpreemptive* M/G/ k with Bernoulli feedback [116]. These results proved (in the latter case, under an additional assumption) constant suboptimality gaps for Gittins in these systems. But both models are somewhat restrictive, excluding, for instance, heavy-tailed job size distributions that are common in computer systems [67, 136, 139, 221, 223]. More recent work, which we discussed in Section 8.1, overcomes these limitations to bound the performance of Gittins in the M/G/ k for general job sizes, including heavy-tailed sizes

[129, 242, 247]. However, all of the above work assumes M/G arrivals with no server unavailability.

8.2.2 Setup Times

Multiserver Models

A significant line of previous work has studied the $M/M/k$ /setup with exponential setup times and FCFS scheduling [14, 100–103, 222]. Among those works, Gandhi and Harchol-Balter [100] and Gandhi and Harchol-Balter [101] also demonstrate that their results generalize to $M/G/k$ /setup with exponential setup times via simulation or analyzing special examples. Recently, Williams et al. [286] go beyond exponential setup times, studying $M/M/k$ /setup with deterministic setup times and FCFS scheduling. However, none of these prior works apply to general setup times, non-Poisson arrivals, or scheduling policies beyond FCFS.

We note that Glazebrook [116], who studies the nonpreemptive $M/G/k$ with Bernoulli feedback, actually studies a more general model that allows for certain types of server unavailability, such as server breakdowns. However, setup times are not covered by Glazebrook [116]. It is likely that more general future work could simultaneously cover setup times, server breakdowns, and other types of server unavailability. See Section 8.10.3 for additional discussion.

Single-Server Models

Compared with multiserver models, single-server models with setup times are better understood [34, 63, 82, 83, 95, 96, 143, 178, 210, 260, 275]. See Doshi [83] for a survey of the work before 1986 and [260] for a more recent survey. These works consider various arrival and service processes, as well as other types of server unavailability in addition to setup times.

However, none of the above works discuss optimal scheduling in the presence of setup times. Progress was made by van Vreumingen [263], who obtains the mean response time of Gittins in the $M/G/1$ /setup as a special case of a more general analysis (Section 8.2.1). But the analysis does not show that other policies might outperform Gittins, nor does it apply to the $G/G/1$ /setup.

8.2.3 Decomposition Laws in Queues

There is a long tradition of proving work decomposition laws for queueing systems [41, 84, 95, 96, 116, 117, 210, 242, 245]. Most of these laws take the form

$$\begin{aligned} \mathbb{E}[\text{work in complex system with M/G arrivals}] \\ = \mathbb{E}[\text{work in M/G/1}] + \mathbb{E}[\text{extra work due to complexity}]. \end{aligned}$$

For example, if the complex system is an M/G/1/setup, the extra work from complexity depends on the setup time distribution. Most work decomposition laws are actually even stronger, holding *distributionally* instead of just in expectation.

We need a work decomposition law where the complexity includes, among other factors, having multiple servers. Such a result for M/G arrivals is relatively recent [242, 245], and no such result exists for G/G arrivals. While there are work decomposition laws for G/G arrivals in the literature [82, 84, 210], to the best of our knowledge, they apply only to single-server models with vacations. To the best of our knowledge, we prove the first work decomposition law for G/G arrivals that holds for multiserver systems like the G/G/ k .

8.3 Model

8.3.1 Core Queueing Models: G/G/ k , G/G/1, M/G/ k , and M/G/1

We consider a G/G/ k queueing model with a single central queue and k identical servers. The system experiences *G/G arrivals*: jobs arrive one-by-one with i.i.d. *interarrival times*, and each job has an i.i.d. *size*, or service requirement. Interarrival times and job sizes are independent of each other. We denote a generic random interarrival time by A and a generic random job size by S .³

³This arrival process is often referred to more specifically as GI/GI arrivals, with the “I” emphasizing the independence assumption. Under this convention, G/G arrivals include even more general stationary arrival processes where independence does not hold. In this work, we focus only on the independent case, so we write simply “G/G” instead of “GI/GI” for brevity.

At any moment of time, a job in the system can be served by one server. Any jobs not in service wait in the queue. Once a job's service is finished, it departs. We follow the convention that each of the k servers has service rate $1/k$. A job of size S thus requires kS time in service to finish. This convention gives all systems we study the same maximum total service rate, namely $k \cdot 1/k = 1$, and thereby the same stability condition.

The name " $G/G/k$ " denotes the fact that the system has G/G arrivals and k servers. When A is exponentially distributed, we write M/G in place of G/G , as in " $M/G/k$ ".

Scheduling Policies

The scheduling policy decides, at every moment in time, which job is in service at which server. We consider a preempt-resume model where preemption occurs without delay or loss of work.

The scheduling objective is minimizing the *mean number of jobs* in the system. We denote the mean number of jobs in system SYS under scheduling policy π by $\mathbb{E}[N]_{\text{SYS}}^{\pi}$, omitting the " SYS " and/or " π " if there is no ambiguity. By Little's law [185], minimizing $\mathbb{E}[N]$ is equivalent to minimizing *mean response time*, the average amount of time a job spends between its arrival and departure.

We use a flexible model of how much the scheduler knows about each job's size (Section 8.3.3). We restrict attention to *non-idling* policies, which are those that never unnecessarily leave servers idle. Nevertheless, our results have implications even for idling policies (Section 8.4.1).

As a consequence of frequent preemption, the server can share one server between multiple jobs. We formalize this in Appendix 8.B.1, but our presentation does not depend on the formal details.

Load and Stability

We write $\lambda = 1/\mathbb{E}[A]$ for the average arrival rate and $\rho = \lambda\mathbb{E}[S]$ for the system's *load*, or utilization. One can think of ρ as the average fraction of servers that are busy. It is clear that $\rho < 1$ is a necessary condition for stability (unless both A and S are deterministic), so we assume this throughout.

Some of our results are stated for the *heavy-traffic limit*. For our purposes, this limit, denoted $\rho \rightarrow 1$, refers to a limit as the job size distribution S remains constant, and the interarrival time distribution A is scaled uniformly

down with its mean approaching the mean job size. That is, the system with load ρ has interarrival time $A_\rho = A_1/\rho$ for some fixed distribution A_1 , where $\mathbb{E}[A_1] = \mathbb{E}[S]$.

It seems intuitive that $\rho < 1$ should be sufficient for stability under non-idling policies, and it is in the G/G/1 [193]. But to the best of knowledge, there are no results characterizing stability of the G/G/ k under complex scheduling policies. Even under FCFS, proving stability of the G/G/ k is not simple, because the system can be stable even when it never empties [166, 215, 251, 278]. Setup times further complicate the matter.

We consider the question of proving stability of the G/G/ k /setup under arbitrary non-idling scheduling policies to be outside the scope of this chapter, so we simply assume stability when $\rho < 1$. We expect this is indeed the case, giving the initial steps of a proof sketch in Appendix 8.D.

Assumption 8.1. For all $\rho < 1$, the G/G/ k /setup is stable under all non-idling scheduling policies, including the Gittins policy.

Additional Assumption on Interarrival Times

Our results for G/G arrivals depend on “how non-Poisson” arrival times are, which we quantify using the following assumption.

Assumption 8.2. There exist $A_{\min}, A_{\max} \in \mathbb{R}_{\geq 0}$ such that $\mathbb{E}[A - a \mid A > a] \in [A_{\min}, A_{\max}]$ for all $a \geq 0$. That is, letting the *interarrival age* A_{age} be the time since the last arrival and *residual interarrival time* A_{res} be the amount of time until the next arrival, we have

$$\mathbb{E}[A_{\text{res}} \mid A_{\text{age}}] \in [A_{\min}, A_{\max}] \quad \text{with probability 1.}$$

One may always use $A_{\min} = \inf_{a \geq 0} \mathbb{E}[A - a \mid A > a]$ and $A_{\max} = \sup_{a \geq 0} \mathbb{E}[A - a \mid A > a]$, so this assumption boils down to the latter being finite.

Our results use Assumption 8.2 via the quantity $\lambda(A_{\max} - A_{\min})$, which we can think of as measuring “how non-Poisson” arrival times are. In the Poisson case, one may use $A_{\min} = A_{\max} = 1/\lambda$, so $\lambda(A_{\max} - A_{\min}) = 0$.

Many interarrival distributions A satisfy Assumption 8.2, such as all phase-type distributions. One can also think of Assumption 8.2 as a relaxation of the well-known *New Better than Used in Expectation* (NBUE) property, which is the special case where $A_{\max} = \mathbb{E}[A]$. The main distributions ruled out by Assumption 8.2 are various classes of heavy-tailed distributions, e.g. power-law tails.

8.3.2 Setup Times

In addition to the basic $G/G/k$ model defined above, we also consider models in which servers require *setup times* to transition from idle to busy. We denote these models with an extra “/setup”, as in $G/G/k/setup$. Whenever a server switches from idle to busy, it must first complete an i.i.d. amount of *setup work*, denoted U . Like work from jobs, servers complete setup work at rate $1/k$, so setup work U results in setup time kU . Setup work amounts are independent of interarrival times and job sizes.

For the purposes of stating our results and proofs in a unified manner, we consider the $G/G/k$ without setup times to be the special case of the $G/G/k/setup$ where $U = 0$ with probability 1.

In our model, a server can be in one of three states:

- *Setting up*, i.e. doing setup work.
- *Busy*, i.e. serving a job.
- *Idle*, i.e. neither serving a job nor doing setup work.

In the $G/G/1/setup$, state transitions are straightforward: the server goes from setting up to busy when it finishes its setup work, from busy to idle when no jobs remain in the system, and from idle to setting up when a job arrives to an empty system. But in the $G/G/k/setup$, the transitions are more complicated. This is because there are several design choices to make, and thus multiple models that might be studied. For example, if we already have one busy server, how many jobs should there be in the queue before we start setting up a second server? For concreteness, we study one particular setup time model, described below, but our work still has implications for alternative models (Sections 8.4.1 and 8.10.3).

In the $G/G/k/setup$, we use the following setup time model: a server transitions

- from setting up to busy when it finishes its setup work,
- from busy to idle when the system has fewer jobs than busy servers, and
- from idle to setting up when the system has fewer busy or setting up servers than jobs.

Thus, transitions to setting up are triggered by arrivals, and transitions to idle are generally triggered by departures. Servers transition “one at a time”, e.g. an arrival triggers at most server to start setting up.

Note that once a setup time begins, it is never canceled, even if the job whose arrival triggered the setup time begins service at another server. Unless another job arrives during the setup time, the server will transition from setting up to busy, then immediately back to idle. Not canceling setup times is a natural modeling choice for some systems, e.g. computer systems where cutting power during startup is undesirable. But our techniques could also be used to analyze setup times that can be canceled (Section 8.10.3).

8.3.3 What the Scheduler Knows About Jobs' Sizes

We consider a flexible model of the scheduler's knowledge called the *Markov-process job model* [242, 243, 245]. In this model, each job has a *state*, which inhabits some *job state space* $\mathbb{X}$, representing what the scheduler knows about that job. Each job's state evolves as an i.i.d. absorbing continuous-time Markov-process $\{X(t)\}_{t \geq 0}$ on some state space $\mathbb{X}$, where $X(t)$ is the state of the job after it has received $t \geq 0$ service. That is, a job's state evolves while it is in service but stays static while it is in the queue. There is an extra absorbing state $\top \notin \mathbb{X}$, corresponding to the job finishing, i.e. jobs exit the system when their state becomes $\top$.

We call $\{X(t)\}_{t \geq 0}$ the *job Markov process*. We can recover the job size from the job Markov process as

$$S = \inf\{t \geq 0 \mid X(t) = \top\}.$$

To clarify, the amount of service t in $X(t)$ is measured in *work* rather than *time*, so jobs evolve at rate $1/k$ when served in a k -server system (Section 8.3.1). As discussed in Appendix 8.B.3, we make some purely technical assumptions on the job Markov process (e.g. r.c.l.l.) to ensure Gittins is well defined.

We assume that the scheduler always knows the state of all jobs in the system, which we denote by $(X_1, \dots, X_N)$. We also assume the scheduler knows the dynamics of the job Markov process, e.g. the size distribution S . A job's state thus encodes everything the scheduler knows about the job. For example, given a job in state x , the scheduler knows its *remaining work*, namely the amount of service the job needs to complete, is distributed as⁴

$$S(x) = (\inf\{t \geq 0 \mid X(t) = \top\} \mid X(0) = x).$$

⁴Abusing notation slightly, we interpret conditioning $X(0) = x$ as the usual notion of starting the job Markov process from state x . This gets around the corner case where the initial state $X(0)$ is never x .

Below are two concrete examples of the Markov-process job model. These are extremes: the first is perfect size information, and the other is zero size information beyond knowing the distribution S . For additional examples, including cases where the scheduler has partial size information, see Scully et al. [245, Section 3].

Example 8.3. The case of *known sizes* is when a job’s state is its remaining work. The state space is $\mathbb{X} = (0, \infty)$, the initial state is distributed as $X(0) \sim S$, and the absorbing state is $\top = 0$. During service, the job’s state decreases at rate 1. That is, $X(t) = (X(0) - t)^+$. In state x , the remaining work is $S(x) = x$.

Example 8.4. The case of *unknown sizes* is when a job’s state is the amount of service it has received so far. The state space is $\mathbb{X} = [0, \infty)$, the initial state is $X(0) = 0$, and the absorbing state is an isolated point $\top$. During service, the job’s state increases at rate 1 and has a chance to jump to $\top$, with the exact chance depending on the distribution of S . That is, $X(t) = t$ until the job completes, after which $X(t) = \top$. In state x , the remaining work is the conditional distribution $S(x) \sim (S - t \mid S > t)$.

8.3.4 The Gittins Policy

The scheduling policy we focus on in this work is the *Gittins* policy (a.k.a. *Gittins index* policy). Gittins is primarily known for the fact that it minimizes $\mathbb{E}[N]$ in the $M/G/1$ [114, 243]. In formulas, we abbreviate Gittins to “Gtn”, as in $\mathbb{E}[N]_{G/G/k/setup}^{\text{Gtn}}$.

The Gittins policy has a relatively simple form. It assigns each job a numerical priority, called a *rank*, where lower rank is better. Gittins always serves the job or jobs of least rank,⁵ and it is non-idling, serving as many jobs as the number of available servers allows. Gittins determines ranks using a *rank function*

$$\text{rank}_{\text{Gtn}} : \mathbb{X} \rightarrow \mathbb{R}_{\geq 0},$$

assigning $\text{rank}_{\text{Gtn}}(x)$ to a job in state $x \in \mathbb{X}$. A job’s rank thus depends only on its own state.

It turns out that our proofs do not directly use the definition of Gittins’s rank function. As such, we specify the Gittins rank function for the concrete

⁵Much literature on the Gittins policy uses the opposite convention, where higher numbers are better. These works typically call a job’s priority its *index* [1, 2, 114], which is the reciprocal of its rank [243].

job Markov processes in Examples 8.3 and 8.4, the latter of which in particular explains the key intuition. We refer the curious reader to Appendix 8.B.2 for the general definition, though we emphasize it does not play a direct role in our proofs.

Example 8.5. In the case of *known* job sizes, it turns out that Gittins reduces to SRPT, which always serves the job of least remaining work. A job’s rank is thus its remaining work. Recalling from Example 8.3 that a job’s state is its remaining work under known sizes, we simply have $\text{rank}_{\text{Gtn}}(x) = x$.

Example 8.6. In the case of *unknown* job sizes, recall from Example 8.4 that a job’s state x is the amount of service it has already received. In this case, the Gittins rank function is [114]

$$\text{rank}_{\text{Gtn}}(x) = \inf_{y > x} \frac{\mathbb{E}[\min\{S, y\} - x \mid S > x]}{\mathbb{P}[S \leq y \mid S > x]}.$$

The intuition for this formula is as follows. Consider a job in state x , and suppose we start serving the job, but decide to “give up” if it reaches state y . On the right-hand side, the numerator is the expected amount of service until we either complete the job or give up, and the denominator is the probability the job completes before we give up. The right-hand side is thus a “service-per-completion” ratio, giving an expected amount of effort it would take to finish one job in expectation. A job’s rank under Gittins is the best service-per-completion ratio one can obtain by optimally choosing the state y in which to give up.

8.4 Main Results

We now state our main results. All of our results hold under the assumptions of Section 8.3, and in particular Assumptions 8.1 and 8.2. As in Section 8.1, we can view a G/G/k/setup system, or any special case thereof, by whether it has (a) multiple servers, (b) non-Poisson arrivals, and (c) setup times. Our bounds use the quantities

$$\begin{aligned} \ell_{(a)} &= C(k-1) \log \frac{1}{1-\rho}, \\ \ell_{(b)} &= \lambda(A_{\max} - A_{\min}), \\ \ell_{(c)} &= \mathbf{1}\{\mathbb{P}[U > 0] > 0\} (2(k-1) + \lambda(A_{\max} + k\mathbb{E}[U_e])), \end{aligned}$$

where $C = \frac{9}{8 \log 1.5} + 1 \approx 3.775$. The idea is that $\ell_{(a)}$ is the loss due to feature (a), as it is nonzero only for systems with $k \geq 2$ servers, and similarly for $\ell_{(b)}$ and $\ell_{(c)}$.⁶

Theorem 8.7. *The performance gap between the Gittins policy in G/G/k/setup and the optimal policy in G/G/1 is bounded by*

$$\mathbb{E}[N]_{G/G/k/setup}^{\text{Gtn}} - \inf_{\pi} \mathbb{E}[N]_{G/G/1}^{\pi} \leq \ell_{(a)} + \ell_{(b)} + \ell_{(c)}.$$

Note that although Theorem 8.7 is not directly about the suboptimality gap of Gittins policy in G/G/k/setup, it still provides an upper bound on the suboptimality gap, because the optimal performance of G/G/1 is a lower bound to G/G/k/setup. This is because servers in G/G/k/setup have speed $1/k$ (Section 8.3.1), so the G/G/1 can mimic any policy in the G/G/k/setup through processor sharing and idling.

With that said, in the special case of the non-idling G/G/1/setup, we can prove a stronger result that drops the $\ell_{(c)}$ term by comparing to a G/G/1/setup instead of a G/G/1.

Theorem 8.8. *In the G/G/1/setup, the performance gap between the Gittins policy and the optimal non-idling policy is bounded by*

$$\mathbb{E}[N]_{G/G/1/setup}^{\text{Gtn}} - \inf_{\pi} \mathbb{E}[N]_{G/G/1/setup}^{\pi} \leq \ell_{(b)}.$$

In particular, in the M/G/1/setup, the Gittins policy minimizes $\mathbb{E}[N]$ among non-idling policies.

The suboptimality gap in Theorem 8.7 is constant when $k = 1$ and $O(\log \frac{1}{1-\rho})$ when $k \geq 2$. In both cases, the gap grows more slowly in the $\rho \rightarrow 1$ limit than $\mathbb{E}[N]_{G/G/1}^{\pi}$, implying heavy-traffic optimality.

Theorem 8.9. *In the G/G/k/setup, if either $k = 1$ or $\mathbb{E}[S^2(\log S)^+] < \infty$, and if either S or A is not deterministic, the Gittins policy is heavy-traffic optimal. Specifically, $\lim_{\rho \rightarrow 1} \mathbb{E}[N]_{G/G/k/setup}^{\text{Gtn}} / \inf_{\pi} \mathbb{E}[N]_{G/G/1}^{\pi} = 1$.*

We prove this result in Section 8.9. The main obstacle is showing a lower bound on $\mathbb{E}[N]_{G/G/1}^{\pi}$. We use SRPT as a lower bound, so the first step of the proof is to analyze SRPT in the heavy-traffic G/G/1 (Theorem 8.28). We find its performance is within a constant factor of SRPT in the heavy-traffic M/G/1.

⁶The reason (8.2) has an $\ell_{(a) \& (c)}$ term instead of an $\ell_{(c)}$ term is because it summarizes both Theorems 8.7 and 8.8.

8.4.1 Remarks on Main Results

Alternative Setup Time Models

Because Theorems 8.7 and 8.9 compare Gittins in the $G/G/k$ /setup to the optimal policy in a $G/G/1$, it also effectively compares Gittins under our setup time model to Gittins in the $G/G/k$ with essentially any other setup time model. This is because the $G/G/1$ serves as a lower bound for alternative setup time models, not just our specific $G/G/k$ /setup. The takeaway is that changing the setup time model does not significantly impact performance in heavy traffic, which makes intuitive sense: servers seldom set up if they are usually busy.

Idling Policies

We say in Section 8.3.1 that we only consider non-idling policies, but Theorem 8.7 still compares Gittins to idling policies. This is because the optimal policy in a $G/G/1$ is clearly non-idling, and, by the discussion above, it gives a lower bound on any policy, idling or non-idling, in the $G/G/k$ /setup.

Why, then, does Theorem 8.8 only compare to the optimal non-idling policy? This is because in the $G/G/1$ /setup, idling the server can change when setup times occur. By idling with one job in the queue, one can effectively control when setup times occur by choosing when to start the job, without waiting for an arrival. Hypothetically, this could improve performance in the $G/G/1$ /setup. That is, idling effectively allows a policy to use an alternative setup time model, so we rule it out.

Opportunities for a Tighter Bound

The bound shown in Theorem 8.7 represents a trade-off between proving a tight bound and stating the result simply. We prioritized making the statement as simple as possible while ensuring the $\ell_{(a)}$ term matches the bound for Gittins in the $M/G/k$ from prior work [242]. But there are at least two clear avenues for tightening our bounds.

First, there are other bounds on Gittins in the $M/G/k$ [129, 245], which can be better than $\ell_{(a)}$ in some cases. We believe that one may take $\ell_{(a)}$ to be the minimum of these bounds, but doing so would complicate the result and proof without substantially changing the main takeaway.

Second, our bound is loose in light traffic. We should have $\mathbb{E}[N] \rightarrow 0$ as $\rho \rightarrow 0$, but our $\ell_{(b)}$ and $\ell_{(c)}$ terms remain nonzero at all loads. One can sharpen our analysis for the special case of the $M/G/k$ /setup to obtain a suboptimality gap that becomes zero in the $\rho \rightarrow 0$ limit. But we doubt even this improved bound is very tight at low loads, so we omit the extra casework.

8.5 Proof Overview

In this section, we give an overview of the proofs of our main results: bounds on Gittins's suboptimality gap (Theorems 8.7 and 8.8) and Gittins's heavy-traffic optimality (Theorem 8.9). At a high level, our proofs work by combining two queueing identities: *WINE*, which is from prior work; and a novel *work decomposition law*, which is built on similar decomposition results from prior work (Section 8.2.3).

The first tool, WINE (Lemma 8.11), expresses the mean number-in-system in terms of *mean r -work* $\mathbb{E}[W_r]$ [242, 243, 245]:

$$\mathbb{E}[N] = \int_0^\infty \frac{\mathbb{E}[W_r]}{r^2} dr.$$

A system's r -work W_r is the total service required to serve all jobs in the system until they all either complete or reach rank greater than r , as determined by rank_{Gtn} (Section 8.3.3). For example, ∞ -work is the total remaining work of all jobs, which we call *total work* or simply *work*. See Section 8.6 for details.

The second tool, the work decomposition law (Theorem 8.17), implies bounds on $\mathbb{E}[W_r]$ under any policy, including Gittins. Combining this with WINE yields bounds on $\mathbb{E}[N]$. Our proof thus boils down to three steps:

- Proving the work decomposition law (Section 8.5.1).
- Using the work decomposition law to bound Gittins's suboptimality gap (Section 8.5.2).
- Using the suboptimality gap bounds to show Gittins is heavy-traffic optimal (Section 8.5.3).

8.5.1 New Tool: Work Decomposition Law for G/G Arrivals

Our work decomposition law, Theorem 8.17, characterizes mean r -work $\mathbb{E}[W_r]$ in the G/G/ k /setup. For simplicity of presentation, below we focus on the special case of mean total work $\mathbb{E}[W]$.

Theorem 8.17 implies that in the G/G/ k /setup under any policy π ,

$$\mathbb{E}[W]^\pi - \mathbb{E}[W]_{G/G/1} \leq \frac{\mathbb{E}[J_I W]^\pi}{1 - \rho} + \frac{\mathbb{E}[J_{\text{setup}} W]^\pi}{1 - \rho} + \rho(A_{\max} - A_{\min}). \quad (8.3)$$

Above, $\mathbb{E}[W]_{G/G/1}$ is the mean work in a non-idling G/G/1, which is policy-invariant; and J_I and J_{setup} are the fraction of idle and setting-up servers, respectively. Flipping the sign on the $\rho(A_{\max} - A_{\min})$ term yields a lower bound instead of an upper bound.

The work decomposition law decomposes work $\mathbb{E}[W]^\pi$ into the policy-invariant term $\mathbb{E}[W]_{G/G/1}$, plus error terms that can depend on the policy π . Each error term characterizes the consequence of a complicating factor that G/G/ k /setup has on the top of the G/G/1 system:

- (a) The first term is due to having multiple servers. It vanishes when $k = 1$, as then $J_I = 0$ if $W > 0$.⁷
- (b) The second term is due to the setup time. It vanishes if servers do not need setup, as then $J_{\text{setup}} = 0$.
- (c) The third term is due to non-Poisson arrivals. It vanishes for Poisson arrivals, as then $A_{\max} = A_{\min}$.

How We Prove the Work Decomposition Law

The proof of work decomposition laws in prior work involves viewing W as a process in the steady state and analyzing its continuous changes and jumps. This strategy works well in M/G systems, because all times have an equal chance of seeing W jump up due to an arrival. But in G/G systems, the chance of having an arrival in the next moment depends on A_{age} , the amount of time since the previous arrival. The jumps of W are thus more complicated to analyze.

⁷When generalizing (8.3) from total work to r -work, there are actually two terms due to having multiple servers. But both vanish when $k = 1$.

The key idea in our proof is to smooth out the non-constant jumping rate of W . Specifically, we consider the process $W - \rho A_{\text{res}}$, which only differs from W by one interarrival time. This process decreases at a constant rate of $1 - \rho$. When an arrival happens, the process jumps, but the expected change is $\mathbb{E}[S] - \rho\mathbb{E}[A] = 0$. Therefore, arrivals only have a “second-order” effect on W , which makes them easier to analyze. This idea builds upon similar smoothing approaches in recent queueing literature [52, 212].

8.5.2 From Work Decomposition to Suboptimality Gap Bounds

We focus here on proving Theorem 8.7, commenting only briefly on the similar proof of Theorem 8.8.

Combining our work decomposition law with WINE gives a formula for Gittins’s suboptimality gap that has the same types of error terms as (8.3). Each error term in the work decomposition law will result in one term in the suboptimality gap $\ell_{(a)} + \ell_{(b)} + \ell_{(c)}$ in Theorem 8.7, after doing the integration and applying some additional treatments that are specific to each term.

Among the three error terms, $\ell_{(a)}$ can be derived similarly to prior work on the $M/G/k$ [129, 242, 245], and $\ell_{(b)}$ follows from Assumption 8.2. But the term corresponding to setup, $\ell_{(c)}$, requires a new analysis. We demonstrate the intuition by bounding $\mathbb{E}[J_{\text{setup}}W]$ in (8.3). First, we write $\mathbb{E}[J_{\text{setup}}W] = \frac{1}{k} \sum_{i=1}^k \mathbb{E}[J_{\text{setup},i}W]$, where $J_{\text{setup},i} = \mathbf{1}\{\text{server } i \text{ is setting up}\}$. Observe that

$$\mathbb{E}[J_{\text{setup},i}W] = \mathbb{P}[J_{\text{setup},i} = 1] \mathbb{E}[W \mid J_{\text{setup},i} = 1].$$

Intuitively, $\mathbb{P}[J_{\text{setup},i} = 1]$ should be diminishing as the load gets heavy because the queue length will get longer the server i will be turned off less frequently. The second factor, $\mathbb{E}[W \mid J_{\text{setup},i} = 1]$, should be bounded because given that the server i is setting up, the work in the system should be no more than the work that arrives during the setup, plus the work that already exists when the setup happens.

For the proof of Theorem 8.8, which gives a tighter bound for the single-server case, we apply WINE and work decomposition law in the same way as above. We will get an expression for $\mathbb{E}[N]_{G/G/1/\text{setup}}^{\text{Gtn}}$ in terms of one $G/G/1$ term, and two error terms corresponding to non-Poisson arrivals and setup times. Instead of analyzing the setup term as in the proof of Theorem 8.7, we make the simple observation that the setup term is the same for all non-idling policies, so it does not contribute to the suboptimality gap.

8.5.3 From Suboptimality Gap Bounds to Heavy-Traffic Optimality

Theorem 8.7 provides an upper bound on the suboptimality gap of Gittins policy in $G/G/k/setup$. To show that the suboptimality gap is small compared with $\inf_{\pi} \mathbb{E}[N]_{G/G/k/setup}^{\pi}$ and establish heavy-traffic optimality of the Gittins policy, we need a lower bound on $\inf_{\pi} \mathbb{E}[N]_{G/G/k/setup}^{\pi}$. This lower bound can be obtained by analyzing $\mathbb{E}[N]_{G/G/1}^{SRPT}$ because SRPT gives the optimal number-in-system in $G/G/1$ with known job sizes [241], which is no more than the optimal number-in-system in $G/G/k/setup$ achievable by a policy that does not know the job size.

We use WINE and work-decomposition law, in a similar way as in the proofs in the suboptimality gaps, to connect SRPT's performance in the $G/G/1$ to its performance in the $M/G/1$. Our end result (Theorem 8.28) shows that $\mathbb{E}[N]_{G/G/1}^{SRPT}$ is a constant factor away from $\mathbb{E}[N]_{M/G/1}^{SRPT}$ as $\rho \rightarrow 1$. This lets us to use the known heavy-traffic asymptotics of SRPT in the $M/G/1$ [183] to lower bound $\mathbb{E}[N]_{G/G/1}^{SRPT}$ and thus show Gittins's heavy-traffic optimality in the $G/G/k/setup$.

8.6 Background on WINE and r -Work

A queueing system's *work* W is the total remaining work of all jobs in the system: $W = \sum_{i=1}^N S(X_i)$, where $S(X_i)$ is the remaining work of job i (Section 8.3.3). We define r -work similarly: we first define the remaining r -work of a job, then define the system's r -work to be the sum of all jobs' remaining r -work.

Definition 8.10. Let $r \geq 0$. The *remaining r -work* of a job in state x , denoted $S_r(x)$, is the amount of service it needs until it either finishes or reaches a state whose rank is at least r :

$$\begin{aligned} S_r(x) &= \text{amount of service a job starting at } x \text{ needs} \\ &\quad \text{to finish or reach rank at least } r \\ &= (\inf\{t \geq 0 \mid X(t) = \top \text{ or } \text{rank}_{\text{Gtn}}(X(t)) \geq r\} \mid X(0) = x). \end{aligned}$$

A system's r -work, denoted W_r , is the sum of the remaining r -work of all jobs in the system: $W_r = \sum_{i=1}^N S_r(X_i)$.

We now present the WINE identity. It holds for any scheduling policy that has access to only the current and past system states (Section 8.3.3). For concreteness, we state WINE for our specific queueing model, but it holds in essentially any system which uses the Markov-process job model.

Lemma 8.11 (WINE [245, Theorem 6.3]). *In the $G/G/k$ /setup under any scheduling policy,*

$$N = \int_0^\infty \frac{\mathbb{E}[W_r \mid X_1, \dots, X_N]}{r^2} dr, \quad \mathbb{E}[N] = \int_0^\infty \frac{\mathbb{E}[W_r]}{r^2} dr.$$

WINE, which integrates the entire system's r -work to get the number of jobs, follows from a more basic identity, sometimes called “single-job WINE” [242], which integrates a single job's remaining r -work.

Lemma 8.12 (Single-Job WINE [245, Lemma 6.2]). *For any job state $x \in \mathbb{X}$, we have $\int_0^\infty \frac{\mathbb{E}[S_r(x)]}{r^2} dr = 1$.*

One subtlety about WINE is that while it applies to any scheduling policy, the definition of r -work uses Gittins's rank function. As a general rule, this makes analyzing Gittins's performance using WINE easier than analyzing other policies' performance using WINE, particularly when proving upper bounds, though there are some exceptions [242, 249]. Our work is no exception: we prove our main results by upper bounding Gittins's r -work and lower-bounding the optimal policies' r -work.

8.6.1 Additional Definitions for Reasoning About r -Work

WINE reduces the problem of analyzing the steady-state mean number of jobs $\mathbb{E}[N]$ to the problem of analyzing steady-state mean r -work $\mathbb{E}[W_r]$. In order to analyze r -work, we need to understand the means by which the amount of r -work in the system changes over time. This section introduces the standard concepts and vocabulary used to discuss r -work [242, 244, 245].

The definitions in this section are parameterized by a rank $r \geq 0$, as denoted by a prefix “ r -”. We often drop this prefix when the rank r is clear from context or not important to the discussion.

Relevant, Irrelevant, Fresh, and Recycled Jobs

We call a job *r -relevant* whenever its rank is less than r . Otherwise, the job is *r -irrelevant*.

Whether a job is r -relevant or r -irrelevant varies over time. Consider one job's journey through the system. When the job arrives, it may be either, depending on its initial state $X(0)$. As the job is served, its rank can go up and down, so it may alternate between r -relevant and r -irrelevant, possibly multiple times, before eventually finishing and exiting the system.

From the above discussion, it is evident that there are two ways for the amount of r -work in a system to increase. Both are important, so we introduce terminology for discussing both.

Definition 8.13. We call an r -relevant job *r -fresh* if it has been r -relevant ever since its arrival. That is, new arrivals that are initially r -relevant are r -fresh until they either finish or become r -irrelevant.

- We write $S_r = S_r(X(0))$ for the random amount of service during which a newly arrived job is r -fresh. Arriving jobs may be r -irrelevant, so it may be that $S_r = 0$ with nonzero probability.
- We call $\rho_r = \lambda \mathbb{E}[S_r]$ the *r -fresh load*. It is the average rate r -work is added by new arrivals.

Definition 8.14. We call an r -relevant job *r -recycled* if it was r -irrelevant at some point in the past. We refer to the moment a job switches from r -irrelevant to r -recycled as an *r -recycling*.

- We write $\lambda_{r\text{-rcy}}$ for the average rate of r -recyclings.
- We write $S_{r\text{-rcy}}$ for the random amount of r -work added by a single r -recycling.
- We call $\rho_{r\text{-rcy}} = \lambda_{r\text{-rcy}} \mathbb{E}[S_{r\text{-rcy}}]$ the *r -recycled load*. It is the average rate r -work is added by r -recyclings.

Note that r -recyclings can only occur when an r -irrelevant job is in service. This is because a job's rank depends only on its state, which only changes during service (Sections 8.3.3 and 8.3.4).

We assume for ease of presentation that $\lambda_{r\text{-rcy}} < \infty$ for all $r \geq 0$. This is actually not always the case, but the assumption can be straightforwardly relaxed by appealing to more advanced Palm calculus techniques. See Scully et al. [245, Appendix E] for details.

Server States

Finally, we need notations for discussing how many servers are doing r -work or setting up.

Definition 8.15.

- We write J_r for the fraction of servers that are serving r -relevant jobs.⁸
Note that $\mathbb{E}[J_r] = \rho_r + \rho_{r\text{-rcy}}$.
- We write J_{setup} for the fraction of servers that are setting up.

8.7 Work Decomposition Law for G/G Arrivals

In this section, we introduce a new formula for the expected r -work in systems with G/G arrivals, which we call the *work decomposition law*. The formula decomposes mean r -work into multiple terms, where each term is either fixed or easy to analyze. To present the formula and its proof, we need to first review a concept called *Palm expectation*. We give an intuitive review that suffices for our purposes below, referring the reader to Miyazawa [211] and Baccelli and Brémaud [26] for more formal treatments.

8.7.1 Palm Expectation and Notation

Consider the amount of work in the system as seen by arriving jobs. Intuitively speaking, to define the long-run average as seen by arrivals, we would like to start with the steady-state average, then condition on the event “an arrival is occurring”. However, the probability that an arrival occurs at any given point in time is 0, so we cannot naively apply the usual conditional expectation formula. Palm expectations solve this problem, giving a way of formally defining a notion that corresponds to conditioning on events like “an arrival is occurring”. We use two main Palm expectations in our analysis of r -work:

- The *Palm expectation at arrivals*, denoted $\mathbb{E}_{\text{arv}}[\cdot]$.
- The *Palm expectation at r -recyclings*, denoted $\mathbb{E}_{r\text{-rcy}}[\cdot]$.

That is, $\mathbb{E}_{\text{arv}}[V]$ is the expectation of random quantity V , which may depend on the system state, sampled at the time “just before” an arrival occurs; and $\mathbb{E}_{r\text{-rcy}}[V]$ is the same, but for r -recyclings instead of arrivals.

Although we define $\mathbb{E}_{\text{arv}}[\cdot]$ and $\mathbb{E}_{r\text{-rcy}}[\cdot]$ as referring to the moment “just before” an event, it is also helpful to be able to refer to the quantity “just after” the event. To facilitate this, we use the following notation:

⁸If processor sharing is occurring, we interpret J_r as the total service rate assigned to r -relevant jobs (Appendix 8.B.1).

- Within $\mathbb{E}_{\text{arv}}[\cdot]$, we denote the remaining r -work of the arriving job by S_r .
 - By our independence assumptions (Section 8.3.1), S_r is independent of the system state.
- Within $\mathbb{E}_{r\text{-rcy}}[\cdot]$, we denote the remaining r -work of the job being recycled by $S_{r\text{-rcy}}$.
 - In general, $S_{r\text{-rcy}}$ is *not* independent of the system state, because recyclings are caused by events happening within the system.

A Notation Shortcut

In addition to the Palm expectation, we also define the notation $\mathbb{E}_{r\text{-acc}}[\cdot]$ as

$$\mathbb{E}_{r\text{-acc}}[V] = \frac{\mathbb{E}[(1 - J_r)V] + \lambda_{r\text{-rcy}}\mathbb{E}_{r\text{-rcy}}[S_{r\text{-rcy}}V]}{1 - \rho_r}, \quad (8.4)$$

where J_r is the fraction of servers that are busy with r -relevant jobs (Section 8.6.1). One can interpret $\mathbb{E}_{r\text{-acc}}[\cdot]$ as a type of Palm expectation. For instance, it behaves like an expectation in the sense that $\mathbb{E}_{r\text{-acc}}[v] = v$ for deterministic values v . But for our purposes, it suffices to understand $\mathbb{E}_{r\text{-acc}}[\cdot]$ as simply a notation shortcut.

Excess Distributions

We introduce a piece of notation that occurs frequently in queueing and renewal theory [16, 137], including in the statement of our work decomposition law below.

Definition 8.16. Given a nonnegative distribution V , we define its *excess*, denoted V_e , to be the distribution with tail⁹

$$\mathbb{P}[V_e > t] = \int_t^\infty \frac{\mathbb{P}[V > u]}{\mathbb{E}[V]} du, \quad \text{which has mean} \quad \mathbb{E}[V_e] = \frac{\mathbb{E}[V^2]}{2\mathbb{E}[V]}.$$

8.7.2 Statement and Proof of Work Decomposition Law

Now we are ready to present the work decomposition law for systems with G/G arrivals. We state the result for r -work W_r , but we can apply the result to total work W by taking an $r \rightarrow \infty$ limit.

⁹As a corner case, if $V = 0$ with probability 1, we let $V_e = 0$ with probability 1.

Theorem 8.17 (Work Decomposition Law for G/G Arrivals). *In the G/G/k/setup under any policy π ,*

$$\begin{aligned} \mathbb{E}[W_r]^\pi = & \frac{\rho_r (\mathbb{E}[(S_r)_e] - \mathbb{E}[S_r] + \mathbb{E}[A_e]) + \rho_{r\text{-rcy}} \mathbb{E}[(S_{r\text{-rcy}})_e]}{1 - \rho_r} \\ & + \mathbb{E}_{r\text{-acc}}[W_r]^\pi - \rho_r \mathbb{E}_{r\text{-acc}}[A_{\text{res}}]^\pi. \end{aligned}$$

Proof. We drop the superscript π throughout. For each r , define

$$Z_r = \frac{1}{2} (W_r - \rho_r A_{\text{res}})^2. \quad (8.5)$$

We use Z_r as a “test function” and extract information about the G/G system by looking at how Z_r changes and applying Miyazawa’s Rate Conservation Law (RCL) [211]. We discuss why Z_r is the right choice of test function in Remark 8.19 below.

Over time, the quantity Z_r changes in the following ways: continuous change as r -work and remaining arrival time decrease over time, jump when a new job arrives, and jump at an r -recycling event. We use Z'_r to denote the continuous change of Z_r , use $\Delta_{\text{arv}} Z_r$ to denote the jumps of Z_r at arrival times, and use $\Delta_{\text{rcy}} Z_r$ to denote jumps of Z_r at recycling times. Analogous notations are used for W_r and A_{res} .

Miyazawa’s RCL [211] implies

$$\mathbb{E}[Z'_r] + \lambda \mathbb{E}_{\text{arv}}[\Delta_{\text{arv}} Z_r] + \lambda_{r\text{-rcy}} \mathbb{E}_{r\text{-rcy}}[\Delta_{r\text{-rcy}} Z_r] = 0. \quad (8.6)$$

This RCL is simply describing the fact that the contribution of continuous changes ($\mathbb{E}[Z'_r]$) and jumps ($\mathbb{E}_{\text{arv}}[\Delta_{\text{arv}} Z_r]$ and $\lambda_{r\text{-rcy}} \mathbb{E}_{r\text{-rcy}}[\Delta_{r\text{-rcy}} Z_r]$) in Z_r cancel out in the long-run average sense.

To extract information about the G/G system, we analyze Z'_r , $\Delta_{\text{arv}} Z_r$, and $\Delta_{r\text{-rcy}} Z_r$ as below.

- At all times, W_r decreases at rate $-W'_r = J_r$, and A_{res} decreases at rate $-A'_{\text{res}} = 1$, so Z_r decreases at rate $-Z'_r = (J_r - \rho_r)(W_r - \rho_r A_{\text{res}})$.
- When a new arrival happens, Z_r jumps as follows. The new job contributes r -work S_r , so W_r jumps up by $\Delta_{\text{arv}} W_r = S_r$. And by definition, the new arrival happens just as A_{res} reaches 0, at which point it jumps up to a newly sampled interarrival time A , so $\Delta_{\text{arv}} A_{\text{res}} = A$. This means that when a new arrival happens, Z_r jumps by $\Delta_{\text{arv}} Z_r = (S_r - \rho_r A)W_r + \frac{1}{2}(S_r - \rho_r A)^2$.

- When an r -recycling happens, the r -work W_r increases by $\Delta_{r\text{-rcy}}W_r = S_{r\text{-rcy}}$, and A_{res} is unaffected, so $\Delta_{r\text{-rcy}}Z_r = S_{r\text{-rcy}}(W_r - \rho_r A_{\text{res}}) + \frac{1}{2}S_{r\text{-rcy}}^2$.

Given the above formulas for Z'_r , $\Delta_{\text{arv}}Z_r$ and $\Delta_{r\text{-rcy}}Z_r$, the terms in (8.6) can be computed one-by-one as follows. For $\mathbb{E}[Z'_r]$, we have

$$\begin{aligned}\mathbb{E}[Z'_r] &= -(1 - \rho_r)\mathbb{E}[W_r] + \rho_r(1 - \rho_r)\mathbb{E}[A_e] \\ &\quad + \mathbb{E}[(1 - J_r)W_r] - \rho_r\mathbb{E}[(1 - J_r)A_{\text{res}}],\end{aligned}$$

where we have used the fact that by basic renewal theory, $\mathbb{E}[A_{\text{res}}] = \mathbb{E}[A_e]$. For $\lambda\mathbb{E}_{\text{arv}}[\Delta_{\text{arv}}Z_r]$, we have

$$\begin{aligned}\lambda\mathbb{E}_{\text{arv}}[\Delta_{\text{arv}}Z_r] &= \lambda\mathbb{E}_{\text{arv}}\left[(S_r - \rho_r A)W_r + \frac{1}{2}(S_r - \rho_r A)^2\right] \\ &= \lambda\mathbb{E}[S_r - \rho_r A]\mathbb{E}_{\text{arv}}[W_r] + \frac{1}{2}\lambda\mathbb{E}[(S_r - \rho_r A)^2] \quad (8.7) \\ &= \rho_r\mathbb{E}[(S_r)_e] + \rho_r^2\mathbb{E}[A_e] - \rho_r\mathbb{E}[S_r],\end{aligned}$$

where the second equality is due to the fact that the new job's r -work S_r and next interarrival time A are independent of the previous amount of r -work W_r , and the third equality uses Definition 8.16 and the fact that $\mathbb{E}[S_r] = \rho_r\mathbb{E}[A]$. Finally, for $\lambda_{r\text{-rcy}}\mathbb{E}_{r\text{-rcy}}[\Delta_{r\text{-rcy}}Z_r]$, we have

$$\begin{aligned}\lambda_{r\text{-rcy}}\mathbb{E}_{r\text{-rcy}}[\Delta_{r\text{-rcy}}Z_r] &= \lambda_{r\text{-rcy}}\mathbb{E}_{r\text{-rcy}}[S_{r\text{-rcy}}W_r] - \rho_r\lambda_{r\text{-rcy}}\mathbb{E}_{r\text{-rcy}}[S_{r\text{-rcy}}A_{\text{res}}] \\ &\quad + \rho_{r\text{-rcy}}\mathbb{E}[(S_{r\text{-rcy}})_e].\end{aligned}$$

Combining the three terms with (8.6) completes the proof. $\square$

Remark 8.18. The proof of Theorem 8.17 does not depend on the details of the Gittins policy. It relies only partitioning the job states, with one part playing the role of states with rank less than r . See Scully [242, Sections 7.2 and 8.3] for an example of this with M/G arrivals.

Remark 8.19. The basic reason for multiplying A_{res} by ρ_r in the definition of Z_r is that we want to avoid having a $\mathbb{E}_{\text{arv}}[W_r]$ term in (8.6), as it is likely to be a large and intractable term. As we can see from (8.7), when computing $\lambda\mathbb{E}_{\text{arv}}[\Delta_{\text{arv}}Z_r]$, the term involving $\mathbb{E}_{\text{arv}}[W_r]$ vanishes because $\mathbb{E}[S_r - \rho_r A] = \mathbb{E}[S_r](1 - \lambda\mathbb{E}[A]) = 0$. Intuitively, ensuring $W_r - \rho_r A_{\text{res}}$ has zero change in expectation when a new job arrives prevents the Palm expectation of arrivals from appearing in the RCL equation. This trick appears throughout the literature on applying the RCL to queues [52, 212].

8.8 Bounding Gittins's Suboptimality Gap

In this section, we prove the main results using WINE and the work decomposition law introduced in Sections 8.6 and 8.7. We first derive a general formula that decomposes the suboptimality gap into four terms and analyze each term based on the specific settings of each theorem. We express the formula in terms of the following quantities.

Definition 8.20. We define *residual interarrival cost* m_{res} , *recycling cost* m_{rcy} , *idleness cost* m_{idle} , and *setup cost* m_{setup} as follows:

$$\begin{aligned} m_{\text{res}} &= \int_0^\infty \frac{-\rho_r \mathbb{E}_{r\text{-acc}}[A_{\text{res}}]}{r^2} dr, \\ m_{\text{idle}} &= \int_0^\infty \frac{\mathbb{E}[(1 - J_r - J_{\text{setup}})W_r]}{r^2(1 - \rho_r)} dr, \\ m_{\text{rcy}} &= \int_0^\infty \frac{\lambda_{r\text{-rcy}} \mathbb{E}_{r\text{-rcy}}[S_{r\text{-rcy}}W_r]}{r^2(1 - \rho_r)} dr, \\ m_{\text{setup}} &= \int_0^\infty \frac{\mathbb{E}[J_{\text{setup}}W_r]}{r^2(1 - \rho_r)} dr. \end{aligned}$$

Lemma 8.21 (Decomposition of Performance Difference). *The performance difference between the Gittins policy in $G/G/k/\text{setup}$ and any policy π in the $G/G/1$ (or $G/G/1/\text{setup}$) can be decomposed as below:*

$$\mathbb{E}[N]_{G/G/k/\text{setup}}^{\text{Gtn}} - \mathbb{E}[N]_{G/G/1}^\pi = (m_{\text{res}}^{\text{Gtn}} - m_{\text{res}}^\pi) + m_{\text{rcy}}^{\text{Gtn}} + m_{\text{idle}}^{\text{Gtn}} + (m_{\text{setup}}^{\text{Gtn}} - m_{\text{setup}}^\pi).$$

Proof. Only the last two terms in Theorem 8.17, namely $\mathbb{E}_{r\text{-acc}}[W_r]$ and $\rho_r \mathbb{E}_{r\text{-acc}}[A_{\text{res}}]$, depend on the specific scheduling policy. After expanding the definitions of the cost terms (Definition 8.20) and $\mathbb{E}_{r\text{-acc}}[\cdot]$ ((8.4) of Section 8.7.1), the result follows immediately from WINE (Lemma 8.11) and the fact that m_{rcy}^π and m_{idle}^π are nonnegative. $\square$

Note that we state Lemma 8.21 in terms of the Gittins policy only because our focus is the optimality of Gittins. The lemma is still true if we replace Gittins with any other policy.

Bounding Gittins's suboptimality gap thus reduces to bounding the four terms in Lemma 8.21. We address one in each of Sections 8.8.1 to 8.8.4, combining the bounds to prove our main results in Section 8.8.5.

8.8.1 Analysis of the Residual Interarrival Cost

Proposition 8.22 (Residual Interarrival Cost). *For any policy π ,*

$$m_{\text{res}}^{\text{Gtn}} - m_{\text{res}}^{\pi} \leq \lambda(A_{\text{max}} - A_{\text{min}}).$$

Proof. Observe that for deterministic v , we have $\mathbb{E}_{r\text{-acc}}[v] = v$ by the computation

$$\mathbb{E}_{r\text{-acc}}[v] = \frac{\mathbb{E}[(1 - J_r)v] + \lambda_{r\text{-rcy}}\mathbb{E}[S_{r\text{-rcy}}v]}{1 - \rho_r} = \frac{1 - \rho_r - \rho_{r\text{-rcy}} + \rho_{r\text{-rcy}}}{1 - \rho_r}v = v.$$

The result follows from the fact that $\mathbb{E}_{r\text{-acc}}[A_{\text{res}}] = \mathbb{E}_{r\text{-acc}}[\mathbb{E}[A_{\text{res}} \mid A_{\text{age}}]]$ and Assumption 8.2. $\square$

8.8.2 Analysis of the Recycling Cost

Proposition 8.23 (Recycling Cost). *In the $G/G/k/\text{setup}$, under the Gittins policy, we have*

$$m_{\text{rcy}}^{\text{Gtn}} \leq (k - 1) \log \frac{1}{1 - \rho}.$$

Proof. The same bound has been shown for $M/G/k$ without setup times, e.g. [242, Proposition 17.9] and [129, Lemma B.5]. It turns out that the prior proofs rely only on the following fact:

Immediately before an r -recycling, the number of r -relevant jobs is at most $k - 1$.

This fact still holds in the $G/G/k/\text{setup}$ under Gittins, so the same proof goes through. The fact holds because immediately before the recycling, the job that is about to be recycled is in service but is r -irrelevant. If there were k jobs that were r -relevant, they would have priority under Gittins, preventing the r -irrelevant job from being in service and thus preventing the r -recycling. $\square$

8.8.3 Analysis of the Idleness Cost

Proposition 8.24 (Idleness Cost). *In the $G/G/k/\text{setup}$, under the Gittins policy, we have*

$$m_{\text{idle}}^{\text{Gtn}} \leq (C - 1)(k - 1) \log \frac{1}{1 - \rho} + (k - 1)\mathbb{1}\{\mathbb{P}[U > 0] > 0\},$$

where $C = \frac{9}{8 \log 1.5} + 1 \approx 3.775$.

The proof of Proposition 8.24 proceeds similarly to the proof of [242, Proposition 17.6], but with a small modification to account for setup times. Given the similarity to prior work, we defer it to Appendix 8.C.

8.8.4 Analysis of the Setup Cost

Proposition 8.25 (Single-Server Setup Cost). *In the $G/G/1/setup$, the setup cost is fixed for any setup-non-idling policy. In particular, since Gtn is also a setup-non-idling policy, we have*

$$m_{\text{setup}}^{\text{Gtn}} - m_{\text{setup}}^{\pi} = 0,$$

for any other setup-non-idling policy π .

Proof of Proposition 8.25. Because $J_{\text{setup}} \in \{0, 1\}$, we have

$$\mathbb{E}[J_{\text{setup}} W_r]^{\pi} = \mathbb{E}[W_r \mid J_{\text{setup}} = 1]^{\pi} \mathbb{P}[J_{\text{setup}} = 1]^{\pi}.$$

Therefore, recalling Definition 8.20, it suffices to show that both $\mathbb{P}[J_{\text{setup}} = 1]^{\pi}$ and $\mathbb{E}[W_r \mid J_{\text{setup}} = 1]^{\pi}$ do not depend on setup-non-idling policy π . Observe that under any setup-non-idling policy, the distributions of busy periods (the continuous periods when there is work in the system) are unaffected by the order of serving specific jobs, and J_{setup} equals 1 only during the first S_{setup} unit of time in each busy period, so $\mathbb{P}[J_{\text{setup}} = 1]^{\pi}$ do not depend on π . As for $\mathbb{E}[W_r \mid J_{\text{setup}} = 1]^{\pi}$, because the server cannot serve jobs when setting up and there is no r -work in the system when the setup begins, W_r is determined by the amount of r -work that has arrived since the setup begins, whose distribution is independent of the policy. $\square$

Proposition 8.26 (Multiserver Setup Cost). *In the $G/G/k/setup$, the setup cost under any setup-non-idling π has the following bound:*

$$m_{\text{setup}}^{\pi} \leq \mathbb{1}\{\mathbb{P}[U > 0] > 0\} (\lambda k \mathbb{E}[U_e] + \lambda A_{\max} + k - 1). \quad (8.8)$$

In particular, the bound holds for Gittins.

To prove Proposition 8.26, we require a helper lemma. The lemma bounds the expected number of jobs in the system during a setup time. To state the lemma, we let $J_{\text{setup},i}$ be the indicator of whether server i is setting up and let $U_{\text{age},i}$ be the age of server i 's setup process for each $i = 1, 2, \dots, n$. We set $U_{\text{age},i}$ to zero if server i is not setting up.

Lemma 8.27. *In the $G/G/k/setup$, for any server i and all $a \geq 0$, we have*

$$\mathbb{E}[N \mid J_{\text{setup},i} = 1, U_{\text{age},i} = a] \leq \lambda a + \lambda A_{\max} + k - 1. \quad (8.9)$$

The proof of Lemma 8.27 is nontrivial but uses standard techniques, so we defer it to Appendix 8.C and move on to proving Proposition 8.26.

Proof of Proposition 8.26. The case where $U = 0$ is clear, so we assume that $U > 0$ with nonzero probability. We first bound r -setup error using the fact that $\rho_r \leq \rho$.

$$\begin{aligned} m_{\text{setup}} &= \int_0^\infty \frac{\mathbb{E}[J_{\text{setup}} W_r]}{r^2(1 - \rho_r)} dr \leq \int_0^\infty \frac{\mathbb{E}[J_{\text{setup}} W_r]}{r^2(1 - \rho)} dr \\ &= \int_0^\infty \frac{\mathbb{E}[\mathbb{E}[J_{\text{setup}} W_r \mid X_1, \dots, X_N]]}{r^2(1 - \rho)} dr, \end{aligned} \quad (8.10)$$

where $X_1, X_2, \dots, X_N$ are the states of all the jobs in the system (Section 8.3.3). Using Tonelli's theorem and WINE (Lemma 8.11), the last expression can be rewritten as

$$\begin{aligned} &\int_0^\infty \frac{\mathbb{E}[\mathbb{E}[J_{\text{setup}} W_r \mid X_1, \dots, X_N]]}{r^2(1 - \rho)} dr \\ &= \mathbb{E} \left[\frac{J_{\text{setup}}}{1 - \rho} \int_0^\infty \frac{\mathbb{E}[W_r \mid X_1, \dots, X_N]}{r^2} dr \right] \\ &= \frac{\mathbb{E}[J_{\text{setup}} N]}{1 - \rho}. \end{aligned} \quad (8.11)$$

By $J_{\text{setup}} = \frac{1}{k} \sum_{i=1}^k J_{\text{setup},i}$, $\mathbb{E}[J_{\text{setup},i} N] = \mathbb{E}[\mathbb{E}[N \mid J_{\text{setup},i} = 1, U_{\text{age},i}]]$ and Lemma 8.27,

$$\frac{\mathbb{E}[J_{\text{setup}} N]}{1 - \rho} \leq \frac{\lambda}{k(1 - \rho)} \sum_{i=1}^k \mathbb{E}[J_{\text{setup},i} U_{\text{age},i}] + \frac{\lambda A_{\max} + k - 1}{k(1 - \rho)} \sum_{i=1}^k \mathbb{E}[J_{\text{setup},i}].$$

Now it remains to compute $\sum_{i=1}^k \mathbb{E}[J_{\text{setup},i}]$ and $\sum_{i=1}^k \mathbb{E}[J_{\text{setup},i} U_{\text{age},i}]$. The mean fraction of servers setting up is no more than the mean fraction of non-busy servers, which is $1 - \rho$, so

$$\frac{1}{k} \sum_{i=1}^k \mathbb{E}[J_{\text{setup},i}] = \mathbb{E}[J_{\text{setup}}] \leq 1 - \rho, \quad (8.12)$$

Basic renewal theory and the $1/k$ service rate (Section 8.3.1) imply the average age of a setup time is $k\mathbb{E}[U_e]$ (Definition 8.16), so

$$\frac{1}{k} \sum_{i=1}^k \mathbb{E}[J_{\text{setup},i} U_{\text{age},i}] = \sum_{i=1}^k \mathbb{E}[J_{\text{setup},i}] \mathbb{E}[U_e]. \quad (8.13)$$

Combining (8.10)–(8.13) finishes the proof. $\square$

8.8.5 Proofs of Main Results

Proof of Theorem 8.7. After expressing the suboptimality gap using Lemma 8.21, we apply Propositions 8.22 to 8.24 and 8.26 and use the fact that m_{setup}^π is non-negative. Grouping the $\log \frac{1}{1-\rho}$ terms to form $\ell_{(a)}$ and grouping the $\mathbb{1}\{\mathbb{P}[U > 0] > 0\}$ terms to form $\ell_{(c)}$ yields the result. $\square$

Proof of Theorem 8.8. After expressing the suboptimality gap using Lemma 8.21, we apply Propositions 8.22 to 8.25. The only nonzero contribution comes from $m_{\text{res}}^{\text{Gtn}} - m_{\text{res}}^\pi$, which Proposition 8.22 shows to be at most $\ell_{(b)}$. $\square$

8.9 Heavy-traffic Optimality

We now turn to prove Theorem 8.9, which amounts to showing that Gittins's suboptimality gap, namely $\mathbb{E}[N]^{\text{Gtn}} - \inf_\pi \mathbb{E}[N]^\pi$ is small relative to the performance of the optimal policy, namely $\inf_\pi \mathbb{E}[N]^i$. It turns out that this is indeed the case: the suboptimality gap is small relative to the performance of SRPT in the G/G/1, which is a lower bound on the performance of any policy in the G/G/k/setup.

We first relate SRPT's performance in the G/G/1 to its performance in the M/G/1, which is known from prior work [183]. We then use this result to prove Theorem 8.9.

Theorem 8.28. *Given Assumption 8.2, in the heavy traffic limit, we have*

$$\lim_{\rho \rightarrow 1} \frac{\mathbb{E}[N]_{G/G/1}^{\text{SRPT}}}{\mathbb{E}[N]_{M/G/1}^{\text{SRPT}}} = \frac{c_S^2 + c_A^2}{c_S^2 + 1},$$

where $c_V^2 = \text{Var}[V]/\mathbb{E}[V]^2$ and the two systems have the same service time distribution and average arrival rate. If $c_S^2 = \infty$, then we interpret the right-hand side as 1.

Proof of Theorem 8.28. If $c_S^2 = c_A^2 = 0$, the result holds because $\mathbb{E}[N] \rightarrow \infty$ in the M/G/1 but not the G/G/1, which in this case are the M/D/1 and D/D/1, respectively (D is for “deterministic”), implying the result. So we focus on the case where $c_S^2 + c_A^2 > 0$. We present the full proof for the $c_S^2 < \infty$ case first, briefly sketching how to adapt the argument to the $c_S^2 = \infty$ case at the end.

Since SRPT is a special case of Gittins (Example 8.5), we can analyze $\mathbb{E}[N]$ in the G/G/1 under SRPT using WINE and the work decomposition law. By Lemma 8.11, Theorem 8.17, and Definition 8.20, we have

$$\begin{aligned} \mathbb{E}[N] = & \int_0^\infty \frac{\rho_r (\mathbb{E}[(S_r)_e] - \mathbb{E}[S_r] + \mathbb{E}[A_e]) + \rho_{r\text{-rcy}} \mathbb{E}[(S_{r\text{-rcy}})_e]}{r^2(1 - \rho_r)} dr \\ & + m_{\text{setup}} + m_{\text{idle}} + m_{\text{rcy}} + m_{\text{res}}. \end{aligned}$$

Since SRPT is a Gittins policy, by Propositions 8.22 to 8.25, we have that in the G/G/1, $m_{\text{setup}} = m_{\text{idle}} = m_{\text{rcy}} = 0$ and $m_{\text{res}} \in [-\lambda A_{\max}, -\lambda A_{\min}]$. We also note that under SRPT, the rank of a job is its remaining size, which together with Definition 8.16 implies

$$\begin{aligned} \mathbb{E}[S_r] &= \mathbb{E}[S \mathbf{1}\{S \leq r\}], \\ \rho_r \mathbb{E}[(S_r)_e] &= \frac{\lambda}{2} \mathbb{E}[S^2 \mathbf{1}\{S \leq r\}], \\ \rho_{r\text{-rcy}} \mathbb{E}[(S_{r\text{-rcy}})_e] &= \frac{\lambda}{2} r^2 \mathbb{P}[S > r]. \end{aligned}$$

Borrowing the notation of Harchol-Balter [137], we write

$$S_{\bar{r}} = \min\{S, r\} \qquad \rho_{\bar{r}} = \lambda \mathbb{E}[S_{\bar{r}}],$$

giving us

$$\mathbb{E}[N] = \int_0^\infty \frac{\rho_{\bar{r}} \mathbb{E}[(S_{\bar{r}})_e] + \rho_r (-\mathbb{E}[S_r] + \mathbb{E}[A_e])}{r^2(1 - \rho_r)} dr + O(1). \quad (8.14)$$

in the G/G/1 under SRPT. Using Definition 8.16 and the fact that $\mathbb{E}[S] = \rho \mathbb{E}[A]$, we can compute

$$\rho (\mathbb{E}[S_e] - \mathbb{E}[S] + \mathbb{E}[A_e]) = \frac{\lambda}{2} (\text{Var}[S] + \text{Var}[A]) + O(1 - \rho).$$

Reasoning similarly and using the fact that $\rho_{\bar{r}} - \rho_r \leq 1 - \rho_r$, we have

$$\rho_{\bar{r}} \mathbb{E}[(S_{\bar{r}})_e] + \rho_r (-\mathbb{E}[S_r] + \mathbb{E}[A_e]) = \frac{\lambda}{2} (\text{Var}[S_{\bar{r}}] + \text{Var}[A]) + O(1 - \rho_r).$$

Because $S_{\bar{r}} = \min\{S, r\}$, we have $\lim_{r \rightarrow \infty} \text{Var}[S_{\bar{r}}] = \text{Var}[S]$. Thus, for all $\epsilon > 0$, there exists r^* such that for any $r \geq r^*$, we have $|\text{Var}[S_{\bar{r}}] - \text{Var}[S]| \leq \epsilon$. Since S does not depend on ρ , we can fix a sufficiently small constant ϵ so that r^* is also a constant independent of ρ . Applying (8.14), we can write $\mathbb{E}[N]$ as

$$\begin{aligned} \mathbb{E}[N] &= \int_0^{r^*} \frac{\rho_{\bar{r}} \mathbb{E}[(S_{\bar{r}})_e] + \rho_r (-\mathbb{E}[S_r] + \mathbb{E}[A_e])}{r^2(1 - \rho_r)} dr \\ &\quad + \int_{r^*}^{\infty} \left(\frac{\frac{\lambda}{2}(\text{Var}[S_{\bar{r}}] + \text{Var}[A])}{1 - \rho_r} + O(1) \right) \frac{1}{r^2} dr + O(1). \end{aligned}$$

Observe that the first integral is non-negative, and it can be uniformly bounded at all loads by substituting $\rho_r \mapsto \mathbb{E}[S_r]/\mathbb{E}[S]$ and $\rho_{\bar{r}} \mapsto \mathbb{E}[S_{\bar{r}}]/\mathbb{E}[S]$, so it is $O(1)$.¹⁰ As for the second integral, note first that we can ignore the $O(1)$ since $\int_{r^*}^{\infty} \frac{O(1)}{r^2} dr = O(1)$. Moreover, by our choice of r^* , we have

$$\int_{r^*}^{\infty} \frac{\frac{\lambda}{2}(\text{Var}[S_r] + \text{Var}[A])}{r^2(1 - \rho_r)} \frac{1}{r^2} dr = \frac{\lambda}{2}(\text{Var}[S] + \text{Var}[A] + \delta) \int_{r^*}^{\infty} \frac{1}{r^2(1 - \rho_r)} dr,$$

where $\delta \in [-\epsilon, \epsilon]$. Therefore, we have that in the G/G/1 under SRPT, for some $\delta \in [-\epsilon, \epsilon]$,

$$\mathbb{E}[N]_{G/G/1}^{\text{SRPT}} = \frac{\lambda}{2}(\text{Var}[S] + \text{Var}[A] + \delta) \int_{r^*}^{\infty} \frac{1}{r^2(1 - \rho_r)} dr + O(1). \quad (8.15)$$

Of course, the M/G/1 is a special case of the G/G/1, so (8.15) also holds for the M/G/1, with $\text{Var}[A] = \mathbb{E}[A]^2$ and δ replaced by some other $\delta' \in [-\epsilon, \epsilon]$. Wierman et al. [284, Theorem 5.8] show $\mathbb{E}[N]_{M/G/1}^{\text{SRPT}} = \Omega(\log \frac{1}{1-\rho})$, so

$$\begin{aligned} \frac{\mathbb{E}[N]_{G/G/1}^{\text{SRPT}}}{\mathbb{E}[N]_{M/G/1}^{\text{SRPT}}} &= \frac{\text{Var}[S] + \text{Var}[A] + \delta}{\text{Var}[S] + \mathbb{E}[A]^2 + \delta'} + O\left(\frac{1}{\log \frac{1}{1-\rho}}\right) \\ &= \frac{\rho^2 c_S^2 + c_A^2 + \delta \lambda^2}{\rho^2 c_S^2 + 1 + \delta' \lambda^2} + O\left(\frac{1}{\log \frac{1}{1-\rho}}\right). \end{aligned}$$

The result follows because c_S^2 and c_A^2 are independent of ρ , and $\delta, \delta' \in [-\epsilon, \epsilon]$ for arbitrarily small ϵ .

¹⁰It is not a priori obvious that the integral converges due to the $r \rightarrow 0$ behavior. This can be verified by direct computation, but for our purposes, it suffices to use the prior knowledge that $\mathbb{E}[N]$ is finite under SRPT.

We have proven the result assuming $c_S^2 < \infty$. If instead $c_S^2 = \infty$, then $\text{Var}[S_r] \rightarrow \infty$ as $r \rightarrow \infty$. This means that for sufficiently large r , the dominant term of W_r is simply $\text{Var}[S_r]/(1 - \rho_r)$, which does not depend on A and is thus the same in the G/G/1 and M/G/1. One can use this fact to show that the performance ratio approaches 1 in heavy traffic by, as in the $c_S^2 < \infty$ case, splitting the WINE integral at large r^* , then observing that the $r > r^*$ part is dominant in heavy traffic. $\square$

Proof of Theorem 8.9. It suffices to show that the suboptimality gap in Theorem 8.7, which is $O(1)$ for the $k = 1$ case and $O(\frac{1}{1-\rho})$ for the $k \geq 2$ case, is dominated by $\mathbb{E}[N]^\pi$ for any policy π .

We begin by observing that for any G/G arrival process, $\mathbb{E}[N]_{G/G/1}^{\text{SRPT}}$ is a lower bound on $\mathbb{E}[N]^\pi$ for any policy π . This is because we can view the G/G/ k /setup as a version of a G/G/1 that imposes extra constraints on the scheduler (Section 8.3.3), and SRPT minimizes $\mathbb{E}[N]$ in the G/G/1 [241].

Next, we observe that by Theorem 8.28 and our assumption that $c_S^2 + c_A^2 > 0$, SRPT's G/G/1 heavy-traffic performance is within a constant factor of its M/G/1 heavy-traffic performance.

It thus suffices to show a lower bound on $\mathbb{E}[N]_{M/G/1}^{\text{SRPT}}$. When $k = 1$, we need only an $\omega(1)$ bound, which always holds [284, Theorem 5.8]. When $k \geq 2$, we need an $\omega(\log \frac{1}{1-\rho})$ bound, which prior work [245, Proof of Theorem 1.3 in Appendix B.2] shows to hold if $\mathbb{E}[S^2(\log S)^+] < \infty$. $\square$

8.10 Potential Extensions

We have seen that combining our new work decomposition law (Theorem 8.17) with WINE (Lemma 8.11) enables the analysis of systems with many complex features, such as the G/G/ k /setup. Thanks to the generality of both systems, we could apply the same technique even beyond the G/G/ k /setup. This section sketches how this can be done for three features: multiserver jobs, batch arrivals, and generalized vacations. We emphasize that our goal here is not to give full proofs, but rather to demonstrate the applicability of our technique to additional systems. We thus say “should be” rather than “is” when stating the end results.

8.10.1 Multiserver Jobs

We study a variation of the model of Grosf et al. [129]. We consider a variant of our $G/G/k$ model where each job has a *server need* $m(x)$, which is a function of its state x . Whenever a job in state x runs, it must occupy exactly $m(x)$ servers. It is thus served at rate $m(x)/k$, thanks to our convention of servers operating at rate $1/k$ (Section 8.3.1). We refer to this model as the $G/G/k/MSJ$, where MSJ stands for “multiserver job”. Grosf et al. [129] study what we would call the $M/G/k/MSJ$.¹¹

One of the main challenges when scheduling in MSJ systems is that it is no longer clear how to stabilize the system. Indeed, analyzing stability even in $M/M/k/MSJ$ systems is an area of current research [127, 238], and optimal scheduling in these systems is an open problem. However, Grosf et al. [129] show that if every possible server need $m(x)$ is a divisor of the number of servers k , then one can ensure stability with a procedure called *DivisorFilling*.

The *DivisorFilling* procedure takes as input any set of k jobs, then outputs a subset of those jobs whose server needs sum to exactly k . *DivisorFilling* can be combined with the Gittins policy by passing the k jobs of least Gittins rank to *DivisorFilling*, resulting in a policy called *DivisorFilling-Gittins* [129].

The $G/G/k/MSJ$ under *DivisorFilling-Gittins* can be analyzed in much the same way as the $G/G/k$ under Gittins. Recalling the structure of the latter analysis from Section 8.8, we encounter the same four “cost terms” (Definition 8.20) to bound.

- The residual interarrival cost m_{setup} can be bounded exactly as in Proposition 8.22.
- The recycling cost m_{rcy} can be bounded exactly as in Proposition 8.23, because *DivisorFilling-Gittins* ensures the key property that makes the proof work [129, Lemma B.5].
- The idleness cost m_{idle} can be bounded by following the same steps as [129, Lemma B.3], because their proof does not rely on Poisson arrivals. The resulting bound is $m_{\text{idle}} \leq e(k-1)\lceil \log \frac{1}{1-\rho} \rceil$.¹²

¹¹Grosf et al. [129] actually consider a slightly more restrictive case in which jobs’ server needs remain constant throughout service, though their proofs could be straightforwardly generalized to handle dynamically changing server needs. The main novelty of our discussion is thus the extension to G/G arrivals.

¹²In the $\rho \rightarrow 1$ limit, this bound below is actually slightly better than that in Proposition 8.24. See Section 8.4.1 for discussion.

- The setup cost m_{setup} is zero, as there are no setup times.

We thus see that the suboptimality gap of DivisorFilling-Gittins in the $G/G/k/\text{MSJ}$ should be at most

$$\mathbb{E}[N]_{G/G/k/\text{MSJ}}^{\text{DivisorFilling-Gittins}} - \inf_{\pi} \mathbb{E}[N]_{G/G/1}^{\pi} \leq e(k-1) \left\lceil \log \frac{1}{1-\rho} \right\rceil + \lambda(A_{\max} - A_{\min}).$$

This is analogous to what Theorem 8.7 says about the $G/G/k$, as $\ell_{(a)}$ is replaced by $e(k-1)\lceil \log \frac{1}{1-\rho} \rceil$.

It is likely that, for a suitable definition of setup times in an MSJ model, one could analyze the same system with setup times, obtaining a result analogous to the one we obtain for the $G/G/k/\text{setup}$.

8.10.2 Batch Arrivals

Scully and Harchol-Balter [243] introduce a general model of batch arrivals that we call *batch- M/G arrivals*. The main notable feature of the model is that it makes few assumptions about what batches look like. For example, it may be that the initial states of jobs in the same batch are correlated with each other. We consider the same type of batches but allow general batch interarrival times, resulting in *batch- G/G arrivals*. All of the results in Section 8.4 should generalize to batch- G/G arrivals. There are three changes needed for the proof, and only the third impacts the end results.

First, one needs to modify the definitions of load, r -fresh load, and other concepts related to the arrival process. But these are straightforward changes. The most important note here is that Assumption 8.2 should refer to the batch interarrival time.

Second, batch arrivals affect the work decomposition law (Theorem 8.17). But they only affect the term that is common to all systems, namely the numerator of the $1/(1-\rho_r)$ term. By Lemma 8.21, this change does not affect suboptimality gaps, which is the basis of all of the results in Section 8.4.

Third, batch arrivals affect the setup cost m_{setup} in multiserver systems. Specifically, we need to slightly modify the statement and proof of Proposition 8.26 to account for the fact that multiple jobs can arrive at once. If more jobs arrive than there are idle servers, this means a single setup is effectively triggered by multiple jobs. The end result is that one must incorporate a term related to the batch size distribution. In contrast, the setup cost in single-server systems is unaffected.

Taken together, these observations imply that our $G/G/k$ and $G/G/1/setup$ results should generalize immediately to batch arrivals. With some effort, the $G/G/k/setup$ results should also generalize.

8.10.3 Generalized Vacations

The term *generalized vacations* refers to a range of models where servers may be unavailable, including:

- Setup times, as studied in this work. This includes models beyond ours, e.g. where we make different decisions about when to start setting up a server, or where a setup time can be canceled if the job that triggered it enters service at another server.
- *Vacations*, where whenever the server goes idle, it goes on vacation for a given amount of time, then only begins serving jobs again when it returns.
- *Server breakdowns*, where servers can become unavailable in the middle of serving a job.
- *Threshold policies*, where servers stay idle until there are a given number of jobs in the system.

These are only a few examples of what generalized vacations can model [83, 96, 210].

One can, in principle, bound Gittins's suboptimality gap in the $G/G/k$ with generalized vacations using essentially the same approach we take for the $G/G/k/setup$. The main change is that we now interpret J_{setup} as the fraction of servers that are unavailable, so we now think of m_{setup} as an *unavailability cost*.

Of course, whether bounding m_{setup} is tractable to bound depends on the specifics of the model. As in the proofs of Propositions 8.25 and 8.26, the key question is: how many jobs might there be in the system while a server is unavailable? Sometimes, this will be very hard to bound, e.g. for server breakdowns. But in other cases, the bound is nearly immediate. For example, consider a threshold policy that does not start serving jobs until there are n jobs present, at which point it serves jobs until the system empties. We would then have $m_{\text{setup}} \in [0, n]$ under any scheduling policy.

One important application of generalized vacations is to more general setup time models. For instance, in practice, it is helpful to not turn servers off as soon as they become idle. One can imagine a wide range of *power*

management policies controlling when servers turn on and off. Provided we do not wait too long to set up servers while there are jobs in the queue, m_{setup} should not be too large, in which case Gittins would have a small suboptimality gap. This means that, in some sense, the power management and job scheduling problems are orthogonal, because a single scheduling policy, namely Gittins, performs well for a wide range of power management policies.

8.11 Conclusion

This work presents the first analysis of the Gittins policy in the $G/G/k/\text{setup}$. We prove simple and explicit bounds on Gittins's suboptimality gap, which are tight enough to imply that Gittins is optimal in heavy traffic in the $G/G/k/\text{setup}$. As a corollary, we find that Gittins is optimal in the $M/G/1/\text{setup}$. Prior to these results, Gittins had not been analyzed in even the $G/G/1$, let alone the $G/G/k/\text{setup}$.

There are several ways in which one might hope to improve our bounds. This is especially true in light traffic, namely the $\rho \rightarrow 0$ limit. Here we have a constant suboptimality gap for mean number of jobs, but by Little's law [185], this corresponds to an *infinite* suboptimality gap for mean response time. We conjecture that Gittins's mean response time suboptimality gap remains bounded in light traffic, but there are significant obstacles to doing so, related to the notorious problem of analyzing the idle period of the $G/G/1$ [179, 287].

Our theoretical results also raise several questions that could be studied with simulations. One such question is related to the additive structure of our suboptimality gap bound in Theorem 8.7, in which each of (a) multiple servers, (b) non-Poisson arrivals, and (c) setup times contributing to the bound via a separate term. If we simulate Gittins in systems with various mixtures of (a), (b), and (c), do we observe an analogous (approximate) additive structure in its empirical performance? We hypothesize the answer is “yes”, because each of the terms in Theorem 8.7 has a distinct cause, and we suspect the interactions between these causes are relatively weak. Investigating this is an interesting direction for future work.

Taking a step back, we might ask: should one use Gittins to minimize mean number of jobs in practice, even beyond the $G/G/k/\text{setup}$ modeling assumptions? While this is clearly a question larger than we can definitively

answer, we believe that our main results, the potential extensions sketched in Section 8.10, and other recent work on Gittins and SRPT in multiserver systems [125, 126, 129, 242, 245] point towards “yes”. Even though the currently known theoretical bounds on Gittins and SRPT are not tight, we have no comparable bounds for other policies, aside from a few close relatives of SRPT [125]. The mere existence of these bounds is thus a point in favor of Gittins. But we are still in the early years of understanding multiserver scheduling.

8.A Summary of Notation

Core $G/G/k$ /setup model (Sections 8.3.1 and 8.3.2)

- k : number of servers
 - Each server has speed $1/k$, so the total service capacity is always $k \cdot 1/k = 1$
- S : job size random variable
 - With k servers, job of size S results in kS service time
- U : setup work random variable (Section 8.3.2)
 - With k servers, setup work U results in kU setup time
- A : interarrival time random variable
- $\lambda = 1/\mathbb{E}[A]$: arrival rate
- $\rho = \lambda\mathbb{E}[S] < 1$: load
- N : number of jobs in the system
- $\mathbb{E}[N]_{\text{SYS}}^{\pi}$: mean number of jobs in system SYS under policy π
 - To reduce clutter, we omit “SYS” and/or “ π ” when there is no ambiguity

Remaining interarrival time (Assumption 8.2)

- A_{res} : residual interarrival time, i.e. time until next arrival
- A_{age} : interarrival age, i.e. time since last arrival
- $A_{\text{min}}, A_{\text{max}}$: bounds on $\mathbb{E}[A_{\text{res}} \mid A_{\text{age}}]$

Markov-process job model (Section 8.3.3)

- See Examples 8.3 and 8.4 for concrete examples of the Markov-process job model
- $\mathbb{X}$: space of possible job states
- $X(t)$: state of a job that has received t service so far
- X_i : state of the i th job currently in the system ($i \in \{1, \dots, N\}$)
- $\top$: terminal state, i.e. a job completes once $X(t) = \top$
- $S(x)$: remaining work of a job in state x
- $W = \sum_{i=1}^N S(X_i)$: (total) work in the system

Gittins policy (Section 8.3.4)

- Gtn: abbreviation for Gittins in formulas, e.g. $\mathbb{E}[N]^{\text{Gtn}}$
- $\text{rank}_{\text{Gtn}}(x)$: rank, i.e. priority (lower is better), of job in state x
 - See Examples 8.5 and 8.6 for concrete examples of the rank function

Main results (Section 8.4)

- $\ell_{(a)}, \ell_{(c)}, \ell_{(c)}$: suboptimality due to (a) multiple servers, (b) non-Poisson arrivals, (c) setup times
- $c_V^2 = \frac{\text{Var}[V]}{\mathbb{E}[V]^2}$: squared coefficient of variation of V

WINE and r -work (Section 8.6)

- $S_r(x)$: remaining r -work of a job in state x
- $W_r = \sum_{i=1}^N S_r(X_i)$: r -work in the system
- r -fresh, r -recycled: see Definitions 8.13 and 8.14
- $S_r = S_r(X(0))$: remaining r -work of a newly arrived job
- $\rho_r = \lambda \mathbb{E}[S_r]$: r -fresh load
- $\lambda_{r\text{-rcy}}$: average rate of r -recyclings
- $S_{r\text{-rcy}}$: remaining r -work of a newly r -recycled job
- $\rho_{r\text{-rcy}} = \lambda_{r\text{-rcy}} \mathbb{E}[S_{r\text{-rcy}}]$: r -recycled load
- J_r : fraction of servers that are serving jobs of rank at most r
- J_{setup} : fraction of servers that are setting up

Stating work decomposition law (Section 8.7.1)

- $\mathbb{E}_{\text{arv}}[\cdot]$: expectation sampled immediately before an arrival
- $\mathbb{E}_{r\text{-rcy}}[\cdot]$: expectation sampled immediately before an r -recycling
- $\mathbb{E}_{r\text{-acc}}[\cdot]$: an expectation defined in terms of $\mathbb{E}_{r\text{-rcy}}[\cdot]$ in (8.4)
- V_e : excess of random variable V
 - Key property: $\mathbb{E}[V_e] = \frac{\mathbb{E}[V^2]}{2\mathbb{E}[V]}$

Proofs of main results (Section 8.8)

- m_{res}^π : contribution to $\mathbb{E}[N]^\pi$ due to residual interarrival time, bounded in Proposition 8.22
- m_{rcy}^π : contribution to $\mathbb{E}[N]^\pi$ due to recyclings, bounded in Proposition 8.23
- m_{idle}^π : contribution to $\mathbb{E}[N]^\pi$ due to idle servers, bounded in Proposition 8.24
- m_{setup}^π : contribution to $\mathbb{E}[N]^\pi$ due to setup times, bounded in Propositions 8.25 and 8.26

8.B Additional Model Details

8.B.1 Processor Sharing

Thus far, we have discussed the scheduler as assigning jobs to servers in one-to-one fashion. However, because we are in a preemptive setting, it is possible for the scheduler to effectively share a server between multiple jobs by rapidly switching between them. As such, we explicitly allow the scheduler to use *processor sharing*, simultaneously serving multiple jobs at reduced rates. Processor sharing arises naturally when using Gittins [1, 2].

Exactly how processor sharing works with multiple servers and setup times has some subtle corner cases, so for completeness, we give the details below. But for ease of discussion, we do not concern ourselves with these details during the proofs, because they complicate the arguments without adding insight. With that said, modifying the proofs to account for processor sharing is a straightforward exercise. For instance, instead of referring to the job at server i , we might refer to an appropriately weighted random choice among jobs currently sharing server i .

In a system without setup times, the scheduler's action is a *service rate vector* $\alpha = (\alpha_1, \dots, \alpha_N)$, where recall that N is the number of jobs. Here α_i is the rate at which job i is being served. In the $G/G/k$, the fact that we have k servers of service rate $1/k$ corresponds to α obeying the constraints $\sum_{i=1}^N \alpha_i \leq 1$ and $\alpha_i \in [0, 1/k]$. From this perspective, a non-idling policy is one such that no α_i can be increased without violating a constraint. To clarify, the service rate vector α is a *gradual* control: it is chosen at every moment in time and can be changed at will.

In a system with setup times, the situation is similar, but we can only assign jobs to busy servers. The first constraint on α thus changes to $\sum_{i=1}^N \alpha_i \leq K_{\text{busy}}/k$, where K_{busy} is the number of busy servers.

8.B.2 General Definition of the Gittins Rank Function

How does Gittins decide what rank to assign each job? The main idea is to assign jobs low rank if they are likely to finish with a small amount of service. To formalize this intuition, consider a job in state x , and suppose we were to serve the job until it enters some state in set $\mathbb{Y} \subseteq \mathbb{X}$. This would involve serving the job for some amount of time, and there would be some chance the job completes. Let

$$\begin{aligned} S_{\mathbb{Y}}(x) &= \text{amount of service a job starting at } x \text{ needs to finish or enter } \mathbb{Y} \\ &= (\inf\{s \geq 0 \mid X(s) \in \mathbb{Y} \cup \{\top\}\} \mid X(0) = x), \end{aligned} \quad (8.16)$$

$$\begin{aligned} C_{\mathbb{Y}}(x) &= \text{event that a job starting at } x \text{ finishes before entering } \mathbb{Y} \\ &= [X(\inf\{s \geq 0 \mid X(s) \in \mathbb{Y} \cup \{\top\}\}) = \top \mid X(0) = x]. \end{aligned} \quad (8.17)$$

We can think of $\mathbb{E}[S_{\mathbb{Y}}(x)]/\mathbb{P}[C_{\mathbb{Y}}(x)]$ as a time-per-completion ratio for jobs in state x , as measured by serving them until they finish or enter $\mathbb{Y}$. The key idea of the Gittins policy is to let a job's rank be its best possible time-per-completion ratio:¹³

$$\text{rank}_{\text{Gtn}}(x) = \inf_{\mathbb{Y} \subseteq \mathbb{X}} \frac{\mathbb{E}[S_{\mathbb{Y}}(x)]}{\mathbb{P}[C_{\mathbb{Y}}(x)]}. \quad (8.18)$$

As an example, consider the case of known job sizes, where a job's state x is its remaining work. The best completion ratio is achieved by always running the job until it finishes, i.e. $\mathbb{Y} = \emptyset$. This results in $S_{\emptyset}(x) = x$ and the event

¹³We consider the ratio to be ∞ , and thus not the infimum, if $\mathbb{P}[C_{\mathbb{Y}}(x)] = 0$.

$C_\emptyset(x)$ occurring with probability 1, so $\text{rank}_{\text{Gtn}}(x) = x$. That is, Gittins always serves the jobs of least remaining work, so it reduces to SRPT (Example 8.5).

8.B.3 Technical Considerations for the Job Markov Process

We have omitted some technical details and assumptions in our definition of the Gittins policy. For example, in (8.16)–(8.18), we must restrict attention to sets $\mathbb{Y} \subseteq \mathbb{X}$ such that $S_{\mathbb{Y}}(x)$ and $C_{\mathbb{Y}}(x)$ are measurable with respect to the natural filtration on the job Markov process. More generally, the theory of the Gittins policy relies on being able to solve an optimal stopping problem which is known in the Markov-process job model literature as the “Gittins game” [242, 243, 245].

Following the convention of prior literature, we consider the technical foundations of the Gittins game to be outside the scope of this chapter. Our results apply to any job Markov process where the foundations can be established. As explained by Scully et al. [245, Appendix D], this is not a restrictive assumption.

8.C Deferred Proofs

Proposition 8.24 (Idleness Cost, restated). *In the $G/G/k$ /setup, under the Gittins policy, we have*

$$m_{\text{idle}}^{\text{Gtn}} \leq (C - 1)(k - 1) \log \frac{1}{1 - \rho} + (k - 1) \mathbf{1}_{\{\mathbb{P}[U > 0] > 0\}},$$

where $C = \frac{9}{8 \log 1.5} + 1 \approx 3.775$.

Proof. At a high level, we follow the same main as the proof of [242, Proposition 17.6], but one of the steps requires modification to account for setup times.

By following the same argument as [242, Lemma 17.4], one can show that

$$m_{\text{idle}}^{\text{Gtn}} = \int_0^\infty \frac{\mathbb{E}[(1 - J_r - J_{\text{setup}})W_r]}{r^2(1 - \rho_r)} dr \leq (k - 1) \inf_{n \in \mathbb{Z}_{\geq 1}} n \left(\frac{1}{1 - \rho} \right)^{1/n}. \quad (8.19)$$

The key fact is that when $1 - J_r - J_{\text{setup}} > 0$, there is an idle server, so there are at most $k - 1$ jobs in the system, as otherwise a server would start setting up. Given this fact, the rest of the proof carries through verbatim. We can also show an alternative bound for $m_{\text{idle}}^{\text{Gtn}}$ that is tighter in light traffic:

$$\begin{aligned} m_{\text{idle}}^{\text{Gtn}} &= \int_0^\infty \frac{\mathbb{E}[(1 - J_r - J_{\text{setup}})W_r]}{r^2(1 - \rho_r)} dr \\ &\leq (k - 1) \frac{\rho}{1 - \rho} + (k - 1) \mathbb{1}\{\mathbb{P}[U > 0] > 0\}. \end{aligned} \quad (8.20)$$

We prove (8.20) below. Combining (8.19) and (8.20) yields

$$\begin{aligned} m_{\text{idle}}^{\text{Gtn}} &\leq (k - 1) \min \left\{ \inf_{n \in \mathbb{Z}_{\geq 1}} n \left(\frac{1}{1 - \rho} \right)^{1/n}, \frac{\rho}{1 - \rho} + \mathbb{1}\{\mathbb{P}[U > 0] > 0\} \right\} \\ &\leq (k - 1) \min \left\{ \inf_{n \in \mathbb{Z}_{\geq 1}} n \left(\frac{1}{1 - \rho} \right)^{1/n}, \frac{\rho}{1 - \rho} \right\} + (k - 1) \mathbb{1}\{\mathbb{P}[U > 0] > 0\} \\ &\leq (C - 1)(k - 1) \log \frac{1}{1 - \rho} + (k - 1) \mathbb{1}\{\mathbb{P}[U > 0] > 0\}, \end{aligned}$$

where the last inequality follows from the same computation as in [242, Proposition 17.6].

It remains only to prove (8.20). The proof follows the same steps as [242, Lemma 17.5], but setup times add some additional considerations. For the purposes of this proof, we abuse notation slightly by defining $X_i = \top$ for the state of job i if there is no i th job. We then define the remaining r -work of $\top$ to be $S_r(\top) = 0$.

We begin by noting that if $J_r + J_{\text{setup}} < 1$, then there are less than k jobs in the system, so

$$\begin{aligned} \mathbb{E}[(1 - J_r - J_{\text{setup}})W_r] &= \sum_{i=1}^k \mathbb{E}[S_r(X_i)(1 - J_r - J_{\text{setup}})] \\ &\leq \frac{k - 1}{k} \sum_{i=1}^k \mathbb{E}[S_r(X_i)], \end{aligned} \quad (8.21)$$

where the last equality follows from the fact that $J_r + J_{\text{setup}} < 1$ is an integer multiple of $1/k$. Applying Lemma 8.12 and (8.21) to the definition of $m_{\text{idle}}^{\text{Gtn}}$ yields

$$m_{\text{idle}}^{\text{Gtn}} = \int_0^\infty \frac{\mathbb{E}[(1 - J_r - J_{\text{setup}})W_r]}{r^2(1 - \rho_r)} dr$$

$$\begin{aligned}
&\leq \frac{k-1}{k(1-\rho)} \int_0^\infty \sum_{i=1}^k \frac{\mathbb{E}[S_r(X_i)]}{r^2} dr \\
&\leq \frac{k-1}{k(1-\rho)} \sum_{i=1}^k \mathbb{P}[X_i \neq \top] \\
&= \frac{(k-1)\mathbb{E}[\min\{N, k\}]}{k(1-\rho)}. \tag{8.22}
\end{aligned}$$

The fact that we are considering the setup-non-idling version of Gittins (Section 8.3.2) implies that the number of non-idle servers (i.e. busy or setting-up) is always at least $\min\{N, k\}$. So it suffices to bound the mean fraction of non-idle servers.

- The mean fraction of busy servers is ρ .
- The mean fraction of setting-up servers is at most $(1-\rho)\mathbb{1}\{\mathbb{P}[U > 0] > 0\}$.
 - If there are no setup times, it is clearly zero.
 - If there are setup times, it is at most the fraction of non-busy servers, namely $1-\rho$.

This means

$$\frac{(k-1)\mathbb{E}[\min\{N, k\}]}{k(1-\rho)} \leq (k-1)\frac{\rho}{1-\rho} + (k-1)\mathbb{1}\{\mathbb{P}[U > 0] > 0\}. \tag{8.23}$$

Combining (8.22) and (8.23) yields (8.20), as desired. $\square$

Lemma 8.27 (restated). *In the $G/G/k$ /setup, for any server i and all $a \geq 0$, we have*

$$\mathbb{E}[N \mid J_{\text{setup},i} = 1, U_{\text{age},i} = a] \leq \lambda a + \lambda A_{\max} + k - 1.$$

Proof. Observe that during setup, the number of jobs in the system can be divided into three sets: jobs present when setup started, jobs triggered the setup and jobs that have arrived since setup started. Therefore, to bound $\mathbb{E}[N \mid J_{\text{setup},i} = 1, U_{\text{age},i} = a]$, it is sufficient to bound the total number of jobs in the three sets on the right. Just before the setup starts, there is at least one idle server, so there are at most $k-1$ jobs present. And there is exactly

1 job that triggers the setup. Let $N_{\text{arv}}(a)$ be the number of jobs that have arrived since a unit of time ago, then the above argument shows that

$$\mathbb{E}[N \mid J_{\text{setup},i} = 1, U_{\text{age},i} = a] \leq \mathbb{E}[N_{\text{arv}}(a) \mid J_{\text{setup},i} = 1, U_{\text{age},i} = a] + k.$$

Therefore, to prove the lemma, it is sufficient to show that for any $a \geq 0$,

$$\mathbb{E}[N_{\text{arv}}(a) \mid J_{\text{setup},i} = 1, U_{\text{age},i} = a] \leq \lambda a + \lambda A_{\text{max}} - 1. \quad (8.24)$$

Conditioned on $J_{\text{setup},i} = 1$ and $U_{\text{age},i} = a$, we know that server i started setup process a unit of time ago. Since the setup process is triggered by an arrival, if we denote the sequence of interarrival times since the setup as $A_1, A_2, \dots, A_n, \dots$, then we have the equation

$$\mathbb{E} \left[\sum_{n=1}^{N_{\text{arv}}(a)+1} A_n \mid J_{\text{setup},i} = 1, U_{\text{age},i} = a \right] = \mathbb{E}[a + A_{\text{res}} \mid J_{\text{setup},i} = 1, U_{\text{age},i} = a]. \quad (8.25)$$

The right-hand side of (8.25) can be upper bounded using Assumption 8.2 and the fact that A_{res} is conditionally independent of the past given the age:

$$\mathbb{E}[a + A_{\text{res}} \mid J_{\text{setup},i} = 1, U_{\text{age},i} = a] = \mathbb{E}[a + A_{\text{res}} \mid U_{\text{age},i} = a] \leq a + A_{\text{max}}. \quad (8.26)$$

To bound the left-hand side of (8.25), observe that $J_{\text{setup},i} = 1$ and $U_{\text{age},i} = a$ is equivalent to having a setup process start a unit of time ago and last until now, which is determined by the setup time and is independent of arrival process in the recent a unit of time, so we have

$$\mathbb{E} \left[\sum_{n=1}^{N_{\text{arv}}(a)+1} A_n \mid J_{\text{setup},i} = 1, U_{\text{age},i} = a \right] = \mathbb{E} \left[\sum_{n=1}^{N_{\text{arv}}(a)+1} A_n \right]. \quad (8.27)$$

We further observe that $A_1, A_2, \dots, A_n, \dots$ are a sequence of i.i.d. variables, and $N_{\text{arv}} + 1$ is the smallest n such that the partial sum $A_1 + \dots + A_n \geq a$, by Wald's equation,

$$\mathbb{E} \left[\sum_{n=1}^{N_{\text{arv}}(a)+1} A_n \right] = \mathbb{E}[N_{\text{arv}}(a) + 1] \mathbb{E}[A] = \frac{\mathbb{E}[N_{\text{arv}}(a) + 1]}{\lambda}. \quad (8.28)$$

Combining (8.25)–(8.28), we get $\mathbb{E}[N_{\text{arv}}(a) + 1]/\lambda \leq a + A_{\text{max}}$, implying (8.24). $\square$

8.D Towards Stability of the $G/G/k$ /setup under Work-Conserving Policies

In this section, we sketch some initial ideas that we believe could be used to prove the stability of the $G/G/k$ /setup under Gittins and other complex scheduling policies. For simplicity, we focus on the $G/G/k$ under stationary scheduling policies.

8.D.1 Partial Proof Sketch

Assume we are using some scheduling policy that makes the system a Markov process. Consider the embedded Markov process of our system at arrival instants. We construct a Lyapunov function and show its negative drift for this embedded Markov process when the total work in the system or the largest job size is outside a bounded region. The negative drift of the Lyapunov function should imply the stability of the system after applying a certain version of the Foster-Lyapunov theorem, although there are still some technical obstacles, as we discuss in the next section.

At any time $t \geq 0$, we let $W(t)$ be the total work in the system, let $S(t)$ be the size of the largest job in the system, and let $A_{\text{res}}(t)$ be the residual arrival time.

Fix large constants $\alpha > 0$ and $\beta > 0$. Define the Lyapunov function $V = W - \rho A_{\text{res}} + k(S - \alpha)^+$. Note that at each arrival point, $A_{\text{res}} = 0$ and $V \geq 0$. The continuous rate of change of V is

$$\begin{aligned} \dot{V} &= -(1 - \rho) + 1 - J - k\mathbf{1}\{S > \alpha\}\dot{S}_{\text{big}}(t) \\ &\leq -(1 - \rho) + 1 - J - \mathbf{1}\{S > \alpha, J < 1\} \\ &\leq -(1 - \rho) + (1 - J)\mathbf{1}\{S \leq \alpha\}, \end{aligned}$$

where in the first inequality, we use the fact that when $J < 1$, all jobs are in service, so the size of the largest job decreases at a rate $1/k$; and in the second inequality, we use the fact that $\mathbf{1}\{J < 1\} \geq 1 - J$. The expected jump of V at an arrival point is

$$\begin{aligned} \Delta V &= \mathbb{E}[S - \rho A + k \max((S - \alpha)^+, (S - \alpha)^+) - k(S - \alpha)^+] \\ &\leq k\mathbb{E}[(S - \alpha)^+], \end{aligned}$$

where we have used the fact that $\mathbb{E}[S] = \rho\mathbb{E}[A]$. Consider the change of V between two consecutive arrivals:

$$\begin{aligned} & \mathbb{E}[V(A)] - V(0) \\ &= \int_0^A \dot{V}(t) dt + \lambda\Delta V \\ &\leq \mathbb{E}\left[\int_0^A (-(1-\rho) + (1-J(t))\mathbb{1}\{S(t) \leq \alpha\}) dt\right] + \lambda k\mathbb{E}[(S-\alpha)^+] \\ &= -\lambda(1-\rho) + \mathbb{E}\left[\int_0^A (1-J(t))\mathbb{1}\{S(t) \leq \alpha\} dt\right] + \lambda k\mathbb{E}[(S-\alpha)^+]. \end{aligned}$$

Let $\mathcal{R} = [0, k(\alpha + \beta) + \beta] \times [0, \alpha + \beta]$, a compact region. We show that $\mathbb{E}[V(A)] - V(0) < 0$ as long as $(W(0), S(0)) \notin \mathcal{R}$. There are two cases to consider:

- If $S(0) > \alpha + \beta$, then $S(t) > \alpha + \beta - t$. As a result,

$$\mathbb{E}[V(A)] - V(0) \leq -\lambda(1-\rho) + \mathbb{E}[(A-\beta)^+] + \lambda k\mathbb{E}[(S-\alpha)^+].$$

- If $S(0) \leq \alpha + \beta$ but $W(0) > k(\alpha + \beta) + \beta$, then we have $S(t) \leq \alpha + \beta$, and $W(t) \geq k(\alpha + \beta) + \beta - t$. For $t \leq \beta$, $N(t) \geq W(t)/S \geq k$, implying $J(t) = 1$. As a result,

$$\mathbb{E}[V(A)] - V(0) \leq -\lambda(1-\rho) + \mathbb{E}[(A-\beta)^+] + \lambda k\mathbb{E}[(S-\alpha)^+].$$

Taking sufficiently large α and β , we get $\mathbb{E}[V(A)] - V(0) < 0$ in both cases.

8.D.2 Remaining Obstacles

We have given a Lyapunov function that has negative drift for the embedded Markov process when (W, S) is outside a compact region $\mathcal{R}$. However, (W, S) is not a complete state description of the Markov process. Moreover, the set of states such that $(W, S) \in \mathcal{R}$ might not be compact. We believe one can show that this set of states is *petite* [206], even if it is not compact, which will suffice. However, this seems to require reasoning about the details of the job state space.

Of course, so far we have just been focusing on the existence of a stationary distribution for the embedded process at arrival instants. Once this is established, we need to use it to construct a stationary distribution for the continuous-time process, but this step is more standard.

Chapter 9

Sharp Waiting-Time Bounds for Multiserver Jobs

9.1 Introduction

In today’s large-scale computing clusters behind cloud platforms, *multiserver jobs* have become increasingly prevalent, where a multiserver job is a job that demands to occupy multiple “servers” (which can be multiple physical servers, multiple CPU cores, etc.) simultaneously during its runtime [4, 184, 258, 267]. For example, cloud platforms allow users to specify the number of CPU cores in their virtual machines or containers, and this information can be utilized by centralized schedulers to make scheduling decisions (see, e.g., Verma et al. [267], Google Kubernetes Engine [122]). Moreover, the number of “servers” that a multiserver job requests, which we refer to as the *server need*, is becoming increasingly large. This trend is driven by machine learning jobs from applications like TensorFlow in Abadi et al. [4], where the jobs are highly parallel and require synchronization. According to the statistics from Google’s Borg Scheduler in Verma et al. [267], the server needs in Borg can vary across six orders of magnitudes.

In this chapter, we study the impact of multiserver jobs on the delay performance of large-scale computing systems using queueing models. Queueing models with multiserver jobs have been studied in the literature, but quantifying the delay performance is notoriously hard. Exact steady-state distributions can only be derived in highly simplified settings with two servers [55, 91], while the majority of prior work has focused on characterizing sta-

bility conditions [5, 127, 214, 219, 238]. However, even for stability, exact conditions are known only for the special cases where all jobs have the same service rate or where there are two job classes. We comment that concurrent to the conference version of our work [147], Grosz et al. [128] and Grosz et al. [129] study the delay performance of multiserver jobs in the traditional heavy-traffic regime. A more detailed review of the related work is provided in Section 9.2.

A recent advance in understanding the delay of multiserver jobs is a characterization of the *queueing probability* in a *large system* by Wang et al. [271], where the queueing probability is the probability that an arriving job has to queue rather than entering service immediately. Specifically, Wang et al. [271] consider a multiserver job system with n servers, and study the asymptotic scaling regimes where n becomes large. The scaling regimes allow different job types to have different arrival rates, server needs and service rates. Among those parameters, server needs and arrival rates can scale up with n . Such scaling regimes capture the trend that different multiserver jobs can be highly heterogeneous, especially in terms of server needs. They establish an upper bound on the queueing probability, based on which they give a sufficient condition for the queueing probability to diminish as n goes to infinity.

Although the work of Wang et al. [271] identifies when the queueing probability diminishes in large systems, which is a much desirable operating scenario, it does not provide much insight for differentiating between scheduling policies. In particular, their queueing probability upper bound holds for any scheduling policy that is reasonably work-conserving (although the bound is presented only for the First-Come-First-Served policy). Moreover, queueing probability does not directly translate to delay of jobs.

In this chapter, we focus on the *waiting time* of jobs, which is the total time a job spends waiting in the queue (not receiving any service), under various scheduling policies. The waiting time is a performance metric that is directly related to job delay. Our goal is to establish bounds on the mean waiting time that are *order-wise tight* as the number of servers, n , scales. Such tight bounds will enable us to differentiate between policies based on their delay performance. We comment that there has been a line of work in the literature [186–189, 261, 276, 277] that focuses on quantifying *when* the mean waiting time diminishes in large systems for various queueing models. However, little is known on *how fast* the mean waiting time diminishes due to the lack of lower bounds. Our results provide the rate of diminishing when

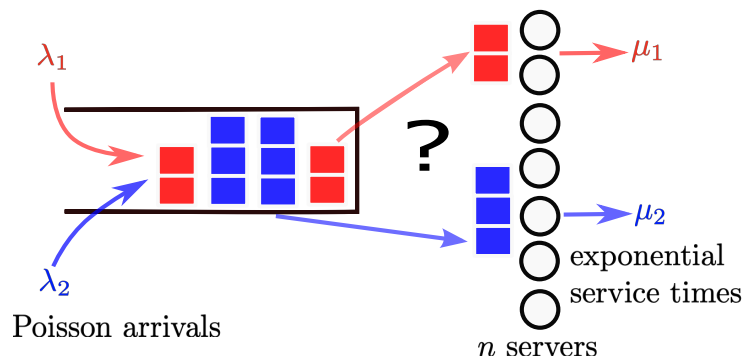


Figure 9.1: A multiserver-job system with two types of jobs. Type 1 jobs have arrival rate λ_1 , service rate μ_1 , and server need $\ell_1 = 2$. Type 2 jobs have arrival rate λ_2 , service rate μ_2 , and server need $\ell_2 = 3$.

the mean waiting time does diminish, but our tight bounds on the mean waiting time are not limited to the “diminishing” scenario.

Since the First-Come-First-Served (FCFS) policy is widely used as a default policy in practice and also receives the most attention from theoretical studies of multiserver jobs [5, 55, 91, 127, 214, 238], in this chapter, we will first examine FCFS and understand the exact order of the mean waiting time under it. Then a natural question that arises is: *can any policy outperform FCFS in terms of the mean waiting time?* More generally, we aim to answer the following fundamental questions:

- *What is the optimal order of the mean waiting time as the system scales?*
- *Which policy achieves the optimal order?*

9.1.1 Model and Performance Metric

We consider a system that consists of n servers and I types of jobs. An example is illustrated in Figure 9.1. Suppose type i jobs need the simultaneous service of ℓ_i servers. We sort the job types such that their *server needs* ℓ_i 's satisfy $\ell_1 \leq \ell_2 \leq \dots \leq \ell_I$. Let the *maximal server need* $\ell_{\max}$ be $\ell_{\max} = \max_{i \in \{1, 2, \dots, I\}} \ell_i = \ell_I$, and we call type I jobs the *maximal-need jobs*.

The dynamics of the system are as follows. For each $i = 1, 2, \dots, I$, type i jobs arrive to the system following a Poisson process with *arrival rate* λ_i . Upon arrival, a job either starts service immediately or waits in a centralized

queue. When a type i job starts service, it leaves the queue and makes exclusive use of ℓ_i servers. The job leaves the system after receiving enough service. The service time of a type i job follows an exponential distribution with *service rate* μ_i . The service times and arrival events are independent. During the operation of the system, a scheduling policy is used to determine which set of jobs to serve at any time. The scheduling policy is allowed to be preemptive, i.e., we can put a job in service back to the queue and resume its service later.

We measure the performance of our scheduling policy based on *mean waiting time* as defined below: let $T_i^w(\infty)$ denote the waiting time of type i jobs in the steady state, then the mean waiting time is defined as the steady-state expected waiting time averaged over all job types, i.e.,

$$\mathbb{E}[T^w(\infty)] = \frac{1}{\lambda} \sum_{i=1}^I \lambda_i \mathbb{E}[T_i^w(\infty)],$$

where $\lambda \triangleq \sum_{i=1}^I \lambda_i$ is the total arrival rate.

9.1.2 Scaling Regimes

We study job delay in scaling regimes where the number of servers, n , goes to infinity. Specifically, we consider a sequence of systems with parameters scaling up jointly with n , and analyze the growth/decrease rate of the mean waiting time. In the considered scaling regimes, the arrival rates λ_i and server needs ℓ_i are allowed to scale with n , while the service rate μ_i and the number of job types I stay constant. One key parameter for specifying a scaling regime is the *slack capacity* δ , defined as $\delta = n - \sum_{i=1}^I \frac{\lambda_i \ell_i}{\mu_i}$, which is the expected number of idle servers in the steady state. Slack capacity is used to specify the heaviness of traffic, which is alternatively specified by *load* ρ given by $\rho = \sum_{i=1}^I \frac{\lambda_i \ell_i}{n \mu_i}$ in literature.

For expositional purposes, we now parameterize the scaling regimes in a specific way below to demonstrate our results. Our general model is presented in Section 9.3. Suppose $\ell_{\max} = n^\gamma$, $\delta = n^\alpha$ for some exponents $0 \leq \alpha, \gamma < 1$, and the total arrival rate $\lambda = \Theta(n)$. We refer to such scaling regimes as *polynomial scaling regimes*. To classify and visualize different polynomial scaling regimes, we aggregate all scaling regimes with the same (α, γ) pair into one point and plot all such points, as shown in Figure 9.2. We partition the set of exponent pairs $(\alpha, \gamma) \in [0, 1]^2$ into three triangles, using the lines

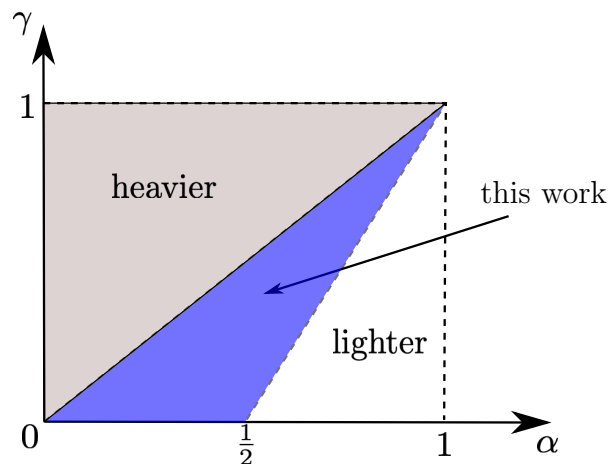


Figure 9.2: Scaling regimes under the special parameterization with slack capacity $\delta = n^\alpha$ and maximum server need $\ell_{\max} = n^\gamma$. The traffic is heavier as we move to the upper left. The three triangles are partitioned by lines $\alpha = \gamma$ and $\alpha = \frac{1+\gamma}{2}$.

$\alpha = \gamma$ and $\alpha = \frac{1+\gamma}{2}$. The corresponding scaling regimes to the upper left are in general “heavier” than the scaling regimes to the lower right, since the former regimes have larger work variability and smaller slack capacity. We comment that the point $(\alpha, \gamma) = (\frac{1}{2}, 0)$ is analogous to the celebrated Halfin-Whitt regime introduced in Halfin and Whitt [135] and $(\alpha, \gamma) = (0, 0)$ is analogous to the Non-Degenerate Slowdown (NDS) regime in Atar [19] in traditional multiclass M/M/n models.

In the polynomial scaling regimes where $\frac{1+\gamma}{2} < \alpha < 1$ (indicated in white in Figure 9.2), Wang et al. [271] proved that the queueing probability diminishes under FCFS when n gets large. In this chapter, we further show that in these regimes, both the queueing probability and the mean waiting time diminish at a fast rate as n get large, under a large class of policies.

While all reasonable policies perform well when $\frac{1+\gamma}{2} < \alpha < 1$, as we move on to polynomial scaling regimes with $\gamma < \alpha < \frac{1+\gamma}{2}$ (marked in blue in Figure 9.2), different policies begin to show separations in their mean waiting times, as presented in our main results below.

9.1.3 Results

Here we present our main results specialized to the polynomial scaling regimes. These results hold under several assumptions: $\gamma < \alpha < \frac{\gamma+1}{2}$, $\lambda = \Theta(n)$, and $\lambda_I \ell_I = \Theta(n)$.¹ In Section 9.4, we will present the general results and comment on how they imply the results here.

- **Mean waiting time under FCFS.** The exact order of the mean waiting time under FCFS is given by

$$\mathbb{E}[T^w(\infty)]^{\text{FCFS}} = \Theta(n^{\gamma-\alpha}). \quad (9.1)$$

- **Mean waiting time lower bound.** Under any policy, the mean waiting time satisfies

$$\mathbb{E}[T^w(\infty)] = \Omega(n^{-\alpha}), \quad (9.2)$$

- **Order-wise optimal policy.** Consider a static priority policy that we call the *Smallest-Need-First (SNF) policy*, which preemptively prioritizes the jobs with smaller server needs. Then the mean waiting time under SNF achieves the lower bound in (9.2), i.e.,

$$\mathbb{E}[T^w(\infty)]^{\text{SNF}} = \Theta(n^{-\alpha}). \quad (9.3)$$

Therefore, the SNF policy is order-wise optimal in the mean waiting time.

Comparing the mean waiting times under FCFS and under SNF, we can see that FCFS is strictly suboptimal when $\gamma > 0$, and SNF improves upon FCFS by a factor of $\Theta(n^\gamma)$.

A key to proving the mean waiting time results above is the *order-wise tight* bounds on the expected workload we establish (Lemma 9.7 and Lemma 9.9). In addition, although we consider the system under the traffic regime where $\gamma < \alpha < \frac{1+\gamma}{2}$, we still need to analyze “subsystems” that are in the lighter traffic regime. In this lighter regime, we show that the total server need decays faster than any polynomial (Lemma 9.17). All these lemmas hold under a very general class of policies, so they could be potentially relevant when we study policies other than FCFS and SNF.

¹We use the standard Bachmann–Landau notation. Consider two sequences $a(n)$ and $b(n)$ (or simply a and b), where $b(n)$ is positive for large enough n . Then $a = O(b)$ if $\limsup_{n \rightarrow \infty} \frac{|a|}{b} < \infty$; $a = o(b)$ if $\lim_{n \rightarrow \infty} \frac{a}{b} = 0$; $a = \Omega(b)$ if $\liminf_{n \rightarrow \infty} \frac{a}{b} > 0$, which is equivalent to $b = O(a)$; $a = \omega(b)$ if $\lim_{n \rightarrow \infty} \frac{|a|}{b} = \infty$, which is equivalent to $b = o(a)$; $a = \Theta(b)$ if a satisfies both $a = O(b)$ and $a = \Omega(b)$.

Results on Queueing Probability.

As a by-product to our analysis, we further derive an upper bound on the queueing probability under any work-conserving policy, presented in Corollary 9.18, which significantly improves upon the queueing probability upper bound in Wang et al. [271] in a slightly more constrained traffic regime.

Simulation Experiments.

The SNF policy we consider in our analysis is a *preemptive* priority policy. Preemption is sometimes undesirable in practice. Therefore, we use simulation experiments to also explore a *non-preemptive* version of the priority policy, which we call the non-preemptive Smallest-Need-First (SNF-NP) policy. SNF-NP serves a job with the smallest server need in the queue when enough number of servers free up. Our simulation experiments compare the mean waiting times under FCFS, SNF, and SNF-NP, along with a few other policies. The simulation results, presented in Section 9.10, show that SNF-NP has comparable performance with SNF, and both of them perform well relative to FCFS and other policies. We also investigate the weighted mean waiting times where jobs with larger server needs have larger weights. We find that SNF and SNF-NP remain competitive when the weights moderately depend on the server needs.

9.1.4 Technical Challenges

The main technical challenges in analyzing the considered multi-server-job system are rooted in the *heterogeneity* among job types in both their service rates and their server needs. Such heterogeneity makes the system dynamics multidimensional: neither the total number of jobs in service nor the total number of busy servers determines the current job departure rate. We comment that even for a classical multi-class M/M/n system, where there are multiple job types with different service rates but all job types have a server need of 1, finding an optimal scheduling policy is known to be a hard problem, and solutions are available mostly in the so-called Halfin-Whitt heavy-traffic regime through the diffusion control problem [18, 21, 141]. Compared with the classical multi-class M/M/n system, our multiserver-job system has an additional layer of intricacy due to the heterogeneous server needs, which

makes it possible for the system to have servers idling while there are jobs waiting in the queue.

To address the challenges due to heterogeneity, our analysis relies on various state-space concentration results. State-space concentration is a phenomenon where the state concentrates around a *subset* of the state space in steady state, observed in queueing systems in heavy-traffic or large-system regimes [186–189, 272, 277]. In the multiserver-job system we consider, state-space concentration results are crucial for analyzing the system dynamics when the queue is nonempty. The scenario when the queue is nonempty is especially important to our scaling regimes since the queueing probability may not be diminishing even when the mean waiting time is diminishing. This contrasts with the analysis in the prior work of Wang et al. [271], which focuses on diminishing queueing probability. Furthermore, our performance goal is to achieve the *optimal order* of the mean waiting time in large systems, deviating from the traditional performance goal of minimizing delay or certain long-run cost.

9.1.5 Organization of the Chapter

This chapter is organized as follows. In Section 9.2, we review some additional related work that has not been discussed in the introduction. We present our model and assumptions in Section 9.3 and then formally state our three main theorems in Section 9.4. In Section 9.5, we give an overview of the proof structure and preliminaries of our main proof technique, the drift method. In Section 9.6, we state and prove Lemmas 9.7 and 9.9, which will be used to prove the three theorems in Sections 9.7, 9.8 and 9.9. In Section 9.10, we present the simulation results. Finally, we conclude the chapter in Section 9.11.

9.2 Related Work

In this section, we give a more detailed review of the prior work on multiserver-job models as well as some related models that are not covered in the introduction.

Multiserver-Job Model.

As mentioned in the introduction, the majority of prior work on the multiserver-job model has either focused on characterizing stability conditions [5, 127, 214, 219, 238], or have been restricted to the highly specialized settings with two servers [55, 91]. There are relatively fewer papers that study the delay of the multiserver-job model, with different performance metrics. [271] characterizes the queueing probability in a large multiserver-job model. [306] studies the optimal cumulative holding cost in a finite-horizon setting. The papers whose performance metrics are closest to ours are [128] and [129]. Grosz et al. [128] characterizes the mean response time in a multiserver-job model under a policy called ServerFilling. Grosz et al. [129] proposes a variant of ServerFilling called ServerFilling-SRPT, which optimizes the mean response time in the traditional heavy traffic regime. The biggest distinction between their work and our work lies in the scaling regimes: in their work, the analysis of mean response time is asymptotically tight when the load of the system goes to one and the number of servers remains *fixed*; in contrast, we consider the scaling regimes where the load, number of servers and server needs *scale jointly*. Another distinction is the assumptions on the server needs: their work assumes that the server needs can divide the total number of servers, while our work assumes that the maximal server need is small compared with the slack capacity.

Virtual Machine (VM) Scheduling.

A problem related to the multiserver-job scheduling problem studied in this chapter is the virtual machine (VM) scheduling problem (see, e.g., [194, 196, 197, 225, 226, 256, 291]). For the VM scheduling problem, typically the system consists of multiple servers, where each server has certain units of each type of resource (e.g., CPU, memory, storage). A VM job demands to occupy multiple units of each type of resource. Each VM job will be served on a single server. Some results for the VM scheduling problem in the traditional heavy-traffic regime can be specialized to the multiserver job scheduling problem. To see this, consider a VM scheduling problem where the system consists of a single server and there is a single resource type. Then each unit of resource can be viewed as a server in the multiserver-job scheduling problem. With this specialization, the results in [197] provide bounds on a linear combination of the queue lengths of different types of jobs.

However, these bounds do not directly translate into bounds on mean job response time.

Multitask-Job Model.

A multitask job is a job that consists of a batch of tasks that can run on servers in parallel, which is similar to a multiserver job in that both can occupy multiple servers at the same time. However, unlike a multiserver job, the tasks of a multitask job can have different runtimes and do not need to be executed simultaneously. Multitask job model has been considered under a wide variety of settings, and is sometimes referred to as batch arrival model [see, e.g., 77–79, 208]. Recently, multitask job model is also extensively studied under the setting of parallel computing due to the popularity of large-scale data processing systems such as MapReduce, Apache Hadoop and Apache Spark. [see, e.g., 276, 305]. The work closest to our work is Weng and Wang [276], which shows diminishing queueing time for multi-server jobs in a load-balancing system where tasks of a job need to be dispatched to the queues at the servers upon arrival.

Dropping Model.

When the multiserver-job system does not have any queueing space and allows incoming jobs to be dropped, it becomes a model that has been extensively studied in the literature and we refer to it as the *dropping model*. In this model, one can design a dropping policy that decides whether to drop an incoming job or not based on the types of the incoming job and of the jobs currently in service. Under the policy that drops an incoming job only when it cannot fit into the servers, i.e., when its server need is larger than the number of available servers, the stationary distribution has a product form under exponentially distributed service times, as observed by Arthurs and Kaufman [15]. The results have been generalized by Whitt [279] to allow jobs to demand multiple resource types (e.g., both CPU and I/O) and by van Dijk [262] to allow general service time distributions. Tikhonenko [257] further combined aspects of Whitt [279] and van Dijk [262]. Different dropping policies, which mostly fall within the class of trunk reservation policies, have been designed to minimize the cost associated with dropping [29, 156, 157].

Streaming Model.

The *streaming model* for a communication network resembles the multiserver-job model in many aspects. In a streaming model, the “servers” correspond to the bandwidth in the network and the “jobs” are data flows such as audio or video flows. Then flows that require a fixed amount of bandwidth [30, 76, 203, 224], sometimes referred to as streaming flows, can be viewed as multiserver jobs. However, a communication network also features a network structure that the multiserver-job model does not have. A communication network usually has both streaming flows and flows that are flexible in their bandwidth needs, and streaming flows again operate in the dropping model. The performance metric in such a system typically combines the cost associated with dropping for streaming flows and the cost associated with delay for other flows.

9.3 Model

A basic description of the system parameters and dynamics has been given in the introduction section. In this section, we provide formal descriptions of the scheduling policies, the system states, the scaling regime, and the concept of subsystems used in our analysis.

Scheduling policies. A scheduling policy decides which jobs to put into service at any moment of time. We are interested in the following two policies:

- *First-Come-First-Served (FCFS)*: Jobs are placed onto servers in a First-Come-First-Served fashion until either the next job in queue does not fit or all the jobs are in service.
- *Smallest-Need-First (SNF)*: Recall that the job types are indexed in a way such that $\ell_1 \leq \ell_2 \leq \dots \leq \ell_I$. We assign priorities to job types such that a smaller index has a higher priority. Whenever there is a job arrival or departure, SNF preempts all the jobs in service and determines a new schedule from scratch. SNF starts from job type 1 and places as many type 1 jobs as possible onto servers. After this, if there are still servers available, SNF goes to the next priority level, type 2, and places as many type 2 jobs as possible onto servers. This procedure continues until no more jobs in the queue can fit into the servers.

System state. Under FCFS or SNF, a Markovian representation of the system state can be described as follows. The state $\mathbf{u}$ of the Markov chain is an ordered list of the jobs in the system, sorted in their order of arrival, and each entry of $\mathbf{u}$ describes the type of the corresponding job and whether the job is in service or not. Let the state space be denoted as $\mathcal{U}$. Although the state space is infinite dimensional, in our analysis, we typically only need to focus on three I -dimensional vectors defined below.

For any time t and each job type i , let $X_i(t)$ denote the number of type i jobs in the system, $Z_i(t)$ denote the number of type i jobs in service, and $Q_i(t) \triangleq X_i(t) - Z_i(t)$ denote the number of type i jobs waiting in the queue. Note that since the total number of servers in use cannot exceed n , and we cannot serve more jobs than there are in the system, we have the following constraints:

$$\sum_{i=1}^I \ell_i Z_i(t) \leq n \quad \text{for all } t \geq 0, \quad (9.4)$$

$$Z_i(t) \leq X_i(t) \quad \text{for all } t \geq 0, i \in [I],$$

where $[I]$ denotes the index set $\{1, 2, \dots, I\}$.

Let $X_i(\infty)$, $Z_i(\infty)$, and $Q_i(\infty)$ be random variables that follow the corresponding steady-state distributions when they exist. We sometimes use vector representations of these quantities for convenience. For example, we write $\mathbf{X}(t) = (X_1(t), X_2(t), \dots, X_I(t))$. We define the vectors $\mathbf{Z}(t)$, $\mathbf{Q}(t)$, $\mathbf{X}(\infty)$, $\mathbf{Z}(\infty)$, and $\mathbf{Q}(\infty)$ in a similar way. Note that these random elements correspond to the n server system and thus their distributions depend on n . Throughout this chapter, for conciseness, we often omit the (∞) in the steady-state random elements except in theorem or lemma statements.

Recall that the waiting time for a job is the total time that the job spends in the queue. Our performance metric is the mean waiting time $\mathbb{E}[T^w(\infty)]$, given by

$$\mathbb{E}[T^w(\infty)] = \frac{1}{\lambda} \sum_{i=1}^I \lambda_i \mathbb{E}[T_i^w(\infty)],$$

where $T_i^w(\infty)$ is the waiting time of a type i job in the steady state. Note that the waiting time of a job is the summation of multiple periods if the job

is preempted. In any case, we can still apply Little's law [see, e.g., 137] and rewrite the mean waiting time as

$$\mathbb{E}[T^w(\infty)] = \frac{1}{\lambda} \sum_{i=1}^I \mathbb{E}[Q_i(\infty)].$$

Bounding the mean waiting time is thus equivalent to bounding the expected total queue length.

Scaling regimes. Recall that we consider scaling regimes where number of servers, n , goes to infinity, and the arrival rates λ_i and server needs ℓ_i are allowed to scale with n , *while the service rate μ_i and the number of job types I stay constant*. The scaling regimes are specified by the slack capacity $\delta \triangleq n - \sum_{i=1}^I \frac{\lambda_i \ell_i}{\mu_i}$, the maximal server need $\ell_{\max} \triangleq \max_{i \in [I]} \ell_i = \ell_I$ and another parameter called the *work variability*: $\sigma^2 \triangleq \sum_{i=1}^I \frac{\lambda_i \ell_i^2}{\mu_i^2}$. Work variability reflects the variability of the “work” caused by job arrivals in terms of server–time product, which is $\frac{\ell_i}{\mu_i}$ in expectation for each type i job. To help later presentation, we also define the *load brought by type i jobs* ρ_i as $\rho_i = \frac{\lambda_i \ell_i}{n \mu_i}$.

We state our assumptions below. Throughout the chapter, $\log n$ denotes natural logarithm.

Assumption 9.1 (Heavy traffic assumption). The slack capacity δ is small relative to $\sqrt{\sigma^2}$:

$$\delta = o\left(\frac{\sqrt{\sigma^2}}{\log n}\right). \quad (9.5)$$

Assumption 9.2 (Maximal server need assumption). There exists a constant $\epsilon_0 \in (0, 1)$ s.t.

$$\ell_{\max} \leq \epsilon_0 \delta. \quad (9.6)$$

Assumption 9.3 (Commonness assumption). The load brought by the maximal-need jobs is not too small (note that $\ell_1 \leq \ell_2 \leq \dots \leq \ell_I = \ell_{\max}$):

$$\rho_I \triangleq \frac{\lambda_I \ell_I}{n \mu_I} = \omega \left(\sqrt{\frac{\delta \log n}{\sqrt{\sigma^2}}} \cdot \frac{\ell_{\max}}{n} \log n \right). \quad (9.7)$$

Assumption 9.1 guarantees that the traffic is not too light, while Assumption 9.2 guarantees that the system is stable under FCFS and SNF. In the

simplified setting of Section 9.1 where $\ell_{\max} = n^\gamma$ and $\delta = n^\alpha$, the first two assumptions correspond to $\alpha < \frac{1+\gamma}{2}$ and $\alpha > \gamma$, which exclude the white and grey parts in Figure 9.2, respectively. Assumption 9.3 states that the load brought by the maximal-need jobs are not too small. To understand the right hand side expression in Assumption 9.3, note that it is automatically satisfied when $\rho_I = \omega \left(\sqrt{\frac{\ell_{\max}}{n}} \log n \right)$. For example, when $\ell_{\max} = \Theta(\sqrt{n})$, then it suffices to have $\rho_I = \omega \left(n^{-1/4} \log n \right)$. However, when the traffic becomes heavier, i.e., when $\frac{\delta \log n}{\sqrt{\sigma^2}}$ becomes smaller, Assumption 9.3 in (9.7) can be much weaker than $\rho_I = \omega \left(\sqrt{\frac{\ell_{\max}}{n}} \log n \right)$.

To have an intuitive view of the magnitudes of the parameters, we give the following asymptotics: $\sigma^2 = O(n\ell_{\max})$, $\delta = o\left(\frac{n}{(\log n)^2}\right)$, and $\ell_{\max} \leq \epsilon_0 \delta = o\left(\frac{n}{(\log n)^2}\right)$; they can be verified from the assumptions.

Subsystems. When we analyze SNF and prove the lower bound for all policies, we frequently use the concept of the i -th *subsystem*, which is the system that has all type j jobs in the original system with $j \leq i$ and removes all type k jobs with $k \geq i$. In the i -th subsystem, the slack capacity becomes $\delta_i = n - \sum_{j=1}^i \frac{\lambda_j \ell_j}{\mu_j}$, and the work variability becomes $\sigma_i^2 = \sum_{j=1}^i \frac{\lambda_j \ell_j^2}{\mu_j^2}$. Note that $\delta = \delta_I$ and $\sigma_I^2 = \sigma^2$. The maximal server need in the i -th system is ℓ_i since $\ell_1 \leq \ell_2 \leq \dots \leq \ell_i$.

As i increases, the load of the i -th subsystem gets heavier since δ_i becomes smaller. There is a *critical index* i^* such that

$$i^* = \min \left\{ i \in [I] \mid \delta_i = o\left(\frac{\sqrt{\sigma_i^2}}{\log n}\right) \right\}, \quad (9.8)$$

i.e., the i^* -th subsystem is the smallest subsystem whose traffic regime is as heavy as that of the original system. Recall that we have assumed $\delta = o\left(\frac{\sqrt{\sigma^2}}{\log n}\right)$, so the set in (9.8) contains at least the index I and thus i^* is well-defined. Note that δ_i is monotonically decreasing while σ_i^2 is monotonically increasing. Thus the index i^* serves as a division point: for any i with $i^* \leq i \leq I$, we have $\delta_i = o\left(\frac{\sqrt{\sigma_i^2}}{\log n}\right)$, resulting in a heavier traffic regime; and for any i with $1 \leq i < i^*$, we have $\delta_i = \Omega\left(\frac{\sqrt{\sigma_i^2}}{\log n}\right)$, resulting in a lighter traffic regime.

9.4 Main Results

In this section, we first present our main results under the scaling regimes we specify in Section 9.3 as Theorems 9.4, 9.5, and 9.6. Then we comment on how the theorems subsume the results in Section 9.1.3.

Theorem 9.4 (Mean waiting time under FCFS). *Consider the multiserver-job system with n servers satisfying Assumptions 9.1 and 9.2. Under the FCFS policy, for each $i \in [I]$, the expected waiting time of type i jobs satisfies*

$$\mathbb{E}[T_i^w(\infty)]^{\text{FCFS}} \geq \frac{\sigma^2}{n(\delta + \ell_{\max})} \cdot (1 - o(1)), \quad (9.9)$$

$$\mathbb{E}[T_i^w(\infty)]^{\text{FCFS}} \leq \frac{\sigma^2}{n(\delta - \ell_{\max})} \cdot (1 + o(1)). \quad (9.10)$$

Consequently,

$$\mathbb{E}[T^w(\infty)]^{\text{FCFS}} = \Theta\left(\frac{\sigma^2}{n\delta}\right), \quad \mathbb{E}[T_i^w(\infty)]^{\text{FCFS}} = \Theta\left(\frac{\sigma^2}{n\delta}\right). \quad (9.11)$$

Theorem 9.5 (Mean waiting time lower bound). *Consider the multiserver-job system with n servers satisfying Assumptions 9.1 and 9.2. Under any policy, the mean waiting time satisfies*

$$\mathbb{E}[T^w(\infty)] \geq \max_{i^* \leq i \leq I} \frac{1}{\lambda} \frac{\mu_{\min} \sigma_i^2}{\ell_i \delta_i} \cdot (1 - o(1)) = \Omega\left(\max_{i^* \leq i \leq I} \frac{1}{\lambda} \frac{\sigma_i^2}{\ell_i \delta_i}\right), \quad (9.12)$$

where i^* is the critical index defined in (9.8) and $\mu_{\min} = \min_{i \in [I]} \mu_i$, and the expression represented by $o(1)$ is independent of the policies.

Theorem 9.6 (Mean waiting time under SNF). *Consider the multiserver-job system with n servers satisfying Assumptions 9.1, 9.2, and 9.3. Under the SNF policy, the mean waiting time satisfies*

$$\mathbb{E}[T^w(\infty)]^{\text{SNF}} \leq \frac{1}{\lambda} \sum_{i=i^*}^I \frac{\mu_{\max} \sigma_i^2}{\ell_i(\delta_i - \ell_i)} \cdot (1 + o(1)) = O\left(\max_{i^* \leq i \leq I} \frac{1}{\lambda} \frac{\sigma_i^2}{\ell_i \delta_i}\right), \quad (9.13)$$

where i^* is the critical index defined in (9.8) and $\mu_{\max} = \max_{i \in [I]} \mu_i$. Consequently, the SNF policy achieves the optimal order of the mean waiting time.

We have a more general bound for the SNF policy that holds without Assumption 9.3. Interested readers can refer to Appendix 9.A of the online companion.

The results appearing in Section 9.1.3 are direct consequences of the above theorems. Recall that in the polynomial scaling regimes, we have $\ell_{\max} = \Theta(n^\gamma)$ and $\delta = \Theta(n^\alpha)$, and $\lambda = \Theta(n)$. The requirement $\gamma < \alpha < \frac{\gamma+1}{2}$ implies Assumptions 9.1 and 9.2; the requirement $\lambda_I \ell_I = \Theta(n)$ implies Assumption 9.3. Under these assumptions, the bounds in Theorem 9.4, 9.5 and 9.6 hold. It can be verified that $\sigma^2 = \Theta(n^{1+\gamma})$, so the expression $\frac{\sigma^2}{n\delta}$ in (9.11) has the order $\Theta(n^{\gamma-\alpha})$; the critical index $i^* = I$, so the expression $\max_{i^* \leq i \leq I} \frac{1}{\lambda} \frac{\sigma_i^2}{\ell_i \delta_i}$ in (9.12) and (9.13) equals $\frac{1}{\lambda} \frac{\sigma^2}{\ell_{\max} \delta} = \Theta(n^{-\alpha})$.

9.5 Proof Roadmap and Drift Method Preliminaries

We organize our proofs of the main results as follows: we first prove two important bounds for a quantity called *workload* given by $\sum_{i=1}^I \frac{\ell_i}{\mu_i} Q_i$, in Lemma 9.7 and Lemma 9.9, respectively. Then we convert the workload bounds to the waiting time bounds in Theorem 9.4 and Theorem 9.5 using properties of FCFS and a linear programming relaxation. For Theorem 9.6, we analyze SNF by considering each i -th subsystems for $i \in [I]$. Some subsystems only need Lemma 9.7 and 9.9, while others require an additional Lemma 9.17.

Our proof approach is closely related to the recently developed drift method [89, 195]. The drift method allows us to extract information from a continuous-time Markov chain $\{\mathbf{S}(t)\}_{t \geq 0}$ on state-space $\mathcal{S}$ by computing the *drift* of different test functions. Because $S(t)$ is a Markov chain with countable state space and bounded transition rates, we can define the *drift* of any function f as

$$Gf(\mathbf{s}) \triangleq \lim_{t \rightarrow 0} \mathbb{E} \left[\frac{f(\mathbf{S}(t)) - f(\mathbf{s})}{t} \mid \mathbf{S}(0) = \mathbf{s} \right]. \quad (9.14)$$

We call the operator G the *generator* of the Markov chain.

For a multiserver-job system, let I -dimensional real vectors $\mathbf{x}, \mathbf{z} \in \mathbb{R}_+^I$ be possible realizations of state descriptors $\mathbf{X}(t)$ and $\mathbf{Z}(t)$, where recall that $\mathbf{X}(t)$ is the vector of the number of jobs in the system at time t , and $\mathbf{Z}(t)$

is the vector of the number of jobs in service at time t . We focus on f that only depends on $\mathbf{x}$, i.e., $f : \mathbb{R}_+^I \rightarrow \mathbb{R}$. The drift of f is of the form

$$Gf(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^I \lambda_i (f(\mathbf{x} + \mathbf{e}_i) - f(\mathbf{x})) + \sum_{i=1}^I \mu_i z_i (f(\mathbf{x} - \mathbf{e}_i) - f(\mathbf{x})), \quad (9.15)$$

where $\mathbf{e}_i \in \mathbb{R}_+^I$ is the vector whose i -th entry is 1 and all the other entries are 0. Note that although f and Gf are functions of the system state $\mathbf{u}$, we write $f(\mathbf{x})$ and $Gf(\mathbf{x}, \mathbf{z})$ to highlight the variables that affect their values.

We frequently use the following relation regarding the drift

$$\mathbb{E}[Gf(\mathbf{X}, \mathbf{Z})] = 0. \quad (9.16)$$

Heuristically, this is because when $\mathbf{X}(0)$ and $\mathbf{Z}(0)$ follow the stationary distribution, $\mathbf{X}(t)$ and $\mathbf{Z}(t)$ also follow the stationary distribution, so $f(\mathbf{X}(t))$ and $f(\mathbf{X}(0))$ have the same expectation. Rigorously speaking, this relation only holds for well behaved functions and Markov processes. The conditions under which the relation holds are discussed in detail in Appendix 9.B of the online companion. Throughout the chapter, we assume (9.16) holds for all f that we consider.

9.6 Workload Bounds

In this section, we prove two bounds for a quantity called *workload* given by $\sum_{i=1}^I \frac{\ell_i}{\mu_i} Q_i$. These bounds are fundamental to the proofs of the main theorems. In Lemma 9.7, we give a lower bound on the expected workload applicable to *any policy*. In Lemma 9.9, we give upper bounds on the expected workload under any δ' -*work-conserving policy*, a class of policies defined in Definition 9.8. After stating these two lemmas, we go through preliminaries for the proofs. Finally, we show the full proof of the two lemmas at the end of the section.

Lemma 9.7 (Workload lower bound). *Consider the multiserver-job system with n servers satisfying Assumptions 9.1 and 9.2. Under any policy, the expected workload is lower bounded as*

$$\mathbb{E} \left[\sum_{i=1}^I \frac{\ell_i}{\mu_i} Q_i(\infty) \right] \geq \frac{\sigma^2}{\delta} \cdot (1 - o(1)), \quad (9.17)$$

where the expression represented by $o(1)$ is independent of the policies.

Definition 9.8. We call a policy δ' -work-conserving, if the following equation holds

$$\sum_{i=1}^I \ell_i Z_i(t) \geq \min \left(\sum_{i=1}^I \ell_i X_i(t), n - \delta' \right) \quad \forall t \geq 0. \quad (9.18)$$

Here $\sum_{i=1}^I \ell_i Z_i(t)$ is equal to the number of busy servers at time t , while $\sum_{i=1}^I \ell_i X_i(t)$, which we call the *total server need*, is the potential number of busy servers if we can put all jobs at time t into service. Therefore, under a δ' -work-conserving policy, either all jobs are in service, or there are at most δ' idling servers. Under any $\ell_{\max}$ -work-conserving policy, one can show that the system is stable when $\ell_{\max} \leq \epsilon_0 \delta$ (Assumption 9.2) holds. In particular, the system is stable under both FCFS and SNF.

Lemma 9.9 (Workload upper bound). *Consider the multiserver-job system with n servers under a δ' -work-conserving policy with $\delta' \leq \epsilon_0 \delta$, where $\epsilon_0 \in (0, 1)$ is the parameter in Assumption 9.2. Then when $\delta = o\left(\frac{\sqrt{\sigma^2}}{\log n}\right)$,*

$$\mathbb{E} \left[\sum_{i=1}^I \frac{\ell_i}{\mu_i} Q_i(\infty) \right] \leq \frac{\sigma^2}{\delta - \delta'} \cdot (1 + o(1)) = O\left(\frac{\sigma^2}{\delta}\right); \quad (9.19)$$

when $\delta = \Omega\left(\frac{\sqrt{\sigma^2}}{\log n}\right)$,

$$\mathbb{E} \left[\sum_{i=1}^I \frac{\ell_i}{\mu_i} Q_i(\infty) \right] = O(\sqrt{\sigma^2} \log n). \quad (9.20)$$

Remark 9.10. When Assumption 9.1 is satisfied, i.e., when $\delta = o\left(\frac{\sqrt{\sigma^2}}{\log n}\right)$, the workload upper bound in Lemma 9.9 coincides with the workload lower bound in Lemma 9.7 order-wise, which implies that the expected workload $\mathbb{E} \left[\sum_{i=1}^I \frac{\ell_i}{\mu_i} Q_i(\infty) \right] = \Theta\left(\frac{\sigma^2}{\delta}\right)$. Note that in this case, although the expected workload under all δ' -work-conserving policies has the same order, the mean waiting time can vary among policies, as shown for FCFS and SNF in Theorems 9.4 and 9.6.

9.6.1 Preliminaries for Lemma 9.7 and Lemma 9.9

Our proofs focus on bounding the *normalized work*, defined as

$$\overline{W} \triangleq \sum_{i=1}^I \frac{\ell_i}{\mu_i} (X_i - \bar{x}_i),$$

where we write $\bar{x}_i \triangleq \frac{\lambda_i}{\mu_i}$ for notational simplicity. We claim that normalized work has the same expectation as the workload, i.e., $\mathbb{E}[\bar{W}] = \mathbb{E}[\sum_{i=1}^I \frac{\ell_i}{\mu_i} Q_i]$. To see this, recall that $Q_i = X_i - Z_i$. Now consider the drift of X_i , given by $GX_i = \lambda_i - \mu_i Z_i$. One can verify that X_i satisfies $\mathbb{E}[GX_i] = 0$, and thus $\mathbb{E}[Z_i] = \frac{\lambda_i}{\mu_i}$. Therefore, the expected workload can be written as:

$$\mathbb{E}\left[\sum_{i=1}^I \frac{\ell_i}{\mu_i} Q_i\right] = \mathbb{E}\left[\sum_{i=1}^I \frac{\ell_i}{\mu_i} (X_i - Z_i)\right] = \mathbb{E}\left[\sum_{i=1}^I \frac{\ell_i}{\mu_i} (X_i - \bar{x}_i)\right] = \mathbb{E}[\bar{W}]. \quad (9.21)$$

Therefore, bounding the expected workload is equivalent to bounding the steady-state expectation of the normalized work $\mathbb{E}[\bar{W}]$.

We break $\mathbb{E}[\bar{W}]$ into three terms:

$$\mathbb{E}[\bar{W}] = \bar{r} + \mathbb{E}[(\bar{W} - \bar{r})^+] - \mathbb{E}[(\bar{W} - \bar{r})^-], \quad (9.22)$$

where $\bar{r} \in \mathbb{R}$ is up to our choice; $(\bar{W} - \bar{r})^+ \triangleq \max\{\bar{W} - \bar{r}, 0\}$ denotes the positive part, and $(\bar{W} - \bar{r})^- \triangleq -\min\{\bar{W} - \bar{r}, 0\}$ denotes the negative part.

The major difficulty during the proofs is bounding the expectation of the positive part $\mathbb{E}[(\bar{W} - \bar{r})^+]$. This relies on the relation that $\mathbb{E}[Gf(\mathbf{X}, \mathbf{Z})] = 0$ as introduced in Section 9.5. In our proofs, we choose f to be piecewise quadratic functions to bound the term

$$\mathbb{E}\left[\sum_{i=1}^I \ell_i (Z_i - \bar{x}_i) \cdot (\bar{W} - \bar{r})^+\right].$$

Since $\sum_{i=1}^I \ell_i (Z_i - \bar{x}_i) = \sum_{i=1}^I \ell_i Z_i - n + \delta$, we will be able to bound $\mathbb{E}[(\bar{W} - \bar{r})^+]$ if we are able to give an accurate estimate of the number of busy servers $\sum_{i=1}^I \ell_i Z_i$ when the normalized work $\bar{W} \geq \bar{r}$. To get a precise estimate, we exploit the state-space concentration result that says for each $i \in [I]$, X_i cannot be much smaller than $\bar{x}_i$, i.e., $(X_i - \bar{x}_i)^-$ is small with high probability. Formally, this state-space concentration is established by Lemma 9.11, whose proof uses a sample-path coupling argument and is given in Appendix 9.C of the online companion.

Lemma 9.11. *Consider the multiserver-job system with n servers. For any nonnegative vector $c = (c_1, \dots, c_I) \in \mathbb{R}_+^I$ independent of n , let $c_{\max} = \max_{i \in [I]} c_i$, $\mu_{\max} = \max_{i \in [I]} \mu_i$ and let*

$$\Phi = \sum_{i=1}^I c_i \ell_i (X_i(\infty) - \bar{x}_i),$$

where $\bar{x}_i = \frac{\lambda_i}{\mu_i}$. Then we have the three bounds below.

(a) For any $K \geq 0$,

$$\mathbb{P}(\Phi \leq -K) \leq \exp\left(-\frac{K^2}{2c_{\max}^2\mu_{\max}\sigma^2}\right). \quad (9.23)$$

(b) For any $\alpha \geq 0$ and $\beta \geq 0$ such that $\alpha\beta \geq c_{\max}^2\mu_{\max}\sigma^2$ and any $j \geq 0$,

$$\mathbb{P}(\Phi \leq -\alpha - \beta j) \leq e^{-j}. \quad (9.24)$$

(c) Let $\Phi^- = \max\{-\Phi, 0\}$ to be the negative part of Φ . Then

$$\mathbb{E}[\Phi^-] \leq \sqrt{c_{\max}^2\mu_{\max}\sigma^2}. \quad (9.25)$$

9.6.2 Proofs of Lemma 9.7 and Lemma 9.9

In this section we show the full proof of Lemma 9.7 and Lemma 9.9.

Proof of Lemma 9.7. Recall that in Section 9.6.1, we have shown that $\mathbb{E}\left[\sum_{i=1}^I \frac{\ell_i}{\mu_i} Q_i\right] = \mathbb{E}[\bar{W}]$, where $\bar{W}$ is the *normalized work* given by $\bar{W} \triangleq \sum_{i=1}^I \frac{\ell_i}{\mu_i} (X_i - \bar{x}_i)$ and $\bar{x}_i \triangleq \frac{\lambda_i}{\mu_i}$. Therefore, our goal is equivalent to lower bounding $\mathbb{E}[\bar{W}]$. To do this, we first perform the following decomposition:

$$\mathbb{E}[\bar{W}] = \bar{r}_1 + \mathbb{E}[(\bar{W} - \bar{r}_1)^+] - \mathbb{E}[(\bar{W} - \bar{r}_1)^-].$$

where $\bar{r}_1$ is a properly chosen small number to be specified later. We bound the positive part $\mathbb{E}[(\bar{W} - \bar{r}_1)^+]$ and the negative part $\mathbb{E}[(\bar{W} - \bar{r}_1)^-]$ separately using different techniques.

We bound $\mathbb{E}[(\bar{W} - \bar{r}_1)^+]$ by analyzing the Lyapunov drift of a function $f: \mathbb{R}_+^I \rightarrow \mathbb{R}$ defined below. Let $\mathbf{x}, \mathbf{z} \in \mathbb{R}_+^I$ denote possible realizations of the state descriptors $\mathbf{X}(t), \mathbf{Z}(t)$. Then f is defined as

$$f(\mathbf{x}) = \varphi_M(w(\mathbf{x}) - \bar{r}_1),$$

where $w(\mathbf{x}) \triangleq \sum_{i=1}^I \frac{\ell_i}{\mu_i} (x_i - \bar{x}_i)$ is a possible realization of $\bar{W}$, $\varphi_M(s): \mathbb{R} \rightarrow \mathbb{R}$ is defined as

$$\varphi_M(s) = \begin{cases} 0 & \text{if } s \leq 0 \\ s^2 & \text{if } 0 < s \leq M \\ 2Ms - M^2 & \text{if } s > M \end{cases},$$

and M is a positive number preventing $\varphi_M(s)$ from growing too fast as s gets large. The derivative of $\varphi_M(s)$ is $\varphi'_M(s) = 2 \min\{s^+, M\}$.

We will utilize the relation $\mathbb{E}[Gf(\mathbf{X}, \mathbf{Z})] = 0$, which is implied by Lemma 9.20 in Appendix 9.B of the online companion if we have $\mathbb{E}[|f(\mathbf{X})|] < \infty$. Since $f(\mathbf{x})$ grows linearly fast as $\mathbf{x}$ gets large, $\mathbb{E}[f(\mathbf{x})] < \infty$ follows if we have $\mathbb{E}[X_i] < \infty$ for all $i \in [I]$. On the other hand, the lemma statement holds trivially if $\mathbb{E}[X_i] = \infty$ for some i .

To calculate $Gf(\mathbf{x}, \mathbf{z})$, we first decompose the drift $Gf(\mathbf{x}, \mathbf{z})$ formula (9.15) in the following way:

$$\begin{aligned} Gf(\mathbf{x}, \mathbf{z}) &= \sum_{i=1}^I (\lambda_i - \mu_i z_i) \frac{\partial f}{\partial x_i}(\mathbf{x}) \\ &\quad + \sum_{i=1}^I \lambda_i \left(f(\mathbf{x} + \mathbf{e}_i) - f(\mathbf{x}) - \frac{\partial f}{\partial x_i}(\mathbf{x}) \right) \\ &\quad + \sum_{i=1}^I \mu_i z_i \left(f(\mathbf{x} - \mathbf{e}_i) - f(\mathbf{x}) + \frac{\partial f}{\partial x_i}(\mathbf{x}) \right). \end{aligned} \quad (9.26)$$

It is easy to see that the partial derivatives of f appearing in the first term is given by

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) = \frac{\partial w}{\partial x_i}(\mathbf{x}) \varphi'_M(w(\mathbf{x}) - \bar{r}_1) = \frac{2\ell_i}{\mu_i} \min((w(\mathbf{x}) - \bar{r}_1)^+, M).$$

To bound the remaining two terms, observe that

$$\begin{aligned} f(\mathbf{x} + \mathbf{e}_i) - f(\mathbf{x}) - \frac{\partial f}{\partial x_i}(\mathbf{x}) &= \int_{\xi \in [\mathbf{x}, \mathbf{x} + \mathbf{e}_i]} \left(\frac{\partial f}{\partial x_i}(\xi) - \frac{\partial f}{\partial x_i}(\mathbf{x}) \right) d\xi \\ &= \frac{\ell_i}{\mu_i} \int_{\xi \in [\mathbf{x}, \mathbf{x} + \mathbf{e}_i]} (\varphi'_M(w(\xi) - \bar{r}_1) - \varphi'_M(w(\mathbf{x}) - \bar{r}_1)) d\xi \\ &\geq \frac{2\ell_i}{\mu_i} \int_{\xi \in [\mathbf{x}, \mathbf{x} + \mathbf{e}_i]} (w(\xi) - w(\mathbf{x})) d\xi \cdot \mathbf{1}_{\{\bar{r}_1 + \frac{\ell_{\max}}{\mu_{\min}} < w(\mathbf{x}) < \bar{r}_1 + M - \frac{\ell_{\max}}{\mu_{\min}}\}} \\ &= \frac{\ell_i^2}{\mu_i^2} \cdot \mathbf{1}_{\{\bar{r}_1 + \frac{\ell_{\max}}{\mu_{\min}} < w(\mathbf{x}) < \bar{r}_1 + M - \frac{\ell_{\max}}{\mu_{\min}}\}}, \end{aligned}$$

where the inequality is due to the fact that for any $s_1 \geq s_2$,

$$\varphi'_M(s_1) - \varphi'_M(s_2) \geq 2(s_1 - s_2) \cdot \mathbf{1}_{\{0 \leq s_2 \leq s_1 \leq M\}}, \quad (9.27)$$

which can be verified by brute force calculation. Similarly,

$$\begin{aligned}
f(\mathbf{x} - \mathbf{e}_i) - f(\mathbf{x}) + \frac{\partial f}{\partial x_i}(\mathbf{x}) &= \int_{\xi \in [\mathbf{x} - \mathbf{e}_i, \mathbf{x}]} \left(-\frac{\partial f}{\partial x_i}(\xi) + \frac{\partial f}{\partial x_i}(\mathbf{x}) \right) d\xi \\
&= \frac{\ell_i}{\mu_i} \int_{\xi \in [\mathbf{x} - \mathbf{e}_i, \mathbf{x}]} \left(-\varphi'_M(w(\xi) - \bar{r}_1) + \varphi'_M(w(\mathbf{x}) - \bar{r}_1) \right) d\xi \\
&\geq \frac{2\ell_i}{\mu_i} \int_{\xi \in [\mathbf{x} - \mathbf{e}_i, \mathbf{x}]} (-w(\xi) + w(\mathbf{x})) d\xi \cdot \mathbf{1}_{\{\bar{r}_1 + \frac{\ell_{\max}}{\mu_{\min}} < w(\mathbf{x}) < \bar{r}_1 + M - \frac{\ell_{\max}}{\mu_{\min}}\}} \\
&= \frac{\ell_i^2}{\mu_i^2} \cdot \mathbf{1}_{\{\bar{r}_1 + \frac{\ell_{\max}}{\mu_{\min}} < w(\mathbf{x}) < \bar{r}_1 + M - \frac{\ell_{\max}}{\mu_{\min}}\}} \\
&\geq \frac{\ell_i^2}{\mu_i^2} \mathbf{1}_{\{\bar{r}_1 + \frac{\ell_{\max}}{\mu_{\min}} < w(\mathbf{x}) < \bar{r}_1 + M - \frac{\ell_{\max}}{\mu_{\min}}\}}.
\end{aligned}$$

Plugging the above inequalities into the decomposition of the drift (9.26) and taking expectation on both sides, because $\mathbb{E}[Gf(\mathbf{X}, \mathbf{Z})] = 0$, we have

$$\begin{aligned}
0 \geq \mathbb{E} \left[\sum_{i=1}^I \ell_i (\bar{x}_i - Z_i) \cdot 2 \min((\bar{W} - \bar{r}_1)^+, M) \right. \\
\left. + \sum_{i=1}^I \left(\frac{\lambda_i \ell_i^2}{\mu_i^2} + \frac{Z_i \ell_i^2}{\mu_i} \right) \mathbf{1}_{\{\bar{r}_1 + \frac{\ell_{\max}}{\mu_{\min}} < \bar{W} < \bar{r}_1 + M - \frac{\ell_{\max}}{\mu_{\min}}\}} \right]. \quad (9.28)
\end{aligned}$$

The inequality still holds when we let $M \rightarrow \infty$ inside the expectation:

$$0 \geq \mathbb{E} \left[\sum_{i=1}^I \ell_i (\bar{x}_i - Z_i) \cdot 2(\bar{W} - \bar{r}_1)^+ + \sum_{i=1}^I \left(\frac{\lambda_i \ell_i^2}{\mu_i^2} + \frac{Z_i \ell_i^2}{\mu_i} \right) \mathbf{1}_{\{\bar{W} > \bar{r}_1 + \frac{\ell_{\max}}{\mu_{\min}}\}} \right]. \quad (9.29)$$

Here we are implicitly exchanging $\lim_{M \rightarrow \infty}$ and $\mathbb{E}$, which is legal by the dominated convergence theorem, given that the random variable inside the expectation is dominated by another random variable with a finite expectation, $2n\ell_{\max}(\bar{W} - \bar{r}_1)^+ + 2n\ell_{\max}/\mu_{\min}$. Using the facts that $\mathbb{E}[\sum_{i=1}^I (\frac{\lambda_i \ell_i^2}{\mu_i^2} + \frac{Z_i \ell_i^2}{\mu_i})] = 2\sigma^2$ and $\sum_{i=1}^I (\frac{\lambda_i \ell_i^2}{\mu_i^2} + \frac{Z_i \ell_i^2}{\mu_i}) \leq \frac{2n\ell_{\max}}{\mu_{\min}}$, we get

$$\mathbb{E} \left[\sum_{i=1}^I \ell_i (Z_i - \bar{x}_i) \cdot (\bar{W} - \bar{r}_1)^+ \right] \geq \sigma^2 - \frac{n\ell_{\max}}{\mu_{\min}} \cdot \mathbb{P} \left(\bar{W} \leq \bar{r}_1 + \frac{\ell_{\max}}{\mu_{\min}} \right). \quad (9.30)$$

By Lemma 9.11 with $\Phi = \bar{W}$, $c_i = \frac{1}{\mu_i}$, $K = 2\sqrt{\mu_{\max}\sigma^2 \log n}/\mu_{\min}$, we have $\mathbb{P}(\bar{W} \leq -K) \leq \frac{1}{n^2}$. Note that this K satisfies $K = O(\sqrt{\sigma^2 \log n})$. We take

$\bar{r}_1 = -K - \frac{\ell_{\max}}{\mu_{\min}}$, then $\mathbb{P}(\bar{W} \leq \bar{r}_1 + \frac{\ell_{\max}}{\mu_{\min}}) \leq \frac{1}{n^2}$. Moreover, observe that $\sum_{i=1}^I \ell_i (Z_i - \bar{x}_i) \leq n - (n - \delta) = \delta$, so

$$\mathbb{E}[(\bar{W} - \bar{r}_1)^+] \geq \frac{\sigma^2}{\delta} \cdot (1 - o(1)). \quad (9.31)$$

The lower bound for negative part $\mathbb{E}[(\bar{W} - \bar{r}_1)^-]$ follows from Lemma 9.11(c), which shows that $\mathbb{E}[(\bar{W})^-] = O(\sqrt{\sigma^2})$. Because $s \mapsto s^-$ is a monotonically decreasing function and $\bar{r}_1 \leq 0$,

$$\mathbb{E}[(\bar{W} - \bar{r}_1)^-] \leq \mathbb{E}[(\bar{W})^-] = O(\sqrt{\sigma^2}). \quad (9.32)$$

Therefore, combining (9.31), (9.32) and the fact that $\bar{r}_1 = -O(\sqrt{\sigma^2} \log n + \ell_{\max})$, we have

$$\mathbb{E}[\bar{W}] = \frac{\sigma^2}{\delta} \cdot (1 - o(1)) - O(\sqrt{\sigma^2} \log n + \ell_{\max}) = \frac{\sigma^2}{\delta} \cdot (1 - o(1)),$$

This finishes the proof of Lemma 9.7. $\square$

Proof of Lemma 9.9. Recall that in Section 9.6.1 we have shown that $\mathbb{E}[\sum_{i=1}^I \frac{\ell_i}{\mu_i} Q_i] = \mathbb{E}[\bar{W}]$, where $\bar{W}$ is the *normalized work* given by $\bar{W} \triangleq \sum_{i=1}^I \frac{\ell_i}{\mu_i} (X_i - \bar{x}_i)$ and $\bar{x}_i \triangleq \frac{\lambda_i}{\mu_i}$. Moreover, we have

$$\mathbb{E}[\bar{W}] \leq \bar{r}_2 + \mathbb{E}[(\bar{W} - \bar{r}_2)^+], \quad (9.33)$$

for any number $\bar{r}_2$. Next, we will bound the term $\mathbb{E}[(\bar{W} - \bar{r}_2)^+]$ for a suitably chosen $\bar{r}_2$.

We bound $\mathbb{E}[(\bar{W} - \bar{r}_2)^+]$ by analyzing the Lyapunov drift of a function $f: \mathbb{R}_+^I \rightarrow \mathbb{R}$ defined below. Let $\mathbf{x}, \mathbf{z} \in \mathbb{R}_+^I$ denote possible realizations of the state descriptors $\mathbf{X}(t), \mathbf{Z}(t)$. We define f as

$$f(\mathbf{x}) = \varphi(w(\mathbf{x}) - \bar{r}_2),$$

where $\varphi(s) = (s^+)^2$, and $w(\mathbf{x}) \triangleq \sum_{i=1}^I \frac{\ell_i}{\mu_i} (x_i - \bar{x}_i)$.

This proof relies on the relation that $\mathbb{E}[Gf(\mathbf{X}, \mathbf{Z})] = 0$, which is justified by Lemma 9.22 in Appendix 9.B of the online companion. To calculate $Gf(\mathbf{x}, \mathbf{z})$, we decompose the drift formula (9.15) in the following way:

$$\begin{aligned} Gf(\mathbf{x}, \mathbf{z}) &= \sum_{i=1}^I (\lambda_i - \mu_i z_i) \frac{\partial f}{\partial x_i}(\mathbf{x}) \\ &\quad + \sum_{i=1}^I \lambda_i \left(f(\mathbf{x} + \mathbf{e}_i) - f(\mathbf{x}) - \frac{\partial f}{\partial x_i}(\mathbf{x}) \right) \\ &\quad + \sum_{i=1}^I \mu_i z_i \left(f(\mathbf{x} - \mathbf{e}_i) - f(\mathbf{x}) + \frac{\partial f}{\partial x_i}(\mathbf{x}) \right). \end{aligned} \quad (9.34)$$

It is easy to see that $\frac{\partial f}{\partial x_i}(\mathbf{x})$ in the first term of (9.34) is given by

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) = \frac{\partial r}{\partial x_i}(\mathbf{x}) \varphi'(w(\mathbf{x}) - \bar{r}_2) = \frac{2\ell_i}{\mu_i} (w(\mathbf{x}) - \bar{r}_2)^+.$$

The other two terms in (9.34) can be bounded by constants independent of $\mathbf{x}$ and $\mathbf{z}$ as below:

$$\begin{aligned} f(\mathbf{x} + \mathbf{e}_i) - f(\mathbf{x}) - \frac{\partial f}{\partial x_i}(\mathbf{x}) &= \int_{\xi \in [\mathbf{x}, \mathbf{x} + \mathbf{e}_i]} \left(\frac{\partial f}{\partial x_i}(\xi) - \frac{\partial f}{\partial x_i}(\mathbf{x}) \right) d\xi \\ &= \frac{\ell_i}{\mu_i} \int_{\xi \in [\mathbf{x}, \mathbf{x} + \mathbf{e}_i]} (\varphi'(w(\xi) - \bar{r}_2) - \varphi'(w(\mathbf{x}) - \bar{r}_2)) d\xi \\ &\leq \frac{2\ell_i}{\mu_i} \int_{\xi \in [\mathbf{x}, \mathbf{x} + \mathbf{e}_i]} (w(\xi) - w(\mathbf{x})) d\xi \\ &= \frac{\ell_i^2}{\mu_i^2}, \end{aligned}$$

where the inequality is because $\varphi'(s) = 2s^+$ is 2-Lipschitz continuous. Similarly,

$$f(\mathbf{x} - \mathbf{e}_i) - f(\mathbf{x}) + \frac{\partial f}{\partial x_i}(\mathbf{x}) \leq \frac{\ell_i^2}{\mu_i^2}.$$

Because $\mathbb{E}[Gf(\mathbf{X}, \mathbf{Z})] = 0$, plugging the last inequality into the decomposition of the drift, (9.34), and taking expectation on both sides of (9.34), we get

$$0 \leq \mathbb{E} \left[\sum_{i=1}^I \ell_i (\bar{x}_i - Z_i) \cdot 2(\bar{W} - \bar{r}_2)^+ \right] + \mathbb{E} \left[\sum_{i=1}^I \left(\frac{\lambda_i \ell_i^2}{\mu_i^2} + \frac{Z_i \ell_i^2}{\mu_i} \right) \right]. \quad (9.35)$$

Observe that because $\mathbb{E}[Z_i] = \frac{\lambda_i}{\mu_i}$, the second term in (9.35) can be straightforwardly computed as

$$\mathbb{E}\left[\sum_{i=1}^I \left(\frac{\lambda_i \ell_i^2}{\mu_i^2} + \frac{Z_i \ell_i^2}{\mu_i}\right)\right] = 2 \sum_{i=1}^I \frac{\lambda_i \ell_i^2}{\mu_i^2} = 2\sigma^2. \quad (9.36)$$

Rearranging the terms, we get the following key equation: for any real number $\bar{r}_2$,

$$\mathbb{E}\left[\sum_{i=1}^I \ell_i (Z_i - \bar{x}_i) \cdot (\bar{W} - \bar{r}_2)^+\right] \leq \sigma^2. \quad (9.37)$$

Now suppose we are able to show that

$$\gamma \cdot \mathbb{E}[(\bar{W} - \bar{r}_2)^+] \leq \mathbb{E}\left[\sum_{i=1}^I \ell_i (Z_i - \bar{x}_i) \cdot (\bar{W} - \bar{r}_2)^+\right] + o(1), \quad (9.38)$$

for some $\gamma > 0$, then by (9.37), $\mathbb{E}[(\bar{W} - \bar{r}_2)^+] \leq \frac{\sigma^2 + o(1)}{\gamma}$, so $\mathbb{E}[\bar{W}] \leq \bar{r}_2 + \frac{\sigma^2 + o(1)}{\gamma}$.

We devote the remainder of this proof to proving (9.38). The idea here is to use the following two events to further partition the probability space:

$$\mathcal{E}_1 = \left\{ \sum_{i=1}^I \left(\frac{1}{\mu_{\min}} - \frac{1}{\mu_i}\right) \ell_i (X_i - \bar{x}_i) > -K_2 \right\}, \quad \mathcal{E}_2 = \left\{ \bar{W} \leq \frac{n}{\mu_{\min}} \right\},$$

where K_2 is a suitable number such that $\mathcal{E}_1$ happens with high probability. We break the term on the left of (9.37) based on the three cases $\mathcal{E}_1$, $\mathcal{E}_1^c \cap \mathcal{E}_2$ and $\mathcal{E}_1^c \cap \mathcal{E}_2^c$ to analyze them separately:

$$\begin{aligned} & \mathbb{E}\left[\sum_{i=1}^I \ell_i (Z_i - \bar{x}_i) \cdot (\bar{W} - \bar{r}_2)^+\right] \\ &= \mathbb{E}\left[\sum_{i=1}^I \ell_i (Z_i - \bar{x}_i) \cdot (\bar{W} - \bar{r}_2)^+ \mathbf{1}_{\{\bar{W} > \bar{r}_2, \mathcal{E}_1\}}\right] \end{aligned} \quad (9.39)$$

$$+ \mathbb{E}\left[\sum_{i=1}^I \ell_i (Z_i - \bar{x}_i) \cdot (\bar{W} - \bar{r}_2)^+ \mathbf{1}_{\{\bar{W} > \bar{r}_2, \mathcal{E}_1^c, \mathcal{E}_2\}}\right] \quad (9.40)$$

$$+ \mathbb{E}\left[\sum_{i=1}^I \ell_i (Z_i - \bar{x}_i) \cdot (\bar{W} - \bar{r}_2)^+ \mathbf{1}_{\{\bar{W} > \bar{r}_2, \mathcal{E}_1^c, \mathcal{E}_2^c\}}\right]. \quad (9.41)$$

Case 1: event $\mathcal{E}_1$ happens. Observe that the term in (9.39) is non-zero only when $\bar{W} > \bar{r}_2$ and event $\mathcal{E}_1$ happens, which implies that

$$\sum_{i=1}^I \frac{\ell_i}{\mu_i} (X_i - \bar{x}_i) > \bar{r}_2,$$

$$\sum_{i=1}^I \left(\frac{1}{\mu_{\min}} - \frac{1}{\mu_i} \right) \ell_i (X_i - \bar{x}_i) > -K_2.$$

Adding up the above two inequalities and rearranging the terms yield

$$\sum_{i=1}^I \ell_i (X_i - \bar{x}_i) > \mu_{\min} (\bar{r}_2 - K_2). \quad (9.42)$$

When the above inequality holds, we can invoke the definition of δ' -work-conserving policy and the fact that $\sum_{i=1}^I \ell_i \bar{x}_i = n - \delta$ to get

$$\begin{aligned} \sum_{i=1}^I \ell_i (Z_i - \bar{x}_i) &\geq \min \left(\sum_{i=1}^I \ell_i X_i, n - \delta' \right) - \sum_{i=1}^I \ell_i \bar{x}_i \\ &= \min \left(\sum_{i=1}^I \ell_i (X_i - \bar{x}_i), \delta - \delta' \right) \\ &\geq \min (\mu_{\min} (\bar{r}_2 - K_2), \delta - \delta'). \end{aligned} \quad (9.43)$$

$$\begin{aligned} &\mathbb{E} \left[\sum_{i=1}^I \ell_i (Z_i - \bar{x}_i) \cdot (\bar{W} - \bar{r}_2)^+ \mathbf{1}_{\{\bar{W} > \bar{r}_2, \mathcal{E}_1\}} \right] \\ &\geq \min (\mu_{\min} (\bar{r}_2 - K_2), \delta - \delta') \cdot \mathbb{E} \left[(\bar{W} - \bar{r}_2)^+ \mathbf{1}_{\{\bar{W} > \bar{r}_2, \mathcal{E}_1\}} \right]. \end{aligned} \quad (9.44)$$

Case 2: event $\mathcal{E}_1^c \cap \mathcal{E}_2$ happens. To bound the term in (9.40), we need to analyze the probability of event $\mathcal{E}_1^c$. Consider Lemma 9.11 with $\Phi = \sum_{i=1}^I \left(\frac{1}{\mu_{\min}} - \frac{1}{\mu_i} \right) \ell_i (X_i - \bar{x}_i)$, $c_i = \frac{1}{\mu_{\min}} - \frac{1}{\mu_i}$, $\alpha = \beta = \sqrt{\frac{\mu_{\max}}{\mu_{\min}^2} \sigma^2}$, $j = 3 \log n$. It can be verified that $\alpha\beta \geq c_{\max}^2 \mu_{\max} \sigma^2$. Let $K_2 = \alpha + \beta j$. Then we have

$$\mathbb{P}(\mathcal{E}_1^c) = \mathbb{P} \left(\sum_{i=1}^I \left(\frac{1}{\mu_{\min}} - \frac{1}{\mu_i} \right) \ell_i (X_i - \bar{x}_i) \leq -K_2 \right) \leq \frac{1}{n^3}.$$

Note that this choice of K_2 satisfies $K_2 = O(\sqrt{\sigma^2 \log n})$.

With the upper bound $\mathbb{P}(\mathcal{E}_1^c) \leq \frac{1}{n^3}$, we bound the term in (9.40) in the following way. Observe that $\sum_{i=1}^I \ell_i(Z_i - \bar{x}_i) \geq -n$. Furthermore, when the event $\mathcal{E}_2$ occurs, $\bar{W} \leq \frac{n}{\mu_{\min}}$ and $0 \leq (\bar{W} - \bar{r}_2)^+ \leq \frac{n}{\mu_{\min}}$. Consequently, we have the following two inequalities

$$\begin{aligned} \mathbb{E} \left[\sum_{i=1}^I \ell_i(Z_i - \bar{x}_i) \cdot (\bar{W} - \bar{r}_2)^+ \mathbf{1}_{\{\bar{W} > \bar{r}_2, \mathcal{E}_1^c, \mathcal{E}_2\}} \right] &\geq -n \cdot \frac{n}{\mu_{\min}} \cdot \mathbb{P}(\mathcal{E}_1^c) \geq -\frac{1}{\mu_{\min} n}, \\ \mathbb{E} \left[(\bar{W} - \bar{r}_2)^+ \mathbf{1}_{\{\bar{W} > \bar{r}_2, \mathcal{E}_1^c, \mathcal{E}_2\}} \right] &\leq \frac{n}{\mu_{\min}} \cdot \mathbb{P}(\mathcal{E}_1^c) \leq \frac{1}{\mu_{\min} n^2}. \end{aligned}$$

We can rearrange the above inequalities into a similar form as (9.44):

$$\begin{aligned} &\mathbb{E} \left[\sum_{i=1}^I \ell_i(Z_i - \bar{x}_i) \cdot (\bar{W} - \bar{r}_2)^+ \mathbf{1}_{\{\bar{W} > \bar{r}_2, \mathcal{E}_1^c, \mathcal{E}_2\}} \right] \\ &\geq \min(\mu_{\min}(\bar{r}_2 - K_2), \delta - \delta') \cdot \mathbb{E} \left[(\bar{W} - \bar{r}_2)^+ \mathbf{1}_{\{\bar{W} > \bar{r}_2, \mathcal{E}_1^c, \mathcal{E}_2\}} \right] - \frac{n + \delta}{\mu_{\min} n^2}, \end{aligned} \quad (9.45)$$

where we have used the fact that $\min(\mu_{\min}(\bar{r}_2 - K_2), \delta - \delta') \leq \delta$.

Case 3: event $\mathcal{E}_1^c \cap \mathcal{E}_2^c$ happens. Lastly, we bound the term in (9.41). Observe that $\mathcal{E}_2^c$ implies that $\sum_{i=1}^I \ell_i X_i \geq n$, so an δ' -work-conserving policy will make sure that $\sum_{i=1}^I \ell_i Z_i \geq n - \delta'$. Therefore,

$$\begin{aligned} &\mathbb{E} \left[\sum_{i=1}^I \ell_i(Z_i - \bar{x}_i) \cdot (\bar{W} - \bar{r}_2)^+ \mathbf{1}_{\{\bar{W} > \bar{r}_2, \mathcal{E}_1^c, \mathcal{E}_2^c\}} \right] \\ &\geq (\delta - \delta') \cdot \mathbb{E} \left[(\bar{W} - \bar{r}_2)^+ \mathbf{1}_{\{\bar{W} > \bar{r}_2, \mathcal{E}_1^c, \mathcal{E}_2^c\}} \right]. \end{aligned} \quad (9.46)$$

Combining the results in three cases (9.44) (9.45) (9.46), we have

$$\begin{aligned} &\mathbb{E} \left[\sum_{i=1}^I \ell_i(Z_i - \bar{x}_i) \cdot (\bar{W} - \bar{r}_2)^+ \right] \\ &\geq \min(\mu_{\min}(\bar{r}_2 - K_2), \delta - \delta') \cdot \mathbb{E} [(\bar{W} - \bar{r}_2)^+] - \frac{n + \delta}{\mu_{\min} n^2}. \end{aligned} \quad (9.47)$$

Therefore, we have shown (9.38) with $\gamma = \min(\mu_{\min}(\bar{r}_2 - K_2), \delta - \delta')$. By (9.37), we conclude that

$$\mathbb{E}[\bar{W}] \leq \bar{r}_2 + \mathbb{E}[(\bar{W} - \bar{r}_2)^+] \leq \bar{r}_2 + \frac{\sigma^2}{\min(\mu_{\min}(\bar{r}_2 - K_2), \delta - \delta')} + O\left(\frac{1}{n}\right). \quad (9.48)$$

Next, we simplify (9.48), based on the order of δ . When $\delta = o(\frac{\sqrt{\sigma^2}}{\log n})$, we take $\bar{r}_2 = K_2 + \frac{\delta - \delta'}{\mu_{\min}}$ and get

$$\mathbb{E}[\bar{W}] \leq K_2 + \frac{\delta - \delta'}{\mu_{\min}} + \frac{\sigma^2}{\delta - \delta'} + O\left(\frac{1}{n}\right) = \frac{\sigma^2}{\delta - \delta'} \cdot (1 + o(1)),$$

where we have used the fact that $K_2 = O(\sqrt{\sigma^2} \log n)$ and $\delta = o(\frac{\sqrt{\sigma^2}}{\log n})$.

When $\delta = \Omega(\sqrt{\sigma^2}/\log n)$, we have $\delta \geq C\sqrt{\sigma^2}/\log n$ for some $C > 0$ and n large enough. Because $\delta - \delta' \geq (1 - \epsilon_0)\delta$, we get $\delta - \delta' \geq (1 - \epsilon_0)C\sqrt{\sigma^2}/\log n$ for n large enough. Taking $\bar{r}_2 = K_2 + (1 - \epsilon_0)C\sqrt{\sigma^2}/(\mu_{\min} \log n)$ yields $\min(\mu_{\min}(\bar{r}_2 - K_2), \delta - \delta') \geq (1 - \epsilon_0)C\sqrt{\sigma^2}/\log n$, so

$$\mathbb{E}[\bar{W}] \leq K_2 + \frac{(1 - \epsilon_0)C\sqrt{\sigma^2}}{\mu_{\min} \log n} + \frac{\sqrt{\sigma^2} \log n}{(1 - \epsilon_0)C} + O\left(\frac{1}{n}\right) = O\left(\sqrt{\sigma^2} \log n\right).$$

This finishes the proof. $\square$

9.7 Proof of Theorem 9.4 (Waiting Times under FCFS)

9.7.1 Proof Overview and Lemmas

We prove Theorem 9.4 in this section. We first present some intuition and state two lemmas. Then we give a proof of the theorem using the lemmas. We leave the proofs of the lemmas to Appendix 9.D.

The analysis of FCFS is based on the intuition that, under FCFS, the number of type i jobs in the queue is approximately proportional to its arrival rate λ_i , i.e.

$$\mathbb{E}[Q_i] \approx \frac{\lambda_i}{\lambda} \mathbb{E}[Q_{\Sigma}], \quad (9.49)$$

where $Q_{\Sigma} \triangleq \sum_{i=1}^I Q_i$ is the total queue length. Using this relation, we can easily convert from the expected workload to mean waiting time:

$$\sum_{i=1}^I \frac{\ell_i}{\mu_i} \mathbb{E}[Q_i] \approx \sum_{i=1}^I \frac{\lambda_i \ell_i}{\mu_i} \frac{1}{\lambda} \mathbb{E}[Q_{\Sigma}] = (n - \delta) \mathbb{E}[T^w], \quad (9.50)$$

where the second equality is due to Little's law and the fact that $\sum_{i=1}^I \frac{\lambda_i \ell_i}{\mu_i} = n - \delta$.

This intuition is formalized by considering a *Modified-FCFS* policy, which serves the jobs in a FCFS order, but will not serve the next job until there are at least $\ell_{\max}$ idle servers. In Lemma 9.12, we prove that under Modified-FCFS, (9.49) holds exactly, so the expected queue length of each type of jobs can be computed using the above argument. Moreover, when n is large, we expect Modified-FCFS to be a good approximation of FCFS. In fact, we can define an *upper-bounding system* as the multiserver-job system with n servers under Modified-FCFS, and define a *lower-bounding system* as the multiserver-job system with $n + \ell_{\max}$ servers under Modified-FCFS. In Lemma 9.13, we prove that the queue length of each type of jobs in the original system is sandwiched between those in the two modified systems. Since the two modified systems themselves are very close to each other, we get a tight characterization of FCFS.

Lemma 9.12. *For both the upper-bounding system and the lower-bounding system, we have*

$$\mathbb{E}[Q_i^{(U)}(\infty)] = \frac{\lambda_i}{\lambda} \mathbb{E}[Q_{\Sigma}^{(U)}(\infty)] \quad \forall i \in [I], \quad (9.51)$$

$$\mathbb{E}[Q_i^{(L)}(\infty)] = \frac{\lambda_i}{\lambda} \mathbb{E}[Q_{\Sigma}^{(L)}(\infty)] \quad \forall i \in [I]. \quad (9.52)$$

where $Q_i^{(U)}(\infty)$ and $Q_i^{(L)}(\infty)$ denote the steady-state queue lengths of type i jobs in the upper and lower-bounding systems, respectively; $Q_{\Sigma}^{(U)}(\infty) \triangleq \sum_{i=1}^I Q_i^{(U)}(\infty)$ and $Q_{\Sigma}^{(L)}(\infty) \triangleq \sum_{i=1}^I Q_i^{(L)}(\infty)$ denote the total queue lengths in the steady state.

Lemma 9.13.

$$\mathbb{E}[Q_i^{(L)}(\infty)] \leq \mathbb{E}[Q_i(\infty)] \leq \mathbb{E}[Q_i^{(U)}(\infty)] \quad \forall i \in [I], \quad (9.53)$$

where $Q_i^{(U)}(\infty)$ and $Q_i^{(L)}(\infty)$ denote the steady-state queue lengths of type i jobs in the upper and lower-bounding systems, respectively.

9.7.2 Proof of Theorem 9.4

Proof of Theorem 9.4. Recall that by Little's law, $\mathbb{E}[T_i^w] = \frac{1}{\lambda_i} \mathbb{E}[Q_i]$ and $\mathbb{E}[T^w] = \frac{1}{\lambda} \sum_{i=1}^I \mathbb{E}[Q_i]$. To characterize $\mathbb{E}[Q_i]$, Lemma 9.13 suggests that

we only need to characterize $\mathbb{E}[Q_i^{(L)}]$ and $\mathbb{E}[Q_i^{(U)}]$, the queue lengths in the lower-bounding system and the upper-bounding system, for each $i \in [I]$.

Observe that the lower-bounding system has $n + \ell_{\max}$ servers, with slack capacity $\delta + \ell_{\max}$. Because $\delta + \ell_{\max} \leq (1 + \epsilon_0)\delta = o\left(\frac{\sqrt{\sigma^2}}{\log n}\right)$, we can apply Lemma 9.7 and Lemma 9.12 to get

$$\frac{1}{\lambda} \mathbb{E}[Q_{\Sigma}^{(L)}] \cdot \sum_{i=1}^I \frac{\lambda_i \ell_i}{\mu_i} = \sum_{i=1}^I \frac{\ell_i}{\mu_i} \mathbb{E}[Q_i^{(L)}] \geq \frac{\sigma^2}{\delta + \ell_{\max}} \cdot (1 - o(1)).$$

Recall the facts introduced in Section 9.3 that $\delta = o(n)$ and $\delta = n - \sum_{i=1}^I \frac{\lambda_i \ell_i}{\mu_i}$, so we have $\sum_{i=1}^I \frac{\lambda_i \ell_i}{\mu_i} = n \cdot (1 - o(1))$. Therefore, the above equation implies that $\mathbb{E}[Q_{\Sigma}^{(L)}] \geq \frac{\lambda \sigma^2}{n(\delta + \ell_{\max})} \cdot (1 - o(1))$. By Lemma 9.12,

$$\mathbb{E}[Q_i^{(L)}] \geq \frac{\lambda_i \sigma^2}{n(\delta + \ell_{\max})} \cdot (1 - o(1)) \quad \forall i \in [I]. \quad (9.54)$$

For the upper-bounding system, observe that it has n server and slack capacity δ , and operates under a $\ell_{\max}$ -work-conserving policy. Applying the workload upper bound in Lemma 9.9 yields

$$\frac{1}{\lambda} \sum_{i=1}^I \frac{\ell_i}{\mu_i} \mathbb{E}[Q_i^{(U)}] \leq \frac{\sigma^2}{\delta - \ell_{\max}} \cdot (1 + o(1)).$$

Following a similar argument as in the lower-bounding system, we have

$$\mathbb{E}[Q_i^{(U)}] \leq \frac{\lambda_i \sigma^2}{n(\delta - \ell_{\max})} \cdot (1 + o(1)) \quad \forall i \in [I]. \quad (9.55)$$

Therefore, by Lemma 9.13, we have

$$\frac{\lambda_i \sigma^2}{n(\delta + \ell_{\max})} \cdot (1 - o(1)) \leq \mathbb{E}[Q_i] \leq \frac{\lambda_i \sigma^2}{n(\delta - \ell_{\max})} \cdot (1 + o(1)) \quad \forall i \in [I], \quad (9.56)$$

which implies the waiting time bounds in Theorem 9.4. $\square$

9.8 Proof of Theorem 9.5 (Mean Waiting Time Lower Bound)

Proof of Theorem 9.5. Recall that by Little's law, $\mathbb{E}[T^w] = \frac{1}{\lambda} \sum_{i=1}^I \mathbb{E}[Q_i]$. Therefore, it suffices to show a lower bound on the total queue length. We

fix a policy in the original system. For any i with $i^* \leq i \leq I$, we consider the i -th subsystem by ignoring all job types with index greater than i . In the i -th subsystem, it is always possible to achieve the same $\mathbb{E}[Q_j]$'s for $j \leq i$ by imitating the service decisions taken by the original system. Therefore, we have $\sum_{j=1}^i \frac{\ell_j}{\mu_j} \mathbb{E}[Q_j] \geq \frac{\sigma_i^2}{\delta_i} \cdot (1 - o(1))$, where the right hand side expression is the workload lower bound of the i -th subsystem according to Lemma 9.7. Then the expected waiting time $\mathbb{E}[T^w]$ is lower-bounded by the optimal value of the following linear programming problem:

$$\begin{aligned} & \min_{\{q_j : j \in [I]\}} \quad \frac{1}{\lambda} \sum_{j=1}^I q_j \\ & \text{subject to} \quad \sum_{j=1}^i \frac{\ell_j}{\mu_j} q_j \geq \frac{\sigma_i^2}{\delta_i} \cdot (1 - o(1)) \quad i^* \leq i \leq I \\ & \quad \quad \quad q_j \geq 0 \quad \forall j \in [I], \end{aligned}$$

where q_j corresponds to $\mathbb{E}[Q_j]$. It is easy to see that the objective value satisfies

$$\frac{1}{\lambda} \sum_{j=1}^I q_j \geq \frac{1}{\lambda} \frac{\mu_{\min}}{\ell_i} \sum_{j=1}^i \frac{\ell_j}{\mu_j} q_j \geq \frac{\mu_{\min} \sigma_i^2}{\lambda \ell_i \delta_i} \cdot (1 - o(1)), \quad (9.57)$$

for any $i^* \leq i \leq I$, where in the first inequality we have used the fact that $\mu_j \geq \mu_{\min}$ and $\ell_j \leq \ell_i$ for any $j \leq i$. Note that when q_j is zero for each j with $j \neq i$, the first inequality becomes an equality up to a constant order factor in terms of μ_i and $\mu_{\min}$. Because i in (9.57) can be an arbitrary integer satisfying $i^* \leq i \leq I$, the optimal objective value is no less than

$$\max_{i^* \leq i \leq I} \frac{\mu_{\min} \sigma_i^2}{\lambda \ell_i \delta_i} \cdot (1 - o(1)). \quad (9.58)$$

Therefore, $\mathbb{E}[T^w] \geq \max_{i^* \leq i \leq I} \frac{\mu_{\min} \sigma_i^2}{\lambda \ell_i \delta_i} \cdot (1 - o(1))$. This completes the proof. $\square$

Remark 9.14. The proof of the lower bound provides some intuitions for choosing the SNF policy. We consider the simple case where $i^* = I$. By (9.57), the total queue length order-wise achieves the lower bound when the queue consists of jobs with the largest server needs, suggesting that we should give low priorities to those jobs. This is in a similar spirit to the famous Shortest-Remaining-Service-Time (SRPT) policy for single-server jobs, which leaves the jobs with large remaining service times in the queue [See, e.g., 137].

9.9 Proof of Theorem 9.6 (Mean Waiting Time under SNF)

To understand the behavior under the SNF policy, one key observation is that for each $i \in [I]$, the type i jobs are unaffected by type j jobs with $j > i$. As a result, we can learn about the original system by analyzing each i -th subsystem, which is obtained by removing all jobs of type j with $j > i$. Some subsystems are under relatively heavier traffic, or more precisely, subject to $\delta_i = o\left(\frac{\sqrt{\sigma_i^2}}{\log n}\right)$, while some subsystems are under lighter traffic. For those subsystems under relatively heavier traffic, Lemma 9.7 and Lemma 9.9 are enough for use; for those subsystems under lighter traffic, we sometimes use Lemma 9.17 below, which is a more refined bound on the expected total server need $\mathbb{E}\left[\sum_{i=1}^I \ell_i X_i\right]$, proved based on the two tail bounds in Lemma 9.15 and Lemma 9.16. The proofs of Lemmas 9.15, Lemma 9.16, and 9.17 are in Appendix 9.E of the online companion.

Lemma 9.15. *Consider the multiserver-job system with n servers satisfying $\ell_{\max} \leq \epsilon_0 \delta$ (Assumption 9.2). Letting $\bar{x}_i = \frac{\lambda_i}{\mu_i}$, under any $\ell_{\max}$ -work-conserving policy, the normalized work has the following tail bound: for any ϵ such that $0 < \epsilon < \epsilon_0$, there exists $\alpha_1 = \frac{2\epsilon\delta}{\mu_{\min}} + \Theta\left(\frac{n\ell_{\max}}{\delta} \log n\right)$ and $\beta_1 = \Theta\left(\frac{n\ell_{\max}}{\delta}\right)$ such that for any $j \geq 0$,*

$$\mathbb{P}\left(\sum_{i=1}^I \frac{\ell_i}{\mu_i} (X_i(\infty) - \bar{x}_i) \geq \alpha_1 + \beta_1 \cdot j\right) \leq e^{-j}. \quad (9.59)$$

Lemma 9.16. *Consider the multiserver-job system with n servers satisfying $\ell_{\max} \leq \epsilon_0 \delta$ (Assumption 9.2). Letting $\bar{x}_i = \frac{\lambda_i}{\mu_i}$, under any $\ell_{\max}$ -work-conserving policy, the total server need has the following tail bound: there exists α_2 and β_2 with $\alpha_2 = \frac{\delta}{2} + \Theta\left(\frac{n\ell_{\max}}{\delta} \log n\right)$ and $\beta_2 = \Theta\left(\frac{n\ell_{\max}}{\delta}\right)$ such that for any $j \geq 0$,*

$$\mathbb{P}\left(\sum_{i=1}^I \ell_i (X_i(\infty) - \bar{x}_i) \geq \alpha_2 + \beta_2 \cdot j\right) \leq e^{-j}. \quad (9.60)$$

The proof technique of the two lemmas is using state space concentration successively: Lemma 9.15 relies on the state-space concentration implied by Lemma 9.11, while Lemma 9.16 relies on the state-space concentration implied by Lemma 9.11 and Lemma 9.15.

As a consequence of the first two lemmas, we can give a bound on the expectation of the total server need $\mathbb{E}[\sum_{i=1}^I \ell_i X_i]$, in a different traffic regime than what is assumed in Assumption 9.1. This is useful for analyzing the dynamics of subsystems under SNF.

Lemma 9.17 (Total server need upper bound under a lighter traffic). *Consider the multi-server-job system with n servers satisfying $\ell_{\max} \leq \epsilon_0 \delta$ (Assumption 9.2), and $\delta = \omega(\sqrt{n\ell_{\max}} \log n)$. Letting $\bar{x}_i = \frac{\lambda_i}{\mu_i}$, under any $\ell_{\max}$ -work-conserving policy, the expected total server need has the following upper bound:*

$$\mathbb{E}\left[\sum_{i=1}^I \ell_i (X_i(\infty) - \bar{x}_i)\right] = \exp\left(-\Omega\left(\frac{\delta^2}{n\ell_{\max}}\right)\right). \quad (9.61)$$

As a quick digression, with Lemma 9.16, we can prove an upper bound on the queueing probability in Corollary 9.18. We comment that this queueing probability bound significantly improves on the bound in Wang et al. [271] in a slightly more constrained traffic regime.

Corollary 9.18 (Queueing probability). *Consider the multiserver-job system with n servers satisfying $\ell_{\max} \leq \epsilon_0 \delta$ (Assumption 9.2), and $\delta = \omega(\sqrt{n\ell_{\max}} \log n)$. Then the steady-state probability that a job experiences queueing has the following upper bound:*

$$\mathbb{P}\left(\sum_{i=1}^I \ell_i X_i(\infty) \geq n\right) = \exp\left(-\Omega\left(\frac{\delta^2}{n\ell_{\max}}\right)\right). \quad (9.62)$$

Now we are ready to prove Theorem 9.6.

Proof of Theorem 9.6. Recall that by Little's Law, we have that $\mathbb{E}[T^w] = \frac{1}{\lambda} \sum_{i=1}^I \mathbb{E}[Q_i]$, and thus it suffices to bound $\mathbb{E}[Q_i]$'s. We bound $\mathbb{E}[Q_i]$ separately for each $i \in [I]$ as follows:

$$\mathbb{E}[Q_i] \leq \frac{1}{\ell_i} \mathbb{E}\left[\sum_{j=1}^i \ell_j Q_j\right] \leq \frac{1}{\ell_i} \mathbb{E}\left[\sum_{j=1}^i \ell_j (X_j - \bar{x}_j)\right], \quad (9.63)$$

where we have used the fact that Q_i 's are non-negative, $Q_j = X_j - Z_j$ and $\mathbb{E}[Z_j] = \frac{\lambda_j}{\mu_j} = \bar{x}_j$. Observe that under SNF, the dynamics of the first i types of jobs are unaffected by the rest of the jobs. Therefore, the expectation $\mathbb{E}\left[\sum_{j=1}^i \ell_j (X_j - \bar{x}_j)\right]$ can be viewed as the expected total server need in the

i -th subsystem, which has slack capacity δ_i , maximal server need ℓ_i , and work variability σ_i^2 . We discuss the bound on $\mathbb{E}[Q_i]$ in three cases based on different relationships of δ_i , ℓ_i and σ_i^2 .

Case 1: $\delta_i = o\left(\frac{\sqrt{\sigma_i^2}}{\log n}\right)$. Applying Lemma 9.9 to the i -th subsystem, we have

$$\mathbb{E}\left[\sum_{j=1}^i \ell_j (X_j - \bar{x}_j)\right] \leq \mu_{\max} \mathbb{E}\left[\sum_{j=1}^i \frac{\ell_j}{\mu_j} (X_j - \bar{x}_j)\right] \leq \frac{\mu_{\max} \sigma_i^2}{\delta_i - \ell_i} \cdot (1 + o(1)).$$

Therefore, $\mathbb{E}[Q_i] \leq \frac{\mu_{\max} \sigma_i^2}{\ell_i(\delta_i - \ell_i)} \cdot (1 + o(1))$.

Case 2: $\delta_i = \Omega\left(\frac{\sqrt{\sigma_i^2}}{\log n}\right)$ and $\ell_i = \omega\left(\frac{\delta \log n}{\sqrt{\sigma^2}} \ell_{\max}\right)$. Applying Lemma 9.9 to the i -th subsystem, we have

$$\mathbb{E}\left[\sum_{j=1}^i \ell_j (X_j - \bar{x}_j)\right] \leq \mu_{\max} \mathbb{E}\left[\sum_{j=1}^i \frac{\ell_j}{\mu_j} (X_j - \bar{x}_j)\right] = O\left(\sqrt{\sigma_i^2} \log n\right).$$

Combining the above equation with (9.63) and the facts that $\ell_i = \omega\left(\frac{\delta \log n}{\sqrt{\sigma^2} \ell_{\max}}\right)$ and $\sigma_i^2 \leq \sigma^2$, we have $\mathbb{E}[Q_i] \leq O\left(\frac{\sqrt{\sigma_i^2} \log n}{\ell_i}\right) = o\left(\frac{\sigma^2}{\delta} \cdot \frac{1}{\ell_{\max}}\right)$.

Case 3: $\delta_i = \Omega\left(\frac{\sqrt{\sigma_i^2}}{\log n}\right)$ and $\ell_i = O\left(\frac{\delta \log n}{\sqrt{\sigma^2}} \ell_{\max}\right)$. We first show that $\delta_i = \omega\left(\sqrt{n \ell_i} \log n\right)$. To see this, observe that $\delta_i = \Omega\left(\frac{\sqrt{\sigma_i^2}}{\log n}\right)$ implies $i < I$, so $\delta_i \geq \frac{\lambda_I \ell_I}{\mu_I} = n \rho_I$. By Assumption 9.3, we have

$$\delta_i \geq n \rho_I = \omega\left(\sqrt{\frac{\delta \log n}{\sqrt{\sigma^2}}} \cdot n \ell_{\max} \log n\right) = \omega\left(\sqrt{n \ell_i} \log n\right). \quad (9.64)$$

Therefore, we can apply Lemma 9.17 to the i -th subsystem to get:

$$\mathbb{E}\left[\sum_{j=1}^i \ell_j (X_j - \bar{x}_j)\right] = \exp\left(-\Omega\left(\frac{\delta_i^2}{n \ell_i}\right)\right) \leq \exp\left(-\Omega((\log n)^2)\right),$$

where we have used $\delta_i = \omega\left(\sqrt{n \ell_i} \log n\right)$ in the last inequality. Therefore, $\mathbb{E}[Q_i] \leq \exp(-\Omega((\log n)^2))$, which decays faster than any polynomial in n .

Combining the three cases, we get

$$\mathbb{E}[T^w] = \frac{1}{\lambda} \sum_{i=1}^I \mathbb{E}[Q_i] \leq \frac{1}{\lambda} \sum_{i=i^*}^I \frac{\mu_{\max} \sigma_i^2}{\ell_i(\delta_i - \ell_i)} \cdot (1 + o(1)),$$

where we have used the fact that $\mathbb{E}[T^w] = \frac{1}{\lambda} \sum_{i=1}^I \mathbb{E}[Q_i]$. Here the summation is taken from i^* to I because Case 1 is equivalent to $i^* \leq i \leq I$. The second inequality is because the other two terms are of lower order compared with $\frac{\mu_{\max} \sigma_I^2}{\ell_I(\delta_I - \ell_I)}$.

Finally, note that I and $\mu_{\max}$ are independent of n during scaling, and $\ell_i \leq \ell_{\max} \leq \epsilon_0 \delta \leq \epsilon_0 \delta_i$ for some $\epsilon_0 < 1$. Therefore, $\frac{1}{\lambda} \sum_{i=i^*}^I \frac{\mu_{\max} \sigma_i^2}{\ell_i(\delta_i - \ell_i)} = \Theta\left(\max_{i^* \leq i \leq I} \frac{1}{\lambda} \frac{\sigma_i^2}{\ell_i \delta_i}\right)$. This finishes the proof. $\square$

9.10 Simulation Results

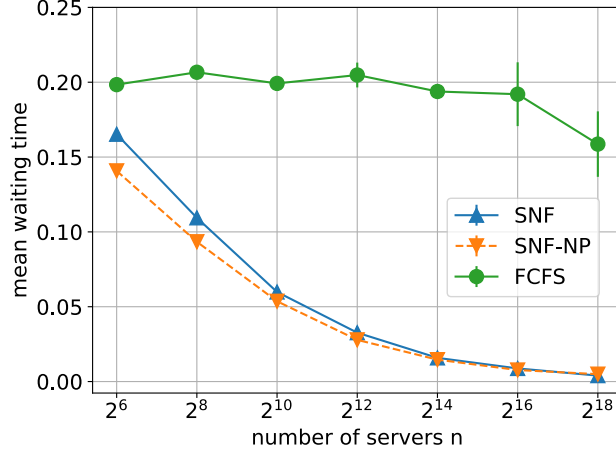
Experiment 1: Comparing SNF and SNF-NP with FCFS.

Our theoretical analysis shows that there is an order gap in mean waiting times under FCFS and SNF, for the scaling regimes satisfying Assumptions 9.1, 9.2 and 9.3. Here, we complement our asymptotic results by comparing the simulated performances of SNF and FCFS in finite systems. We also simulate the non-preemptive variant of SNF which we call SNF-NP. SNF-NP identifies jobs in the queue that can fit into the idle servers and prioritizes those with the smallest server needs for service.

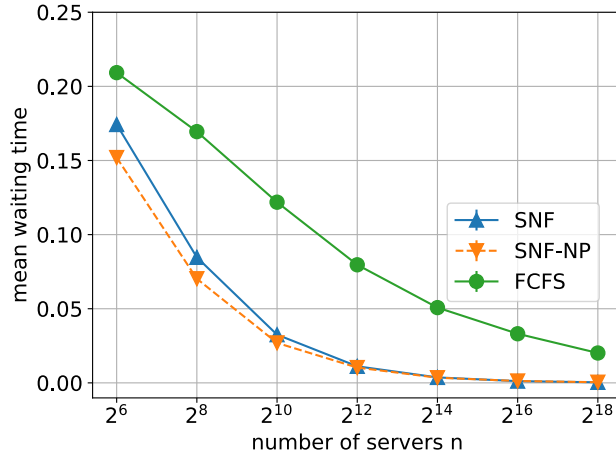
We run the simulation experiments under two sets of parameters.² The Parameter Set One satisfies all three assumptions, while the Parameter Set Two does not satisfy Assumption 9.3. The parameters are specified in the caption of Figure 9.3. We can see that both SNF and SNF-NP outperform FCFS by a large margin. In particular, the ratio between the mean waiting times of FCFS and SNF is always greater than 3 when $n \geq 2^{10}$, and it gets to as large as 40.0 under Parameter Set One, and 54.4 under Parameter Set Two.

A secondary finding is that SNF-NP performs slightly better than SNF. This phenomenon is also observed in the next few experiments, the reason for which is unclear from our current theory. We hypothesize that under SNF, the jobs with large server needs sometimes do not fit into the remaining servers after they are preempted, leading to more idle servers. SNF-NP alleviates

²In all of our experiments, we run a long enough trajectory to estimate the mean and the confidence interval for each data point. The confidence interval is estimated through batch means method [see 17]: we run a long trajectory, divide it into 20 batches, and calculate the standard deviation of the means across the batches; the radius of the confidence interval is 1.96 times of the standard deviation estimated from the batch means.



(a) Mean waiting time under the Parameter Set One. There are three job types, with service rates and server needs given by $\mu_1 = 0.25$, $\ell_1 = 1$; $\mu_2 = 0.5$, $\ell_2 = \lfloor \log_2 n \rfloor$; $\mu_3 = 1$, $\ell_3 = \lfloor \sqrt{n} \rfloor$; slack capacity $\delta = 2\lfloor \sqrt{n} \rfloor$; the arrival rates are chosen so that the loads satisfy $\rho_1 = \rho_2 = \rho_3 = \frac{n-\delta}{3n}$.



(b) Mean waiting time under the Parameter Set Two. There are three job types, with the same service rates, server needs and slack capacity as the Parameter Set One. The arrival rates are chosen so that the loads satisfy $\rho_1 = \rho_2 = \frac{n-\delta-n^{0.7}}{2n}$, $\rho_3 = n^{-0.3}$.

Figure 9.3: The mean waiting times of FCFS, SNF, and SNF-NP under two sets of parameters.

this issue by prohibiting preemption. However, this advantage of SNF-NP is moderate because the maximal server need is smaller than the slack capacity in our scaling regimes.

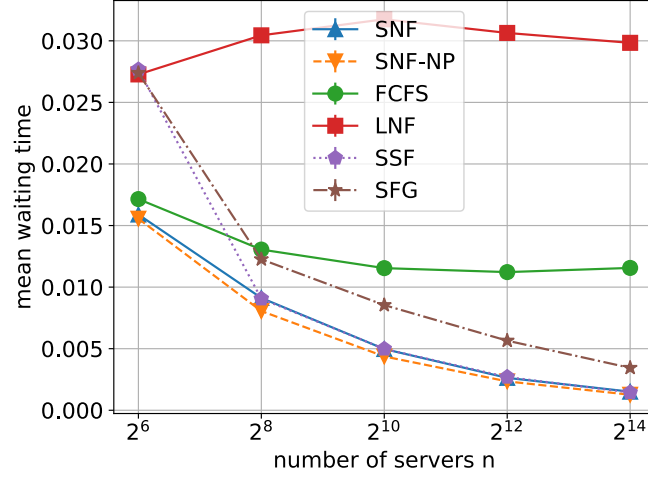
Experiment 2: Comparing with More Policies.

One limitation of our policies, SNF and SNF-NP, is that they do not utilize the service rate information. Although we can show the order-wise optimality of SNF, our analysis is restricted to the scaling regimes where the service rates μ_i 's do not scale with n , i.e. the values of μ_i 's can not be too large or small. When the values of μ_i 's vary significantly across different job types, it becomes questionable whether scheduling jobs solely based on server needs is the most effective way. For example, it could be more natural to prioritize jobs with smaller *expected sizes*, defined as $\frac{\ell_i}{\mu_i}$ for jobs of type i .

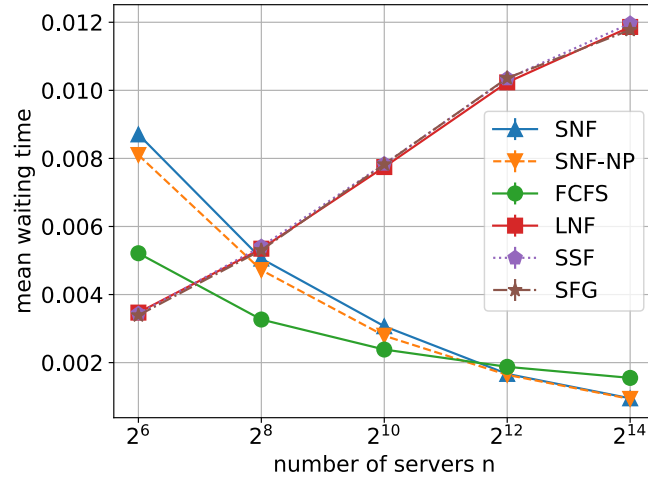
Another aspect not considered in SNF and SNF-NP is the *packing effect*: SNF and SNF-NP do not aim to minimize the number of idle servers, $n - \sum_{i=1}^I \ell_i Z_i$, when choosing the subset of jobs to serve, which might harm the mean waiting times. Furthermore, the approach of SNF and SNF-NP is somewhat contrary to common bin-packing algorithms. For instance, the best-fit algorithm prioritizes packing jobs with larger server needs rather than those with smaller server needs.

Given the motivation discussed, we compare the mean waiting times of SNF and SNF-NP with the following policies that utilize the expected sizes or consider the packing effect:

- **Smallest-Size-First (SSF)**: SSF prioritizes jobs with smaller values of $\frac{\ell_i}{\mu_i}$ preemptively; it is the special case of the $c\mu/m$ -rule in [306] with c being the all-one vector.
- **Larges-Need-First (LNF)**: LNF prioritizes jobs with larger server needs preemptively; it can be seen as following the best-fit algorithm to improve the packing.
- **ServerFilling-Gittins (SFG)**: SFG is a policy designed for multiserver job models with general service time distributions, under the assumption that the number of servers n and the server needs are powers of 2 [129]. Adapted to our setting, SFG schedules jobs based on both the expected sizes and the server needs, aiming to prioritize jobs with smaller expected sizes while optimizing packing.



(a) Mean waiting time under Parameter Set Three. There are three job types, with service rates and server needs $\mu_1 = 1, \ell_1 = 1; \mu_2 = 16, \ell_2 = 8; \mu_3 = 16, \ell_3 = \lfloor \sqrt{n} \rfloor$; slack capacity $\delta = 2\lfloor \sqrt{n} \rfloor$. Arrival rates are chosen according to loads $\rho_1 = \rho_2 = \rho_3 = \frac{n-\delta}{3n}$.



(b) Mean waiting time under Parameter Set Four. There are three job types, with service rates and server needs $\mu_1 = 1, \ell_1 = 1; \mu_2 = 32, \ell_2 = 8; \mu_3 = 256, \ell_3 = \lfloor \sqrt{n} \rfloor$; slack capacity $\delta = 2\lfloor \sqrt{n} \rfloor$. Arrival rates are chosen according to loads $\rho_1 = \rho_2 = \rho_3 = \frac{n-\delta}{3n}$.

Figure 9.4: The mean waiting times of six different policies in systems with different service rates.

We simulate the three policies described above along with SNF, SNF-NP, and FCFS. The parameters are similar to those used in Figure 9.3(a), with two modifications: we set ℓ_2 to be a power of 2 and increase the service rates for job types 2 and 3. The simulations are conducted under two sets of parameters with different service rates, namely Parameter Set Three and Parameter Set Four, as illustrated in Figure 9.4(a) and (b) respectively.

From Figure 9.4, we can see that LNF is always worse than SNF and SNF-NP for $n \geq 2^8$. The performance of SSF depends on its priority order: when it gives type 3 jobs the lowest priority (e.g., when $n \geq 2^8$ in Figure 9.4(a)), it performs similarly to SNF and SNF-NP; otherwise it performs similarly to LNF (for all n in Figure 9.4(b)). The performance of SFG lies between LNF and SSF. None of the three policies outperform SNF and SNF-NP for $n \geq 2^8$.

The comparison between LNF and SFG with SNF and SNF-NP indicates that considering the packing effect does not bring much benefit to the mean waiting time in the scaling regimes of the two examples in Figure 9.4. This outcome aligns with our expectations because the slack capacity here is larger than the maximal server needs, which ensures that any $\ell_{\max}$ -work-conserving policy maintains enough busy servers to quickly reduce the workload even if the packing is not optimal.

On the other hand, the relationships between SSF, LNF and SNF are more interesting: they suggest that in these examples, a policy is likely to have small mean waiting times when *the maximal-need jobs* (i.e., type 3 jobs) are served with the *lowest priority*, even if those jobs may have large service rates and the smallest expected sizes, as we can see from the data points in Figure 9.4(b) with $n \geq 2^8$. Fully understanding this phenomenon is outside the scope of our theory because we assume that the service rates do not scale with n , while in Figure 9.4(b), the service rates can be even larger than the server needs. In Appendix 9.F, we give further details of Experiment 2 and provide some preliminary hypotheses about this phenomenon. There we make an interesting observation that in the setting of Experiment 2, the workload is considerably smaller under SNF than under LNF. Intuitively, SNF make the jobs with large service rates and large server needs stay longer in the system by giving them low priorities; those jobs will help keep more servers busy during periods when there are fewer jobs in the system.

Experiment 3: Mean Waiting Times Weighted by Server Needs.

We have been focusing on the unweighted mean waiting time. However, sometimes we may want to give the jobs with larger server needs larger weights – for example, when each multiserver job consists of many unit-server-need components, and one cares about the mean waiting time averaged over the components.

In Figure 9.5, we consider mean waiting times weighted by $\sqrt{\ell_i}$ (Figure 9.5 (a)) and weighted by ℓ_i (Figure 9.5 (b)) under Parameter Set Three (same model parameters as Figure 9.4 (a)). We can see that when the mean waiting times are weighted by $\sqrt{\ell_i}$, for $n \geq 2^{10}$, SFG outperforms other policies, followed by SNF-NP, SNF, SSF, and lastly LNF and FCFS. When the mean waiting times are weighted by ℓ_i , LNF shows the best performance for all n , followed by SFG and other policies.

These two simulations indicate that generally, when the weights of the mean waiting times depend on the server needs, other policies could outperform SNF and SNF-NP. Nevertheless, as shown in Figure 9.5 (a), SNF and SNF-NP still perform reasonably well when the weights moderately depend on the server needs.

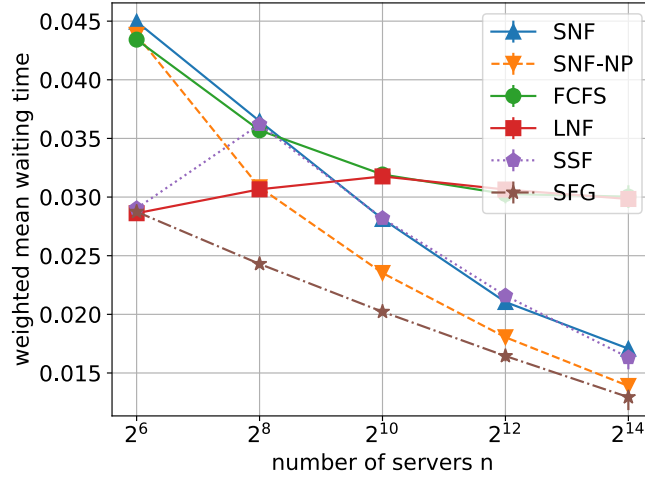
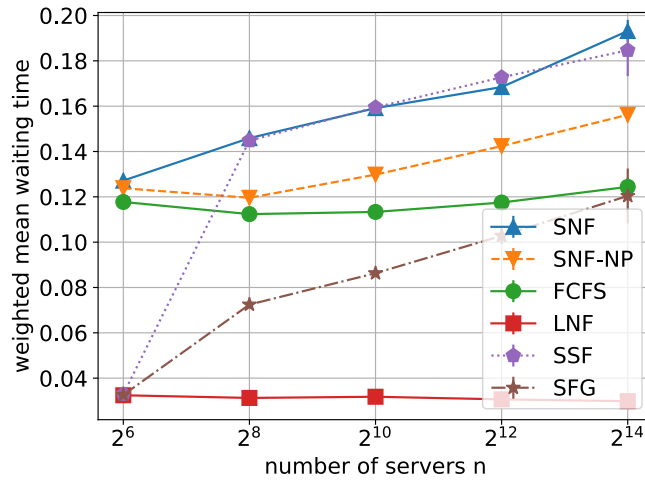
(a) The weight is $\sqrt{\ell_i}$ for each type i (b) The weight is ℓ_i for each type i

Figure 9.5: The weighted mean waiting times of six different policies under Parameter Set Three.

9.11 Conclusion and Future Work

In this chapter, we establish order-wise sharp bounds on the mean waiting times of multiserver jobs under the FCFS policy and the SNF policy. We also prove a lower bound for the mean waiting times under all policies. These bounds imply the order-wise optimality of SNF and the strict suboptimality of FCFS.

Our analysis has two steps: we first prove the workload bounds stated in Lemma 9.7 and Lemma 9.9 for a general class of policies using the drift method, along with some novel state-space concentration arguments. Then we use the workload bounds to analyze the mean waiting times under FCFS and SNF, and prove a lower bound for the mean waiting times under all policies.

Apart from the theoretical analysis, we also perform simulation experiments. We first demonstrate through simulations the performance improvement of SNF compared with FCFS in finite systems, and the fact that SNF-NP, the non-preemptive variant of SNF, has comparable performance with SNF. We also compare SNF and SNF-NP with other policies that utilize the service rate information or consider packing effects. We find that in our examples, when the number of servers is large, SNF and SNF-NP outperform those policies. Finally, we investigate the weighted mean waiting times under these policies, where jobs with larger server needs have larger weights. We find that in this case, other policies could outperform SNF and SNF-NP; but SNF and SNF-NP still perform reasonably well when the weights moderately depend on the server needs.

There are several interesting directions for future work:

1. Derive a tighter bound on the mean waiting time under SNF when Assumption 9.3 is violated.
2. Design a policy that achieves order-wise optimality in the many-server scaling limit without the maximal server need assumption (Assumption 9.2).
3. Understand why the performance of SNF-NP appears to be slightly better than SNF.
4. Extend the analysis to the settings where the service rates scale with n ; this could help us understand why SNF and SNF-NP outperform other policies in Experiment 2.

9.A Discussion on Settings without Assumption 9.3 (Commonness Assumption)

Recall that the bound on the mean waiting time under SNF in Theorem 9.6 relies on Assumption 9.3 (commonness assumption). Here we present a bound on the mean waiting time of SNF without Assumption 9.3. Note that, however, this bound does not imply the order-wise optimality of SNF.

We first define some necessary notation for stating the bound. Let $i_1^* \triangleq \min \{i \in [I] \mid \delta_i = O(\sqrt{n\ell_i \log n})\}$, and let $i_2^* \triangleq i^* = \min \{i \in [I] \mid \delta_i = o(\sqrt{\sigma_i^2 / \log n})\}$. Because of Assumption 9.1, these two sets are non-empty and i_1^*, i_2^* are well-defined. Because $n\ell_i \geq \mu_{\min} \sigma_i^2$ and δ_i 's are monotonically decreasing, we have $i_1^* \leq i_2^*$. Moreover, σ_i^2 and ℓ_i are monotonically increasing, so for any $i_2^* \leq i \leq I$, $\delta_i = o(\sqrt{\sigma_i^2 / \log n})$; for any $i_1^* \leq i < i_2^*$, $\delta_i = O(\sqrt{n\ell_i \log n}) \cap \Omega(\sqrt{\sigma_i^2 / \log n})$; for any $1 \leq i < i_1^*$, $\delta_i = \omega(\sqrt{n\ell_i \log n})$.

Theorem 9.19 (Waiting times of SNF without Assumption 9.3). *Consider the multiserver-job system with n servers satisfying Assumption 9.1 and 9.2. Under SNF, for each $i \in [I]$, the steady-state expected waiting time for type i jobs, $\mathbb{E}[T_i^w(\infty)]^{\text{SNF}}$, is bounded by:*

$$\mathbb{E}[T_i^w(\infty)]^{\text{SNF}} \leq \frac{1}{\lambda_i} \frac{\mu_{\max} \sigma_i^2}{\ell_i(\delta_i - \ell_i)} \cdot (1 + o(1)) \quad \text{for } i \geq i_2^*, \quad (9.65)$$

$$\mathbb{E}[T_i^w(\infty)]^{\text{SNF}} \leq O\left(\frac{1}{\lambda_i} \frac{\sqrt{\sigma_i^2 \log n}}{\ell_i}\right) \quad \text{for } i_1^* \leq i < i_2^*, \quad (9.66)$$

$$\mathbb{E}[T_i^w(\infty)]^{\text{SNF}} \leq \exp\left(-\Omega\left(\frac{\delta_i^2}{n\ell_i}\right)\right) \quad \text{for } 1 \leq i < i_1^*. \quad (9.67)$$

Thus, the mean waiting time is bounded by

$$\mathbb{E}[T^w(\infty)]^{\text{SNF}} \leq \frac{1}{\lambda} \sum_{i=i_2^*}^I \frac{\mu_{\max} \sigma_i^2}{\ell_i(\delta_i - \ell_i)} \cdot (1 + o(1)) + \sum_{k=i_1^*}^{i_2^*-1} O\left(\frac{1}{\lambda} \frac{\sqrt{\sigma_k^2 \log n}}{\ell_k}\right). \quad (9.68)$$

The proof of this theorem is almost identical to the proof of Theorem 9.6. In fact, we have already derived the bounds (9.65), (9.66) and (9.67) in the proof of Theorem 9.6 in Section 9.9 without using Assumption 9.3.

Note that when we have Assumption 9.3, the second term on the right of (9.68) has a strictly smaller order than the first term (see the proof of

Theorem 9.6 in Section 9.9), so the upper bound order-wise matches the lower bound. Without the commonness assumption, it is unclear whether SNF achieves the optimal order. We leave a tighter bound on the mean waiting time under SNF for future work.

9.B Conditions for Applying Drift Method

9.B.1 Applying Drift Method to a Generic Markov Chain

In this subsection, we consider a continuous-time Markov chain $\{\mathbf{S}(t)\}_{t \geq 0}$ on a countable state space $\mathcal{S}$ with the generator G . Let $r_{ss'}$ denote the transition rate from state $\mathbf{s}$ to $\mathbf{s}'$, for $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$ such that $\mathbf{s} \neq \mathbf{s}'$. Let $r(\mathbf{s}) \triangleq \sum_{\mathbf{s}' \in \mathcal{S}, \mathbf{s}' \neq \mathbf{s}} r_{ss'}$. Then for any function $f: \mathcal{S} \rightarrow \mathbb{R}$, the drift of f equals:

$$Gf(\mathbf{s}) = \sum_{\mathbf{s}' \in \mathcal{S}, \mathbf{s}' \neq \mathbf{s}} r_{ss'} (f(\mathbf{s}') - f(\mathbf{s})). \quad (9.69)$$

We assume $\{\mathbf{S}(t)\}_{t \geq 0}$ has a stationary distribution; let $\mathbf{S}(\infty)$ be the random element with this stationary distribution.

We give a justification of the frequently used condition:

$$\mathbb{E}[Gf(\mathbf{S}(\infty))] = 0. \quad (9.70)$$

Lemma 9.20 below, which is a restatement of the Proposition 3 of Glynn and Zeevi [118], gives rigorous conditions under which the relation in (9.70) holds.

Lemma 9.20. *Consider the Markov chain $\{\mathbf{S}(t)\}_{t \geq 0}$ and a function $f: \mathcal{S} \rightarrow \mathbb{R}$. If for all $\mathbf{s} \in \mathcal{S}$,*

$$\mathbb{E}[r(\mathbf{S}(\infty)) \cdot |f(\mathbf{S}(\infty))|] < \infty,$$

then we have $\mathbb{E}[Gf(\mathbf{S}(\infty))] = 0$.

Lemma 9.20 implies that as long as $\sup_{\mathbf{s} \in \mathcal{S}} r(\mathbf{s}) < \infty$ and $\mathbb{E}[|f(\mathbf{S}(\infty))|] < \infty$, we have $\mathbb{E}[Gf(\mathbf{S}(\infty))] = 0$.

Lemma 9.21 below, which is a continuous-time analogue of a result in Hajek [134], provides some functions f that satisfy $\mathbb{E}[|f(\mathbf{S}(\infty))|] < \infty$.

Lemma 9.21. *Consider the Markov chain $\{\mathbf{S}(t)\}_{t \geq 0}$ and let $V : \mathcal{S} \rightarrow \mathbb{R}_+$ be a Lyapunov function. Assume that*

$$\sup_{\mathbf{s} \in \mathcal{S}} r(\mathbf{s}) < \infty \quad \text{and} \quad \sup_{\mathbf{s}, \mathbf{s}' \in \mathcal{S} : r_{\mathbf{s}\mathbf{s}'} > 0} |V(\mathbf{s}') - V(\mathbf{s})| < \infty,$$

and that there exists $\gamma > 0$ and $B \geq 0$, such that when $V(\mathbf{s}) \geq B$,

$$GV(\mathbf{s}) \leq -\gamma.$$

Then for any positive integer m ,

$$\mathbb{E}[V(\mathbf{S}(\infty))^m] < \infty.$$

9.B.2 Applying Drift Method to the Multiserver-Job System

Now we give conditions under which $E[Gf(\mathbf{X}, \mathbf{Z})] = 0$ for some $f : \mathbb{R}^I \rightarrow \mathbb{R}$ in multiserver-job system. Recall that

$$Gf(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^I \lambda_i (f(\mathbf{x} + \mathbf{e}_i) - f(\mathbf{x})) + \sum_{i=1}^I \mu_i z_i (f(\mathbf{x} - \mathbf{e}_i) - f(\mathbf{x})), \quad (9.71)$$

where $\mathbf{x}, \mathbf{z}$ are possible realizations of $\mathbf{X}, \mathbf{Z}$, $\mathbf{e}_i \in \mathbb{R}^I$ is the vector whose i th entry is 1 and all the other entries are 0. It is not hard to see from the formula of $Gf(\mathbf{x}, \mathbf{z})$ that the total transition rate of the multiserver-job system is bounded by $\mu_{\max} n$ in any state. Consequently, by Lemma 9.20, it suffices to have $\mathbb{E}[|f(\mathbf{X})|] < \infty$. In the next lemma, we show that $\mathbb{E}[|f(\mathbf{X})|] < \infty$ as long as f is bounded by some polynomial.

Lemma 9.22. *Consider the multiserver-job system with n servers under a δ' -work-conserving policy with $\delta' \leq \epsilon_0 \delta$, where $\epsilon_0 \in (0, 1)$ is the parameter used in Assumption 9.2. Then for any positive integer m , we have*

$$\mathbb{E} \left[\left(\sum_{i=1}^I \frac{\ell_i}{\mu_i} X_i(\infty) \right)^m \right] < \infty. \quad (9.72)$$

Proof of Lemma 9.22. We prove this lemma by applying Lemma 9.21. Let $V(\mathbf{x}) = \sum_{i=1}^I \frac{\ell_i}{\mu_i} x_i$. We first show that the drift $GV(\mathbf{x}, \mathbf{z})$ is negative when $V(\mathbf{x})$ is larger than a threshold. Observe that $GV(\mathbf{x}, \mathbf{z})$ can be bounded as

$$GV(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^I \left(\frac{\lambda_i \ell_i}{\mu_i} - \ell_i z_i \right)$$

$$\begin{aligned}
&\leq \sum_{i=1}^I \frac{\lambda_i \ell_i}{\mu_i} - \min \left(\sum_{i=1}^I \ell_i x_i, n - \delta' \right) \\
&\leq -\min \left(\sum_{i=1}^I \ell_i x_i - n + \delta, \delta - \delta' \right),
\end{aligned}$$

where the second inequality is due to the use of δ' -work-conserving policy. For any state such that $V(\mathbf{x}) > \frac{n}{\mu_{\min}} + 1$ with $\mu_{\min} = \min_{i \in [I]} \mu_i$, we have

$$\sum_{i=1}^I \ell_i x_i \geq \mu_{\min} \sum_{i=1}^I \frac{\ell_i}{\mu_i} x_i = \mu_{\min} V(\mathbf{x}) > n,$$

so $GV(\mathbf{x}, \mathbf{z}) < -\delta + \delta' \leq -\epsilon_0 \delta < 0$. Moreover, any transition in the system at most changes $V(\mathbf{x})$ by $\max_i \frac{\ell_i}{\mu_i}$, which is also a finite number. Applying Lemma 9.21, we get $\mathbb{E}[(\sum_{i=1}^I \frac{\ell_i}{\mu_i} X_i)^m] < \infty$ for any positive integer m . $\square$

9.C State-Space Concentration by Coupling with Infinite-Server Systems

In this section, we prove Lemma 9.11, the state-space concentration result stated in Section 9.6.

To prove Lemma 9.11, we define a multiserver-job systems with infinitely many servers, in which each job starts service immediately upon arrival. The job characteristics in this infinite-server system are identical to the n -server system under study: there are I job types; for each $i \in [I]$, the job type i has arrival rate λ_i , server need ℓ_i , and service rate μ_i . The infinite-server system serves as a lower-bounding system for the n -server system in the sense of stochastic dominance, as argued below.

Consider the following coupling between the infinite-server system and the n -server system under study: let these two systems have the same arrival sequences: whenever the n -server system has a job arrival, the infinite-server system have a job arrival with the same server need and service time. Since each job in the infinite-server system enters service immediately upon its arrival, each job leaves the infinite-server system no later than the corresponding job does in the n -server system. Consequently, the infinite-server system always has no more jobs than the n -server system does. Let

$X_i^{(\infty)}(t)$ be the number of type i jobs in the infinite-server system at time t , then by the coupling, $X_i^{(\infty)}(t) \leq X_i(t)$ almost surely for all $t \geq 0$ and $i \in [I]$. Therefore, in the steady-state, $X_i^{(\infty)}(\infty)$ is stochastically dominated by $X_i(\infty)$, i.e., $\mathbb{P}(X_i^{(\infty)}(\infty) > x) \leq \mathbb{P}(X_i(\infty) > x)$ for any $x \in \mathbb{R}$. We denote this relationship as $X_i^{(\infty)}(\infty) \leq_{st} X_i(\infty)$.

Next, we restate and prove Lemma 9.11.

Lemma 9.23 (Lemma 9.11). *Consider the multiserver-job system with n servers. For any nonnegative vector $c = (c_1, \dots, c_I) \in \mathbb{R}_+^I$ independent of n , let $c_{\max} = \max_{i \in [I]} c_i$, $\mu_{\max} = \max_{i \in [I]} \mu_i$ and let*

$$\Phi = \sum_{i=1}^I c_i \ell_i (X_i(\infty) - \bar{x}_i),$$

where $\bar{x}_i = \frac{\lambda_i}{\mu_i}$. Then we have the three bounds below.

(a) For any $K \geq 0$,

$$\mathbb{P}(\Phi \leq -K) \leq \exp\left(-\frac{K^2}{2c_{\max}^2 \mu_{\max} \sigma^2}\right). \quad (9.73)$$

(b) For any $\alpha \geq 0$ and $\beta \geq 0$ such that $\alpha\beta \geq c_{\max}^2 \mu_{\max} \sigma^2$ and any $j \geq 0$,

$$\mathbb{P}(\Phi \leq -\alpha - \beta j) \leq e^{-j}. \quad (9.74)$$

(c) Let $\Phi^- = \max\{-\Phi, 0\}$ to be the negative part of Φ . Then

$$\mathbb{E}[\Phi^-] \leq \sqrt{c_{\max}^2 \mu_{\max} \sigma^2}. \quad (9.75)$$

Proof of Lemma 9.11. Consider the infinite-server system defined at the beginning of this section. Based on the coupling argument, we have $X_i^{(\infty)}(\infty) \leq_{st} X_i(\infty)$ for any $i \in [I]$. We define the random variable $\Phi^{(\infty)}$ as

$$\Phi^{(\infty)} = \sum_{i=1}^I c_i \ell_i \left(X_i^{(\infty)}(\infty) - \frac{\lambda_i}{\mu_i} \right).$$

Because $c_i \ell_i \geq 0$, comparing the definitions of $\Phi^{(\infty)}$ and Φ , we get

$$\Phi^{(\infty)} \leq_{st} \Phi. \quad (9.76)$$

By (9.76), it suffices to prove the bounds in (a), (b) and (c) for $\Phi^{(\infty)}$.

We first prove the bound in (a). Note that $X_i^{(\infty)}(\infty)$ has the same distribution as the number of jobs in the M/M/ ∞ system with arrival rate λ_i and service rate μ_i . Consequently, $X_i^{(\infty)}(\infty)$ follows $\text{Poisson}(\frac{\lambda_i}{\mu_i})$ distribution by classic results [see, e.g., 37, p.249]. Furthermore, it not hard to see that $X_i^{(\infty)}(\infty)$'s are independent across different job types i . Therefore, for any $s \geq 0$,

$$\mathbb{E}[e^{-s\Phi^{(\infty)}}] = \exp\left(\sum_{i=1}^I \frac{\lambda_i}{\mu_i} (e^{-c_i \ell_i s} - 1 + c_i \ell_i s)\right).$$

By Markov inequality, for any $s \geq 0$,

$$\mathbb{P}(\Phi^{(\infty)} \leq -K) = \mathbb{P}(e^{-s\Phi^{(\infty)}} \geq e^{sK}) \leq e^{-sK} \mathbb{E}[e^{-s\Phi^{(\infty)}}].$$

To upper-bound $\mathbb{E}[e^{-s\Phi^{(\infty)}}]$, observe that $e^{-s} \leq 1 - s + \frac{s^2}{2}$ for any $s \geq 0$. Then we have

$$\begin{aligned} \mathbb{E}[e^{-s\Phi^{(\infty)}}] &= \exp\left(\sum_{i=1}^I \frac{\lambda_i}{\mu_i} (e^{-c_i \ell_i s} - 1 + c_i \ell_i s)\right) \leq \exp\left(\frac{1}{2} \sum_{i=1}^I \frac{\lambda_i}{\mu_i} (c_i \ell_i s)^2\right) \\ &= \exp\left(\frac{1}{2} \sum_{i=1}^I c_i^2 \mu_i \cdot \frac{\lambda_i \ell_i^2}{\mu_i^2} s^2\right) \leq \exp\left(\frac{1}{2} c_{\max}^2 \mu_{\max} \sigma^2 s^2\right), \end{aligned}$$

where we have used the facts that $c_i^2 \leq c_{\max}^2$ and $\mu_i \leq \mu_{\max}$ for any $i \in [I]$. Therefore, for any $s \geq 0$,

$$\mathbb{P}(\Phi^{(\infty)} \leq -K) \leq \exp\left(\frac{1}{2} c_{\max}^2 \mu_{\max} \sigma^2 s^2 - Ks\right).$$

Taking $s = \frac{K}{c_{\max}^2 \mu_{\max} \sigma^2}$ and combining the fact that $\Phi^{(\infty)} \leq_{st} \Phi$, we get

$$\mathbb{P}(\Phi \leq -K) \leq \mathbb{P}(\Phi^{(\infty)} \leq -K) \leq \exp\left(-\frac{K^2}{2c_{\max}^2 \mu_{\max} \sigma^2}\right), \quad (9.77)$$

which proves the bound in (a).

Next, we prove the bound in (b). Let α and β be any two non-negative numbers such that $\alpha\beta \geq c_{\max}^2 \mu_{\max} \sigma^2$ and let $j \geq 0$. By the bound in (a) with $K = \alpha + \beta j$, we have

$$\begin{aligned} \mathbb{P}(\Phi^{(\infty)} \leq -\alpha - \beta j) &\leq \exp\left(-\frac{\alpha^2 + 2\alpha\beta j + \beta^2 j^2}{2c_{\max}^2 \mu_{\max} \sigma^2}\right) \\ &\leq \exp\left(-\frac{\alpha\beta j}{c_{\max}^2 \mu_{\max} \sigma^2}\right) \\ &\leq e^{-j}. \end{aligned}$$

Because $\Phi^{(\infty)} \leq_{st} \Phi$, we have proved the bound in (b).

To prove the bound in (c), observe that

$$\begin{aligned} \mathbb{E}\left[(\Phi^{(\infty)})^-\right] &\leq \mathbb{E}\left[|\Phi^{(\infty)}|\right] \leq \sqrt{\mathbb{E}\left[(\Phi^{(\infty)})^2\right]}, \\ \mathbb{E}\left[(\Phi^{(\infty)})^2\right] &= \sum_{i=1}^I c_i^2 \ell_i^2 \mathbb{E}\left[\left(X_i^{(\infty)}(\infty) - \frac{\lambda_i}{\mu_i}\right)^2\right] = \sum_{i=1}^I c_i^2 \ell_i^2 \frac{\lambda_i}{\mu_i} \leq c_{\max}^2 \mu_{\max} \sigma^2, \end{aligned}$$

where we have used the fact that the mean and variance of $X_i^{(\infty)}(\infty)$ are both $\frac{\lambda_i}{\mu_i}$, since $X_i^{(\infty)}(\infty)$ follows the $\text{Poisson}\left(\frac{\lambda_i}{\mu_i}\right)$ distribution. Because $s \mapsto s^-$ is a non-increasing function for real number s , we have

$$\mathbb{E}\left[\Phi^-\right] \leq \mathbb{E}\left[(\Phi^{(\infty)})^-\right] \leq \sqrt{c_{\max}^2 \mu_{\max} \sigma^2}. \quad (9.78)$$

This finishes the proof of the bound (c). $\square$

9.D Proofs of the Lemmas for Theorem 9.4

In this section, we prove Lemma 9.12 and Lemma 9.13, which are used in the proof of Theorem 9.4 in Section 9.7. The lemmas are proved using a sample-path coupling argument, so before showing the proofs, we first give a construction of the sample paths of a multiserver-job system. We index all the jobs by their order of arrivals. Let $A_0(k)$, $S_0(k)$, $C(k)$ for $k = 1, 2, \dots$ be three sequences of independent random variables, where $A_0(k)$ follows $\exp(\lambda)$ distribution, $S_0(k)$ follows $\exp(1)$ distribution, and $C(k) = i$ with probability λ_i/λ for $i \in [I]$. For the k -th job, we let it arrive to the system at the time

$A(k) = \sum_{j=1}^k A_0(j)$, with service time $S(k) = S_0(k)/\mu_{C(k)}$ and server need $L(k) = \ell_{C(k)}$.

We can reproduce a sample path under FCFS or Modified-FCFS given the realization of the jobs. Let the k -th job's waiting time be $W(k)$, then the k -th job starts its service at time $A(k) + W(k)$, occupying $L(k)$ servers and leaves the system at $A(k) + W(k) + S(k)$.

We couple the sample paths of the original system, the upper-bounding system, and the lower-bounding system by making them share the same realization of jobs, i.e., the k -th job of the three system arrive at the same time, with the same service times and server needs.

Now we are ready to present the proof. Note that quantities with superscript (U) corresponds to the upper-bounding system, while quantities with (L) belongs to the lower-bounding system.

Lemma 9.24 (Lemma 9.12). *For both the upper-bounding system and the lower-bounding system, we have*

$$\mathbb{E}[Q_i^{(U)}(\infty)] = \frac{\lambda_i}{\lambda} \mathbb{E}[Q_\Sigma^{(U)}(\infty)] \quad \forall i \in [I], \quad (9.51 \text{ restated})$$

$$\mathbb{E}[Q_i^{(L)}(\infty)] = \frac{\lambda_i}{\lambda} \mathbb{E}[Q_\Sigma^{(L)}(\infty)] \quad \forall i \in [I], \quad (9.52 \text{ restated})$$

where $Q_i^{(U)}(\infty)$ and $Q_i^{(L)}(\infty)$ denote the steady-state queue lengths of type i jobs in the upper and lower-bounding systems, respectively; $Q_\Sigma^{(U)}(\infty) \triangleq \sum_{i=1}^I Q_i^{(U)}(\infty)$ and $Q_\Sigma^{(L)}(\infty) \triangleq \sum_{i=1}^I Q_i^{(L)}(\infty)$ denote the total queue lengths in the steady state.

Proof of Lemma 9.12. In the upper-bounding system, the queue length of type i jobs at time t is given by $Q_i^{(U)}(t) = \sum_{k=1}^\infty \mathbf{1}\{C(k) = i, A(k) \leq t, A(k) + W^{(U)}(k) > t\}$. Taking expectation, we get

$$\mathbb{E}[Q_i^{(U)}(t)] = \sum_{k=1}^\infty \frac{\lambda_i}{\lambda} \mathbb{P}(A(k) \leq t, A(k) + W^{(U)}(k) > t \mid C(k) = i). \quad (9.79)$$

Let the σ -algebra $\mathcal{F}_k = \sigma((A_0(j), S_0(j), C(j)) \mid j \leq k)$. From the definitions, it is easy to see that

$$(S(j), A(j), L(j), W^{(U)}(j)) \in \mathcal{F}_k \quad \text{for } j \leq k.$$

Observe that the following two events are identical,

$$\{A(k) \leq t, A(k) + W^{(U)}(k) > t\} = \left\{A(k) \leq t, \sum_{j \in \mathcal{Z}^{(U)}(t; k-1)} L(j) > n - \ell_{\max}\right\}, \quad (9.80)$$

where $\mathcal{Z}^{(U)}(t; k) \triangleq \{j \mid j \leq k, A(j) + W^{(U)}(j) + S(j) > t\}$ is the set of jobs with index $j \leq k$ who has not finished service by time t in the upper-bounding system. Moreover, the RHS event is obviously in the σ -algebra generated by $\mathcal{F}_{k-1}$ and $A_0(k)$, which makes it independent of $C(k)$. This helps us get rid of the conditioning on the RHS of (9.79):

$$\mathbb{E}[Q_i^{(U)}(t)] = \frac{\lambda_i}{\lambda} \sum_{k=1}^{\infty} \mathbb{P}(A(k) \leq t, A(k) + W^{(U)}(k) > t), \quad (9.81)$$

$$\mathbb{E}[Q_{\Sigma}^{(U)}(t)] = \sum_{i=1}^I \mathbb{E}[Q_i^{(U)}(t)] = \sum_{k=1}^{\infty} \mathbb{P}(A(k) \leq t, A(k) + W^{(U)}(k) > t). \quad (9.82)$$

Dividing the above two equalities, we get (9.51).

The proof of (9.52) is almost verbatim with the superscript (U) replaced by (L) , and equation (9.80) replaced by

$$\{A(k) \leq t, A(k) + W^{(L)}(j) > t\} = \left\{A(k) \leq t, \sum_{j \in \mathcal{Z}^{(L)}(t; k-1)} L^{(L)}(j) > n\right\}, \quad (9.83)$$

where $\mathcal{Z}^{(L)}(t; k) \triangleq \{j \mid j \leq k, A(j) + W^{(L)}(j) + S(j) > t\}$. $\square$

Lemma 9.25 (Lemma 9.13).

$$\mathbb{E}[Q_i^{(L)}(\infty)] \leq \mathbb{E}[Q_i(\infty)] \leq \mathbb{E}[Q_i^{(U)}(\infty)] \quad \forall i \in [I], \quad (9.53 \text{ restated})$$

where $Q_i^{(U)}(\infty)$ and $Q_i^{(L)}(\infty)$ denote the steady-state queue lengths of type i jobs in the upper and lower-bounding systems, respectively.

Proof of Lemma 9.13. We prove the following claim by doing induction on k :

$$W(k) \leq W^{(U)}(k) \quad \forall k = 1, 2, \dots \quad (9.84)$$

Case $k = 1$ is trivial because $W(1) = W^{(U)}(1) = 0$. Suppose that we have proved the cases of $1, 2, \dots, k$. For case $k + 1$, to get a contradiction, suppose

that $W(k+1) > W^{(U)}(k+1)$. Let $T = A(k+1) + W^{(U)}(k+1)$, then the $(k+1)$ -th job starts service at time T in the upper-bounding system, but does not start service until $A(k+1) + W(k+1)$ in the original system. This implies that at the moment before time T , the original system has more than $n - L(k+1)$ busy servers, while the upper-bounding system has at most $n - \ell_{\max}$ busy servers. Because $n - L(k+1) \geq n - \ell_{\max}$,

$$\sum_{j \in \mathcal{Z}(T; k)} L(j) > \sum_{j \in \mathcal{Z}^{(U)}(T; k)} L(j), \quad (9.85)$$

where $\mathcal{Z}(T; k) = \{j \mid j \leq k, A(j) + W(j) + S(j) > T\}$, $\mathcal{Z}^{(U)}(T; k) = \{j \mid j \leq k, A(j) + W^{(U)}(j) + S(j) > T\}$. However, by induction hypothesis, $\mathcal{Z}(T; k) \subseteq \mathcal{Z}^{(U)}(T; k)$, which contradicts (9.85). Therefore $W(k+1) \leq W^{(U)}(k+1)$. By induction, (9.84) is true.

Observe that the number of type i jobs in the queues at time t is

$$Q_i(t) = \sum_{k=1}^{\infty} \mathbf{1}\{C(k) = i, A(k) \leq t, A(k) + W(k) > t\}, \quad (9.86)$$

$$Q_i^{(U)}(t) = \sum_{k=1}^{\infty} \mathbf{1}\{C(k) = i, A(k) \leq t, A(k) + W^{(U)}(k) > t\}. \quad (9.87)$$

by (9.84), $Q_i(t) \leq Q_i^{(U)}(t)$ for any t, i . Therefore, in the steady state, we have $\mathbb{E}[Q_i] \leq \mathbb{E}[Q_i^{(U)}]$ for any i .

For the other inequality in the lemma, we perform a similar argument. We prove the following claim by doing induction on k :

$$W(k) \geq W^{(L)}(k) \quad \forall k = 1, 2, \dots \quad (9.88)$$

Case $k = 1$ is trivial because $W(1) = W^{(L)}(1) = 0$. Suppose we have proved case $1, 2, \dots, k$. For case $k+1$, to get a contradiction, suppose $W(k+1) < W^{(L)}(k+1)$. Let $T = A(k+1) + W(k+1)$, then $T < A(k+1) + W^{(L)}(k+1)$. This implies that at time T , the $(k+1)$ -th job in the original system starts its service, while the $(k+1)$ -th job in the lower-bounding system is still waiting at the head of line. According to the policy they use, at the moment right before T , the original system has at most $n - L(k+1)$ busy servers, while the lower-bounding system has at least n busy servers. Therefore,

$$\sum_{j \in \mathcal{Z}(T; k)} L(j) < \sum_{j \in \mathcal{Z}^{(L)}(T; k)} L(j), \quad (9.89)$$

where $\mathcal{Z}^{(L)}(T; k) = \{j \mid j \leq k, A(j) + W^{(L)}(j) + S(j) > T\}$. However, by induction hypothesis, $\mathcal{Z}^{(L)}(T; k) \subseteq \mathcal{Z}(T; k)$, which contradicts (9.89). Therefore $W(k+1) \geq W^{(L)}(k+1)$ and (9.88) is proved.

Because

$$Q_i^{(L)}(t) = \sum_{k=1}^{\infty} \mathbf{1}\{C(k) = i, A(k) \leq t, A(k) + W^{(L)}(k) > t\}, \quad (9.90)$$

combined with (9.86) and (9.88), it follows that $Q_i(t) \geq Q_i^{(L)}(t)$ for all i, t . Therefore, in the steady state, we have $\mathbb{E}[Q_i^{(L)}] \leq \mathbb{E}[Q_i]$ for any i . This finishes the proof of Lemma 9.13. $\square$

9.E Proofs of the Lemmas for Theorem 9.6 (Mean Waiting Time under SNF)

9.E.1 A Tail Bound for Generic Markov Chains Based on Lyapunov Conditions

In this subsection, we prove a tail bound in Markov chains based on a certain Lyapunov drift condition. Consider a continuous-time Markov chain $\{\mathbf{S}(t)\}_{t \geq 0}$ on a countable state space $\mathcal{S}$ with generator G . Recall that $r_{ss'}$ denotes the transition rate from state $\mathbf{s}$ to $\mathbf{s}'$, and let $r(\mathbf{s}) \triangleq \sum_{\mathbf{s}' \in \mathcal{S}, \mathbf{s}' \neq \mathbf{s}} r_{ss'}$. For any function $f: \mathcal{S} \rightarrow \mathbb{R}$, the drift of f is equal to

$$Gf(\mathbf{s}) = \sum_{\mathbf{s}' \in \mathcal{S}, \mathbf{s}' \neq \mathbf{s}} r_{ss'} (f(\mathbf{s}') - f(\mathbf{s})). \quad (9.91)$$

Recall that $\mathbf{S}(\infty)$ denotes the random element that follows the stationary distribution.

For a Lyapunov function $V: \mathcal{S} \rightarrow \mathbb{R}_+$, Lemma 9.26 below gives a tail bound on $V(\mathbf{S}(\infty))$. Lemma 9.26 is a straightforward consequence of Lemma A.1 in Weng and Wang [276], which slightly generalizes the well-known Lyapunov-based tail bounds [see, e.g., 33, 272].

Lemma 9.26 (Lyapunov upper bound). *Consider the Markov chain $\{\mathbf{S}(t)\}_{t \geq 0}$ and let $V: \mathcal{S} \rightarrow \mathbb{R}_+$ be a Lyapunov function such that $\mathbb{E}[V(\mathbf{S}(\infty))] < \infty$. Assume that*

$$\sup_{\mathbf{s} \in \mathcal{S}} r(\mathbf{s}) < \infty,$$

$$v_{\max} \triangleq \sup_{\mathbf{s}, \mathbf{s}' \in \mathcal{S}: r_{\mathbf{s}\mathbf{s}'} > 0} |V(\mathbf{s}') - V(\mathbf{s})| < \infty,$$

$$f_{\max} \triangleq \sup_{\mathbf{s}} \sum_{\mathbf{s}': V(\mathbf{s}') > V(\mathbf{s})} r_{\mathbf{s}\mathbf{s}'} (V(\mathbf{s}') - V(\mathbf{s})) < \infty.$$

If there exists $B \geq 0$, $\gamma > 0$, $\xi > 0$ and a subset $\mathcal{E} \subseteq \mathcal{S}$ such that for all $\mathbf{s}$ with $V(\mathbf{s}) > B + v_{\max}$,

$$GV(\mathbf{s}) \leq \begin{cases} -\gamma, & \forall \mathbf{s} \in \mathcal{E}, \\ \xi, & \forall \mathbf{s} \in \mathcal{S} \setminus \mathcal{E}, \end{cases}$$

then for any real number $j \geq 0$,

$$\mathbb{P}\left(V(\mathbf{S}(\infty)) \geq B + 2v_{\max} \cdot \left(\left(\frac{f_{\max}}{\gamma} + 2\right)j + 1\right)\right) \leq e^{-j + \left(\frac{\xi}{\gamma} + 1\right)} \mathbb{P}(\mathbf{S}(\infty) \in \mathcal{E}^c). \quad (9.92)$$

Proof of Lemma 9.26. According to Lemma 1 of Bertsimas et al. [33], for any integer $j' \geq 0$,

$$\begin{aligned} \mathbb{P}(V(\mathbf{S}(\infty)) > B + 2v_{\max} \cdot j') \\ &\leq \left(\frac{f_{\max}}{f_{\max} + \gamma}\right)^{j'} + \left(\frac{\xi}{\gamma} + 1\right) \mathbb{P}(\mathbf{S}(\infty) \in \mathcal{E}^c) \\ &\leq \exp\left(-\frac{\gamma j'}{f_{\max} + \gamma}\right) + \left(\frac{\xi}{\gamma} + 1\right) \mathbb{P}(\mathbf{S}(\infty) \in \mathcal{E}^c). \end{aligned}$$

For any $j \geq 0$, we set $j' = \lceil (\frac{f_{\max}}{\gamma} + 1)j \rceil$, which is no more than $((f_{\max}/\gamma + 2)j + 1)$. Then (9.92) follows. $\square$

9.E.2 Tail Bounds for the Multiserver-Job System

With the general tool Lemma 9.26 ready, we can now start proving Lemma 9.15, Lemma 9.16 and Lemma 9.17, which are needed for proving Theorem 9.6. As a by-product, we will also prove the bound on queueing probability stated in Corollary 9.18.

Lemma 9.27 (Lemma 9.15). *Consider the multiserver-job system with n servers satisfying $\ell_{\max} \leq \epsilon_0 \delta$ (Assumption 9.2). Letting $\bar{x}_i = \frac{\lambda_i}{\mu_i}$, under any $\ell_{\max}$ -work-conserving policy, the normalized work has the following tail bound:*

for any ϵ such that $0 < \epsilon < \epsilon_0$, there exists $\alpha_1 = \frac{2\epsilon\delta}{\mu_{\min}} + \Theta\left(\frac{n\ell_{\max}}{\delta} \log n\right)$ and $\beta_1 = \Theta\left(\frac{n\ell_{\max}}{\delta}\right)$ such that for any $j \geq 0$,

$$\mathbb{P}\left(\sum_{i=1}^I \frac{\ell_i}{\mu_i} (X_i(\infty) - \bar{x}_i) \geq \alpha_1 + \beta_1 \cdot j\right) \leq e^{-j}. \quad (9.59 \text{ restated})$$

Lemma 9.28 (Lemma 9.16). *Consider the multiserver-job system with n servers satisfying $\ell_{\max} \leq \epsilon_0\delta$ (Assumption 9.2). Letting $\bar{x}_i = \frac{\lambda_i}{\mu_i}$, under any $\ell_{\max}$ -work-conserving policy, the total server need has the following tail bound: there exists α_2 and β_2 with $\alpha_2 = \frac{\delta}{2} + \Theta\left(\frac{n\ell_{\max}}{\delta} \log n\right)$ and $\beta_2 = \Theta\left(\frac{n\ell_{\max}}{\delta}\right)$ such that for any $j \geq 0$,*

$$\mathbb{P}\left(\sum_{i=1}^I \ell_i (X_i(\infty) - \bar{x}_i) \geq \alpha_2 + \beta_2 \cdot j\right) \leq e^{-j}. \quad (9.60 \text{ restated})$$

Proof of Lemma 9.15. The idea of the proof is applying Lemma 9.26 to the Lyapunov function $V_1(\mathbf{x}) = \sum_{i=1}^I \frac{\ell_i}{\mu_i} x_i$, where $\mathbf{x} \in \mathbb{R}^I$ denotes possible realizations of $\mathbf{X}$. To check the conditions of Lemma 9.26, consider the drift of $V_1(\mathbf{x})$:

$$GV_1(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^I \lambda_i (V_1(\mathbf{x} + \mathbf{e}_i) - V_1(\mathbf{x})) + \sum_{i=1}^I \mu_i z_i (V_1(\mathbf{x} - \mathbf{e}_i) - V_1(\mathbf{x})). \quad (9.93)$$

From the formula of $GV_1(\mathbf{x}, \mathbf{z})$, we can immediately see that the two parameters $v_{\max}$ and $f_{\max}$ in Lemma 9.26 are $v_{\max} = \max_{i \in [I]} \frac{\ell_i}{\mu_i}$ and $f_{\max} = \sum_{i=1}^I \frac{\lambda_i \ell_i}{\mu_i}$. These two parameters satisfy $v_{\max} \leq \frac{\ell_{\max}}{\mu_{\min}}$ and $f_{\max} = n - \delta \leq n$.

In the rest of the prove, we show that when $V_1(\mathbf{x})$ is larger than the threshold $\sum_{i=1}^I \frac{\ell_i}{\mu_i} \bar{x}_i + B_1$, i.e.,

$$\sum_{i=1}^I \frac{\ell_i}{\mu_i} x_i = V_1(\mathbf{x}) > \sum_{i=1}^I \frac{\ell_i}{\mu_i} \bar{x}_i + B_1,$$

for some $B_1 \geq 0$, there exists $\gamma, \xi > 0$, such that with high probability, we have $GV_1(\mathbf{x}, \mathbf{z}) < -\gamma$, and in the worst case, we have $GV_1(\mathbf{x}, \mathbf{z}) < \xi$. We prove this by finding a suitable state-space concentration. To see what we need, we write out the form of $GV_1(\mathbf{x}, \mathbf{z})$ explicitly:

$$GV_1(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^I \left(\frac{\lambda_i \ell_i}{\mu_i} - \ell_i z_i \right) \quad (9.94)$$

$$\begin{aligned}
&\leq \sum_{i=1}^I \frac{\lambda_i \ell_i}{\mu_i} - \min \left(\sum_{i=1}^I \ell_i x_i, n - \ell_{\max} \right) \\
&= -\min \left(\sum_{i=1}^I \ell_i (x_i - \bar{x}_i), \delta - \ell_{\max} \right) \\
&\leq -\min \left(\sum_{i=1}^I \ell_i (x_i - \bar{x}_i), \epsilon_0 \delta \right), \tag{9.95}
\end{aligned}$$

where the first inequality is because the policy is $\ell_{\max}$ -work-conserving, and the last inequality used the fact that $\ell_{\max} \leq \epsilon_0 \delta$. The last expression suggests that we need a high probability lower bound on $\sum_{i=1}^I \ell_i X_i$.

We apply Lemma 9.11 to construct a subset $\mathcal{E}_3(j_1) \subseteq \mathbb{R}^I$ such that $\mathbf{X} \in \mathcal{E}_3(j_1)$ with high probability, where j_1 is a non-negative number. We set $c_i = \frac{1}{\mu_{\min}} - \frac{1}{\mu_i}$, $\alpha = \frac{\epsilon \delta}{\mu_{\min}}$, $\beta = \frac{\mu_{\max} n \ell_{\max}}{\mu_{\min}^2 \epsilon \delta}$, $j = 2 \log n + j_1$ in Lemma 9.11. Because $\alpha \beta \geq \frac{\mu_{\max} n \ell_{\max}}{\mu_{\min}^3} \geq c_{\max}^2 \mu_{\max} \sigma^2$, we have

$$\mathbb{P} \left(\sum_{i=1}^I \left(\frac{1}{\mu_{\min}} - \frac{1}{\mu_i} \right) \ell_i (X_i - \bar{x}_i) \leq -\alpha - \beta j_1 \right) \leq e^{-j_1}. \tag{9.96}$$

Letting $\alpha_3 = \alpha + 2\beta \log n$, $\beta_3 = \beta$, we rewrite the last inequality as

$$\mathbb{P} \left(\sum_{i=1}^I \left(\frac{1}{\mu_{\min}} - \frac{1}{\mu_i} \right) \ell_i (X_i - \bar{x}_i) \leq -\alpha_3 - \beta_3 j_1 \right) \leq \frac{1}{n^2} e^{-j_1}. \tag{9.97}$$

Note that $\beta_3 = \Theta \left(\frac{n \ell_{\max}}{\delta} \right)$ and $\alpha_3 = \frac{\epsilon \delta}{\mu_{\min}} + \Theta \left(\frac{n \ell_{\max}}{\delta} \log(n) \right)$. Define the set $\mathcal{E}_3(j_1)$ as

$$\mathcal{E}_3(j_1) = \left\{ \mathbf{x} \in \mathbb{R}^I \mid \sum_{i=1}^I \left(\frac{1}{\mu_{\min}} - \frac{1}{\mu_i} \right) \ell_i (x_i - \bar{x}_i) > -\alpha_3 - \beta_3 j_1 \right\}.$$

Then (9.97) implies that $\mathbb{P}(\mathcal{E}_3(j_1)) \geq 1 - \frac{1}{n^2} e^{-j_1}$.

Next, we divide the state space into two parts based on $\mathcal{E}_3(j_1)$ and bound $GV(\mathbf{x}, \mathbf{z})$ on each part separately. When $V_1(\mathbf{x}) > \sum_{i=1}^I \frac{\ell_i}{\mu_i} \bar{x}_i + B_1$ and $\mathbf{x} \in \mathcal{E}_3(j_1)$, we have

$$\sum_{i=1}^I \frac{\ell_i}{\mu_i} (x_i - \bar{x}_i) > B_1,$$

$$\sum_{i=1}^I \left(\frac{1}{\mu_{\min}} - \frac{1}{\mu_i} \right) \ell_i(x_i - \bar{x}_i) > -\alpha_3 - \beta_3 j_1.$$

Adding up the above two inequality and choosing $B_1 = \alpha_3 + \beta_3 j_1 + \frac{\epsilon\delta}{\mu_{\min}}$ yield

$$\sum_{i=1}^I \ell_i(x_i - \bar{x}_i) \geq \epsilon\delta. \quad (9.98)$$

By (9.95) and the fact that $\epsilon < \epsilon_0$, we conclude that $GV(\mathbf{x}, \mathbf{z}) \leq -\epsilon\delta$.

When $V_1(\mathbf{x}) > \sum_{i=1}^I \frac{\ell_i}{\mu_i} \bar{x}_i + B_1$ and $\mathbf{x} \notin \mathcal{E}_3(j_1)$, it follows from (9.94) that $GV(\mathbf{x}, \mathbf{z}) \leq n$.

Now we can apply Lemma 9.26 with $B = \sum_{i=1}^I \frac{\ell_i}{\mu_i} \bar{x}_i + B_1$, $\gamma = -\epsilon\delta$, $\xi = n$, $v_{\max} = \max_{i \in [I]} \frac{\ell_i}{\mu_i}$, $f_{\max} = \sum_{i=1}^I \frac{\lambda_i \ell_i}{\mu_i}$ and $\mathcal{E} = \mathcal{E}_3(j_1)$ to get

$$\begin{aligned} \mathbb{P} \left(\sum_{i=1}^I \frac{\ell_i}{\mu_i} (X_i - \bar{x}_i) \geq \alpha_3 + \beta_3 j_1 + \frac{\epsilon\delta}{\mu_{\min}} + 2v_{\max} \left(\left(\frac{f_{\max}}{\epsilon\delta} + 2 \right) j_2 + 1 \right) \right) \\ \leq e^{-j_2} + \left(\frac{n}{\epsilon\delta} + 1 \right) \frac{1}{n^2} e^{-j_1}. \end{aligned} \quad (9.99)$$

which holds for any real number $j_1, j_2 \geq 0$. (9.99) can be further simplified to

$$\begin{aligned} \mathbb{P} \left(\sum_{i=1}^I \frac{\ell_i}{\mu_i} (X_i - \bar{x}_i) \geq \alpha_3 + \beta_3 j_1 + \frac{\epsilon\delta}{\mu_{\min}} + 2 \frac{\ell_{\max}}{\mu_{\min}} \left(\frac{(2\epsilon_0 + 1)n}{\epsilon\delta} j_2 + 1 \right) \right) \\ \leq e^{-j_2} + e^{-j_1}. \end{aligned} \quad (9.100)$$

where we have used the facts that $v_{\max} \leq \frac{\ell_{\max}}{\mu_{\min}}$, $f_{\max} \leq n$, $n > \delta$, $\epsilon < \epsilon_0$, and $\left(\frac{n}{\epsilon\delta} + 1 \right) \frac{1}{n^2} \leq 1$ which holds for n large enough. Taking both j_1 and j_2 to be $j + \log(2)$, and letting $\beta_1 = \beta_3 + \frac{(4\epsilon_0 + 2)n\ell_{\max}}{\mu_{\min}\epsilon\delta}$, $\alpha_1 = \alpha_3 + \frac{\epsilon\delta}{\mu_{\min}} + \frac{2\ell_{\max}}{\mu_{\min}} + \beta_1 \log(2)$, we get the inequality in the lemma statement. Note that α_1 and β_1 defined here satisfy $\beta_1 = \Theta\left(\frac{n\ell_{\max}}{\delta}\right)$ and $\alpha_1 = \frac{2\epsilon\delta}{\mu_{\min}} + \Theta\left(\frac{n\ell_{\max}}{\delta} \log(n)\right)$. This proves Lemma 9.15. $\square$

Proof of Lemma 9.16. We prove the result by applying Lemma 9.26 to the Lyapunov function $V_2(\mathbf{x}) = \sum_{i=1}^I \ell_i x_i$. To check the conditions of Lemma 9.26, consider the drift of $V_2(\mathbf{x})$:

$$GV_2(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^I \lambda_i (V_2(\mathbf{x} + \mathbf{e}_i) - V_2(\mathbf{x})) + \sum_{i=1}^I \mu_i z_i (V_2(\mathbf{x} - \mathbf{e}_i) - V_2(\mathbf{x})). \quad (9.101)$$

From the formula of $GV_2(\mathbf{x}, \mathbf{z})$, we can immediately see that the two parameters in Lemma 9.26, $v_{\max}$ and $f_{\max}$, can be taken as $v_{\max} = \ell_{\max}$ and $f_{\max} = \sum_{i=1}^I \lambda_i \ell_i \leq \mu_{\max} n$.

In the rest of the proof, we show that when $V_2(\mathbf{x})$ is larger than the threshold $\sum_{i=1}^I \ell_i \bar{x}_i + B_2$, i.e.,

$$\sum_{i=1}^I \ell_i x_i = V_2(\mathbf{x}) > \sum_{i=1}^I \ell_i \bar{x}_i + B_2, \quad (9.102)$$

for some non-negative B_2 , there exists some $\gamma, \xi > 0$, such that with high probability, we have $GV_2(\mathbf{x}, \mathbf{z}) < -\gamma$, and in the worst case, we have $GV_2(\mathbf{x}, \mathbf{z}) < \xi$. We bound $GV_2(\mathbf{x}, \mathbf{z})$ using a state-space concentration argument. To do this, we write out the form of $GV_2(\mathbf{x}, \mathbf{z})$ explicitly and bound it as below:

$$\begin{aligned} GV_2(\mathbf{x}, \mathbf{z}) &= \sum_{i=1}^I (\lambda_i \ell_i - \mu_i \ell_i z_i) \\ &= \sum_{i=1}^I (\lambda_i \ell_i + (\mu_{\max} - \mu_i) \ell_i z_i) - \sum_{i=1}^I \mu_{\max} \ell_i z_i \\ &\leq \sum_{i=1}^I (\lambda_i \ell_i + (\mu_{\max} - \mu_i) \ell_i x_i) - \mu_{\max} \min \left(\sum_{i=1}^I \ell_i x_i, n - \ell_{\max} \right) \\ &= -\mu_{\max} \min \left(\sum_{i=1}^I \frac{\mu_i}{\mu_{\max}} \ell_i (x_i - \bar{x}_i), \right. \\ &\quad \left. (\delta - \ell_{\max}) - \sum_{i=1}^I \left(1 - \frac{\mu_i}{\mu_{\max}} \right) \ell_i (x_i - \bar{x}_i) \right), \end{aligned} \quad (9.104)$$

where the inequality is due to $z_i \leq x_i$ and the $\ell_{\max}$ -work-conserving condition. The last expression suggests that in order to apply Lemma 9.26, we need to lower bound $\sum_{i=1}^I \frac{\mu_i}{\mu_{\max}} \ell_i (X_i - \bar{x}_i)$ and upper bound $\sum_{i=1}^I (1 - \frac{\mu_i}{\mu_{\max}}) \ell_i (X_i - \bar{x}_i)$ with high probability.

Now we construct a subset $\mathcal{E} \subseteq \mathbb{R}^I$ such that $\mathbf{X} \in \mathcal{E}$ with high probability, and $f(\mathbf{x})$ has negative drift whenever $\mathbf{x} \in \mathcal{E}$ and $f(\mathbf{x}) > B_2$ for some $B_2 \geq 0$. The construction of $\mathcal{E}$ is based on the high probability sets implied

by Lemma 9.26 and Lemma 9.15 that we just proved. For a fixed $j_1 \geq 0$, define the set

$$\mathcal{E}_4(j_1) = \left\{ \mathbf{x} \in \mathbb{R}^I \mid \sum_{i=1}^I \left(\frac{\mu_{\max}}{\mu_i} + \frac{\mu_i}{\mu_{\max}} - 2 \right) \ell_i(x_i - \bar{x}_i) > -\alpha_4 - \beta_4 j_1 \right\},$$

where

$$\beta_4 = \left(\frac{\mu_{\max}}{\mu_{\min}} + \frac{\mu_{\min}}{\mu_{\max}} - 2 \right)^2 \frac{\mu_{\max}}{\mu_{\min}} \cdot \frac{n\ell_{\max}}{\epsilon\delta}, \quad \alpha_4 = \epsilon\delta + 2\beta_4 \log n.$$

Note that $\beta_4 = \Theta\left(\frac{n\ell_{\max}}{\delta}\right)$, $\alpha_4 = \epsilon\delta + \Theta\left(\frac{n\ell_{\max}}{\delta} \log n\right)$. Let

$$\mathcal{E}_1(j_1) = \left\{ \mathbf{x} \in \mathbb{R}^I \mid \sum_{i=1}^I \frac{\ell_i}{\mu_i} (x_i - \bar{x}_i) < \alpha_1 + \beta_1 j_1 \right\},$$

where $\beta_1 = \Theta\left(\frac{n\ell_{\max}}{\delta}\right)$ and $\alpha_1 = \frac{2\epsilon\delta}{\mu_{\min}} + \Theta\left(\frac{n\ell_{\max}}{\delta} \log n\right)$ are given in the tail bound of Lemma 9.15. Define the set $\mathcal{E}$ and non-negative number B_2 as

$$\begin{aligned} \mathcal{E} &= \mathcal{E}_4(j_1) \cap \mathcal{E}_1(j_1 + 2 \log n). \\ B_2 &= \mu_{\max}(\alpha_1 + \beta_1(j_1 + 2 \log n)) + \alpha_4 + \beta_4 j_1 + \epsilon\delta \\ &= 2 \left(1 + \frac{\mu_{\max}}{\mu_{\min}} \right) \epsilon\delta + \Theta\left(\frac{n\ell_{\max}}{\delta} \log(n)\right) + \Theta\left(\frac{n\ell_{\max}}{\delta}\right) j_1. \end{aligned}$$

With the above definition, if $\mathbf{x}$ satisfies $V_2(\mathbf{x}) > \sum_{i=1}^I \ell_i \bar{x}_i + B_2$ and $\mathbf{x} \in \mathcal{E}$, we have

$$\begin{aligned} \sum_{i=1}^I \ell_i (x_i - \bar{x}_i) &> \mu_{\max} (\alpha_1 + \beta_1(j_1 + 2 \log n)) \\ &\quad + \alpha_4 + \beta_4 j_1 + \epsilon\delta, \end{aligned} \tag{9.105}$$

$$\sum_{i=1}^I \frac{\ell_i}{\mu_i} (x_i - \bar{x}_i) < \alpha_1 + \beta_1(j_1 + 2 \log n), \tag{9.106}$$

$$\sum_{i=1}^I \left(\frac{\mu_{\max}}{\mu_i} + \frac{\mu_i}{\mu_{\max}} - 2 \right) \ell_i (x_i - \bar{x}_i) > -\alpha_4 - \beta_4 j_1. \tag{9.107}$$

Taking the linear combination of the three inequalities $-(9.105) + \mu_{\max} \cdot (9.106) - (9.107)$ implies

$$\sum_{i=1}^I \left(1 - \frac{\mu_i}{\mu_{\max}} \right) \ell_i (x_i - \bar{x}_i) \leq -\epsilon\delta < 0; \tag{9.108}$$

Taking the linear combination $2 \cdot (9.105) - \mu_{\max} \cdot (9.106) + (9.107)$ implies

$$\sum_{i=1}^I \frac{\mu_i}{\mu_{\max}} \ell_i(x_i - \bar{x}_i) \geq \mu_{\max} (\alpha_1 + \beta_1(j_1 + 2 \log n)) + \alpha_4 + \beta_4 j_1 + 2\epsilon\delta > \epsilon\delta. \quad (9.109)$$

Note that when deriving (9.108) and (9.109), we have used the fact that all parameters involved in the expressions, $\alpha_1, \beta_1, j_1, \alpha_4, \beta_4, \delta$, are non-negative. Plugging (9.108) and (9.109) into (9.104) and choosing $\epsilon < \min(1 - \epsilon_0, \epsilon_0)$, we get

$$GV_2(\mathbf{x}, \mathbf{z}) \leq -\mu_{\max} \min(\epsilon\delta, \delta - \ell_{\max}) \leq -\mu_{\max} \epsilon\delta.$$

Next, we show that $\mathbf{X} \in \mathcal{E}$ with high probability. We apply Lemma 9.11 with $c_i = \frac{\mu_{\max}}{\mu_i} + \frac{\mu_i}{\mu_{\max}} - 2$, $c_{\max} = \frac{\mu_{\max}}{\mu_{\min}} + \frac{\mu_{\min}}{\mu_{\max}} - 2$, $\alpha = \epsilon\delta$, $\beta = \frac{c_{\max}^2 \mu_{\max} n \ell_{\max}}{\mu_{\min} \epsilon\delta}$. Because $\alpha\beta \geq c_{\max}^2 \mu_{\max} \sigma^2$, we have

$$\mathbb{P}\left(\sum_{i=1}^I \left(\frac{\mu_{\max}}{\mu_i} + \frac{\mu_i}{\mu_{\max}} - 2\right) \ell_i(X_i - \bar{x}_i) \leq -\alpha - \beta j\right) \leq e^{-j}. \quad (9.110)$$

Setting $j = j_1 + 2 \log n$ in the last inequality and recalling the definitions of α_4 and β_4 , we get

$$\mathbb{P}\left(\sum_{i=1}^I \left(\frac{\mu_{\max}}{\mu_i} + \frac{\mu_i}{\mu_{\max}} - 2\right) \ell_i(X_i - \bar{x}_i) \leq -\alpha_4 - \beta_4 j_1\right) \leq \frac{1}{n^2} e^{-j_1}, \quad (9.111)$$

which implies that $\mathbb{P}(\mathbf{X} \in \mathcal{E}_4(j_1)) \geq 1 - \frac{1}{n^2} e^{-j_1}$. Moreover, applying Lemma 9.15 to the set $\mathcal{E}_1(j_1 + 2 \log n)$, we have $\mathbb{P}(\mathbf{X} \in \mathcal{E}_1(j_1 + 2 \log n)) \geq 1 - \frac{1}{n^2} e^{-j_1}$. Because $\mathcal{E} = \mathcal{E}_4(j_1) \cap \mathcal{E}_1(j_1 + 2 \log n)$, by union bound, we have

$$\mathbb{P}(\mathbf{X} \in \mathcal{E}) \geq 1 - \frac{2}{n^2} e^{-j_1}.$$

When $\mathbf{x} \notin \mathcal{E}$, from the formula of $GV_2(\mathbf{x}, \mathbf{z})$ in (9.103), it straightforward that $GV_2(\mathbf{x}, \mathbf{z}) \leq \mu_{\max} n$.

We apply Lemma 9.26 with $B = \sum_{i=1}^I \ell_i \bar{x}_i + B_2$, $\gamma = -\mu_{\max} \epsilon\delta$, $\xi = \mu_{\max} n$, $\mathcal{E} = \mathcal{E}_4(j_1) \cap \mathcal{E}_1(j_1 + 2 \log n)$, $f_{\max} \leq \mu_{\max} n$, $v_{\max} = \ell_{\max}$. This gives us

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^I \ell_i(X_i - \bar{x}_i) \geq B_2 + 2\ell_{\max} \left(\left(\frac{f_{\max}}{\mu_{\max} \epsilon\delta} + 2\right) j_2 + 1\right)\right) \\ \leq e^{-j_2} + \left(\frac{n}{\epsilon\delta} + 1\right) \frac{2}{n^2} e^{-j_1}; \end{aligned} \quad (9.112)$$

after a further simplification, we get

$$\mathbb{P}\left(\sum_{i=1}^I \ell_i(X_i - \bar{x}_i) \geq B_2 + 2\ell_{\max}\left(\frac{(2\epsilon_0 + 1)n}{\epsilon\delta}j_2 + 1\right)\right) \leq e^{-j_2} + e^{-j_1}, \quad (9.113)$$

where we have used the fact that $f_{\max} \leq n\mu_{\max}$, $n > \delta$, $\epsilon < \epsilon_0$, and $(\frac{n}{\epsilon\delta} + 1)\frac{2}{n^2} \leq 1$ which holds for n large enough. Taking j_2 to be j_1 and rearranging the terms, we get

$$\mathbb{P}\left(\sum_{i=1}^I \ell_i(X_i - \bar{x}_i) \geq \alpha_2 + \beta_2 j_1\right) \leq 2 \exp(-j_1), \quad (9.114)$$

where

$$\begin{aligned} \beta_2 &= \frac{(2\epsilon_0 + 1)n\ell_{\max}}{\epsilon\delta} + \mu_{\max}\beta_1 + \beta_4 = \Theta\left(\frac{n\ell_{\max}}{\delta}\right), \\ \alpha_2 &= \mu_{\max}(\alpha_1 + 2\beta_2 \log n) + \alpha_4 + \epsilon\delta + 2\ell_{\max} \\ &= 2\left(1 + \frac{\mu_{\max}}{\mu_{\min}}\right)\epsilon\delta + \Theta\left(\frac{n\ell_{\max}}{\delta} \log(n)\right). \end{aligned}$$

Choosing any ϵ satisfying $0 < \epsilon < \min\left(1 - \epsilon_0, \epsilon_0, \frac{1}{4(1 + \mu_{\max}/\mu_{\min})}\right)$ and $j_1 = j + \log(2)$, we finish the proof. $\square$

Lemma 9.29 (Lemma 9.17 (Total server need upper bound under a lighter traffic)). *Consider the multiserver-job system with n servers satisfying $\ell_{\max} \leq \epsilon_0\delta$ (Assumption 9.2), and $\delta = \omega(\sqrt{n\ell_{\max}} \log n)$. Letting $\bar{x}_i = \frac{\lambda_i}{\mu_i}$, under any $\ell_{\max}$ -work-conserving policy, the expected total server need has the following upper bound:*

$$\mathbb{E}\left[\sum_{i=1}^I \ell_i(X_i(\infty) - \bar{x}_i)\right] = \exp\left(-\Omega\left(\frac{\delta^2}{n\ell_{\max}}\right)\right). \quad (9.61 \text{ restated})$$

Proof of Lemma 9.17. To prove the lemma, we will connect the expectation $\mathbb{E}\left[\sum_{i=1}^I \ell_i(X_i - \bar{x}_i)\right]$ with the tail probability $\mathbb{P}\left(\sum_{i=1}^I \ell_i(X_i - \bar{x}_i) \geq \alpha_2 + \beta_2 j\right)$, where $\alpha_2 = \frac{\delta}{2} + \Theta\left(\frac{n\ell_{\max}}{\delta} \log n\right)$, $\beta_2 = \Theta\left(\frac{n\ell_{\max}}{\delta}\right)$ are given in Lemma 9.16.

We first decompose $\mathbb{E}\left[\sum_{i=1}^I \ell_i(X_i - \bar{x}_i)\right]$ as follows:

$$\mathbb{E}\left[\sum_{i=1}^I \ell_i(X_i - \bar{x}_i)\right] = \mathbb{E}\left[\sum_{i=1}^I \ell_i(X_i - \bar{x}_i) \mathbf{1}_{\{\sum_{i=1}^I \ell_i X_i \geq n\}}\right]$$

$$\begin{aligned}
& + \mathbb{E} \left[\sum_{i=1}^I \ell_i(X_i - \bar{x}_i) \mathbf{1}_{\{\sum_{i=1}^I \ell_i X_i < n\}} \right] \\
& = \mathbb{E} \left[\left(\sum_{i=1}^I \ell_i(X_i - \bar{x}_i) - \delta \right)^+ \right] \tag{9.115}
\end{aligned}$$

$$+ \delta \mathbb{P} \left(\sum_{i=1}^I \ell_i X_i \geq n \right) \tag{9.116}$$

$$+ \mathbb{E} \left[\sum_{i=1}^I \ell_i(X_i - \bar{x}_i) \mathbf{1}_{\{\sum_{i=1}^I \ell_i X_i < n\}} \right], \tag{9.117}$$

where in the second equality we have used the fact that $\sum_{i=1}^I \ell_i \bar{x}_i + \delta = n$.

Next, we bound the three terms separately. The first term $\mathbb{E} \left[\left(\sum_{i=1}^I \ell_i(X_i - \bar{x}_i) - \delta \right)^+ \right]$ can be rewritten as follows:

$$\begin{aligned}
& \mathbb{E} \left[\left(\sum_{i=1}^I \ell_i(X_i - \bar{x}_i) - \delta \right)^+ \right] \\
& = \int_0^\infty \mathbb{P} \left(\sum_{i=1}^I \ell_i(X_i - \bar{x}_i) - \delta \geq t \right) dt \\
& \leq \sum_{j=\lfloor \frac{\delta - \alpha_2}{\beta_2} \rfloor}^\infty \int_{\alpha_2 + \beta_2 j}^{\alpha_2 + \beta_2(j+1)} \mathbb{P} \left(\sum_{i=1}^I \ell_i(X_i - \bar{x}_i) \geq t \right) dt \\
& \leq \sum_{j=\lfloor \frac{\delta - \alpha_2}{\beta_2} \rfloor}^\infty \beta_2 \mathbb{P} \left(\sum_{i=1}^I \ell_i(X_i - \bar{x}_i) \geq \alpha_2 + \beta_2 \cdot j \right) \\
& \leq \sum_{j=\lfloor \frac{\delta - \alpha_2}{\beta_2} \rfloor}^\infty \beta_2 e^{-j},
\end{aligned}$$

where the last inequality is due to Lemma 9.16. Therefore,

$$\mathbb{E} \left[\left(\sum_{i=1}^I \ell_i(X_i - \bar{x}_i) - \delta \right)^+ \right] \leq \frac{e}{e-1} \beta_2 \exp \left(- \left\lfloor \frac{\delta - \alpha_2}{\beta_2} \right\rfloor \right), \tag{9.118}$$

where e is the base for the natural logarithm.

Next, we bound $\delta \mathbb{P}(\sum_{i=1}^I \ell_i X_i \geq n) + \mathbb{E}[\sum_{i=1}^I \ell_i (X_i - \bar{x}_i) \mathbf{1}_{\{\sum_{i=1}^I \ell_i X_i < n\}}]$ by analyzing the Lyapunov drift of the function $f(\mathbf{x}) = \sum_{i=1}^I \frac{\ell_i}{\mu_i} x_i$, given by

$$Gf(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^I \left(\frac{\lambda_i \ell_i}{\mu_i} - \ell_i z_i \right).$$

Because $f(\mathbf{x})$ is a linear function, and the system operates under an $\ell_{\max}$ -work-conserving policy, by Lemma 9.22, we have $\mathbb{E}[Gf(\mathbf{X}, \mathbf{Z})] = 0$, or equivalently,

$$\mathbb{E} \left[\sum_{i=1}^I \left(\frac{\lambda_i \ell_i}{\mu_i} - \ell_i Z_i \right) \right] = 0. \quad (9.119)$$

By the $\ell_{\max}$ -work-conserving condition, $\sum_{i=1}^I \ell_i Z_i \geq n - \ell_{\max}$ when $\sum_{i=1}^I \ell_i X_i \geq n$, and $\sum_{i=1}^I \ell_i Z_i = \sum_{i=1}^I \ell_i X_i$ when $\sum_{i=1}^I \ell_i X_i < n$. Therefore,

$$\begin{aligned} \mathbb{E} \left[\sum_{i=1}^I \left(\frac{\lambda_i \ell_i}{\mu_i} - \ell_i Z_i \right) \mathbf{1}_{\{\sum_{i=1}^I \ell_i X_i \geq n\}} \right] &\leq (-\delta + \ell_{\max}) \mathbb{P} \left(\sum_{i=1}^I \ell_i X_i \geq n \right), \\ \mathbb{E} \left[\sum_{i=1}^I \left(\frac{\lambda_i \ell_i}{\mu_i} - \ell_i Z_i \right) \mathbf{1}_{\{\sum_{i=1}^I \ell_i X_i < n\}} \right] &\leq \mathbb{E} \left[\sum_{i=1}^I \ell_i (\bar{x}_i - X_i) \mathbf{1}_{\{\sum_{i=1}^I \ell_i X_i < n\}} \right]. \end{aligned}$$

Summing up the above two inequalities and rearranging the terms, we get

$$\begin{aligned} &\mathbb{E} \left[\sum_{i=1}^I \ell_i (X_i - \bar{x}_i) \mathbf{1}_{\{\sum_{i=1}^I \ell_i X_i < n\}} \right] + \delta \mathbb{P} \left(\sum_{i=1}^I \ell_i X_i \geq n \right) \\ &\leq \ell_{\max} \mathbb{P} \left(\sum_{i=1}^I \ell_i X_i \geq n \right) = \ell_{\max} \mathbb{P} \left(\sum_{i=1}^I \ell_i (X_i - \bar{x}_i) \geq \delta \right) \\ &\leq \ell_{\max} \exp \left(- \left\lfloor \frac{\delta - \alpha_2}{\beta_2} \right\rfloor \right). \end{aligned} \quad (9.120)$$

where the last inequality follows from Lemma 9.16 with $j = \lfloor \frac{\delta - \alpha_2}{\beta_2} \rfloor$.

Combining (9.118) and (9.120), we get a bound on $\mathbb{E}[\sum_{i=1}^I \ell_i (X_i - \bar{x}_i)]$:

$$\mathbb{E} \left[\sum_{i=1}^I \ell_i (X_i - \bar{x}_i) \right] \leq \left(\frac{e}{e-1} \beta_2 + \ell_{\max} \right) \exp \left(- \left\lfloor \frac{\delta - \alpha_2}{\beta_2} \right\rfloor \right). \quad (9.121)$$

Observe that

$$\frac{e}{e-1}\beta_2 + \ell_{\max} = O\left(\frac{n\ell_{\max}}{\delta}\right). \quad (9.122)$$

Because $\delta = \omega\left(\frac{n\ell_{\max}}{\delta} \log n\right)$, we have $\alpha_2 = \frac{\delta}{2} + o(\delta)$. Consequently,

$$\left\lfloor \frac{\delta - \alpha_2}{\beta_2} \right\rfloor = \Omega\left(\frac{\delta^2}{n\ell_{\max}}\right). \quad (9.123)$$

Therefore, (9.121) can be simplified into

$$\mathbb{E}\left[\sum_{i=1}^I \ell_i(X_i - \bar{x}_i)\right] = O\left(\frac{n\ell_{\max}}{\delta}\right) \exp\left(-\Omega\left(\frac{\delta^2}{n\ell_{\max}}\right)\right).$$

Because $\delta = \omega\left(\sqrt{n\ell_{\max}} \log n\right)$, we have $\exp\left(-\Omega\left(\frac{\delta^2}{n\ell_{\max}}\right)\right) \leq \exp\left(-\Omega((\log n)^2)\right)$, which decays faster than any polynomial. Therefore we can omit the factor $O\left(\frac{n\ell_{\max}}{\delta}\right)$. This finishes the proof of Lemma 9.17. $\square$

Corollary 9.30 (Corollary 9.18 (**Queueing probability**)). *Consider the multiserver-job system with n servers satisfying $\ell_{\max} \leq \epsilon_0\delta$ (Assumption 9.2), and $\delta = \omega\left(\sqrt{n\ell_{\max}} \log n\right)$. Then the steady-state probability that a job experiences queueing has the following upper bound:*

$$\mathbb{P}\left(\sum_{i=1}^I \ell_i X_i(\infty) \geq n\right) = \exp\left(-\Omega\left(\frac{\delta^2}{n\ell_{\max}}\right)\right). \quad (9.62 \text{ restated})$$

Proof of Corollary 9.18. Choosing $j = \frac{\delta - \alpha_2}{\beta_2}$ in Lemma 9.16, we have

$$\mathbb{P}\left(\sum_{i=1}^I \ell_i(X_i - \bar{x}_i) \geq \alpha_2 + \beta_2 \cdot j\right) \leq \exp\left(-\frac{\delta - \alpha_2}{\beta_2}\right). \quad (9.124)$$

Recall that $\alpha_2 = \frac{\delta}{2} + \Theta\left(\frac{n\ell_{\max}}{\delta} \log n\right)$ and $\beta_2 = \Theta\left(\frac{n\ell_{\max}}{\delta}\right)$. Because we have assumed $\delta = \omega\left(\sqrt{n\ell_{\max}} \log n\right)$, $\alpha_2 = \frac{\delta}{2} + o(\delta)$. Therefore, $\delta \geq \alpha_2 + \beta_2 \cdot j$ and $\frac{\delta - \alpha_2}{\beta_2} = \Omega\left(\frac{\delta^2}{n\ell_{\max}}\right)$. It follows that

$$\mathbb{P}\left(\sum_{i=1}^I \ell_i X_i \geq n\right) = \mathbb{P}\left(\sum_{i=1}^I \ell_i(X_i - \bar{x}_i) \geq \delta\right)$$

$$\begin{aligned}
&\leq \mathbb{P}\left(\sum_{i=1}^I \ell_i(X_i - \bar{x}_i) \geq \alpha_2 + \beta_2 \cdot j\right) \\
&\leq \exp\left(-\Omega\left(\frac{\delta^2}{n\ell_{\max}}\right)\right).
\end{aligned}$$

This finishes the proof of Corollary 9.18. $\square$

9.F More Discussions on the Experiment 2 in the Simulations Section

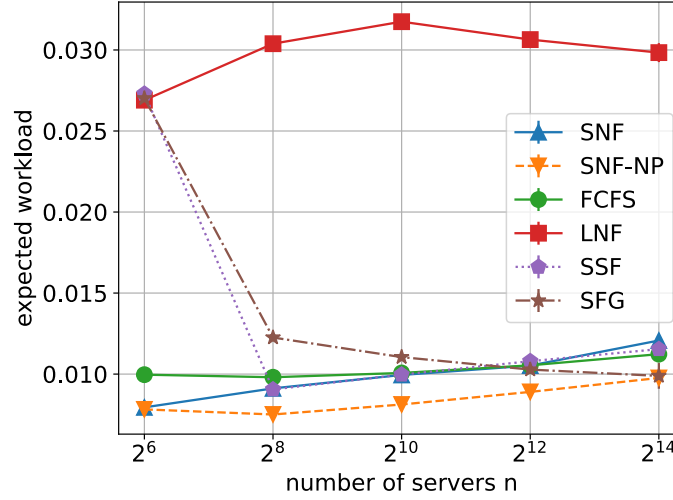
In this section, we give further discussions of Experiment 2 presented in Section 9.10. In the setting of Experiment 2, the jobs with large server needs also have large service rates. A notable observation from the experiment is that under the Smallest-Need-First (SNF) policy or its non-preemptive variant (SNF-NP), the mean waiting time of the jobs is in general much smaller than under the Largest-Need-First (LNF) policy; this is even the case when LNF agrees with the Smallest-Size-First (SSF) policy.

To better understand why SNF and SNF-NP lead to small mean waiting times in Experiment 2, we look into the steady-state expected workload, $\mathbb{E}\left[\sum_{i=1}^I \frac{\ell_i}{\mu_i} Q_i\right]$. In Figure 9.6, we plot the expected workloads for the two experiments in Figure 9.4, where Figure 9.6(a) and (b) corresponds to Figure 9.4(a) and (b), respectively. Notably, the expected workloads have a large variation across different policies: The expected workloads under LNF are much larger than those under SNF and SNF-NP; the expected workloads under SSF are close to those under SNF if SSF gives the maximal-need jobs the least priority, and close to those under LNF otherwise; the expected workloads under ServerFilling-Gittins (SFG) are close to those under SSF.

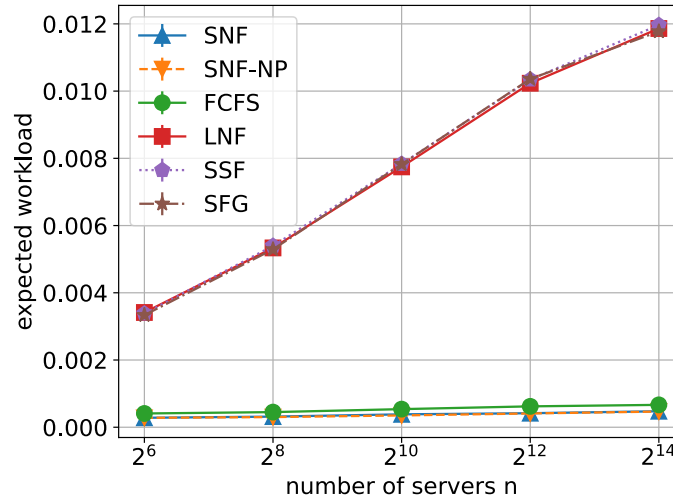
As a brief aside, we compare the simulation results in Figure 9.6 with our theory. By Lemma 9.7 and Lemma 9.9, under Assumptions 9.1-9.3, the expected workloads under any $\ell_{\max}$ -work-conserving policy are bounded as follows:

$$\frac{\sigma^2}{\delta} \cdot (1 - o(1)) \leq \mathbb{E}\left[\sum_{i=1}^I \frac{\ell_i}{\mu_i} Q_i\right] \leq \frac{\sigma^2}{\delta - \ell_{\max}} \cdot (1 + o(1)).$$

These bounds suggest that for sufficiently large n , the expected workloads under any two $\ell_{\max}$ -work-conserving policies differ by at most a factor of



(a) Expected workloads under Parameter Set Three. We can observe that there is a large difference in the expected workloads under LNF and the expected workloads under other policies for $n \geq 2^8$.



(b) Expected workloads under Parameter Set Four. Observe that under Parameter Set Four, the variation in the expected workloads across policies is even larger than that under Parameter Set Three in Figure 9.6a.

Figure 9.6: The expected workloads in Experiment 2 from the Simulations Section.

$\ell_{\max}/(\delta - \ell_{\max})$ — a factor of 2 in the setting of Figure 9.6 where $\ell_{\max} = \lfloor \sqrt{n} \rfloor$ and $\delta = 2\lfloor \sqrt{n} \rfloor$. The significant variations in the expected workloads under different policies in Figure 9.6 are due to the large μ_i values, necessitating a much larger n for the $o(1)$ terms in the bounds to become negligible.

Why do SNF and SNF-NP lead to smaller expected workloads than LNF in Figure 9.6? At first glance, it seems counterintuitive, because LNF packs the jobs more compactly and leaves fewer servers idle when there are lots of jobs in the system. However, when there are not many jobs in the system, we conjecture that SNF and SNF-NP might be able to keep more servers busy than LNF and thus result in smaller expected workloads in the steady state. Here is our *heuristic argument*: consider the normalized work defined in Section 9.6, $\sum_{i=1}^I \frac{\ell_i}{\mu_i}(X_i - \bar{x}_i)$, where $\bar{x}_i = \frac{\lambda_i}{\mu_i}$, whose steady-state expectation equals the expected workload. We can show that the drift of normalized work equals $-\sum_{i=1}^I \ell_i(Z_i - \bar{x}_i)$, the number of busy servers subtracting a constant. Under a $\ell_{\max}$ -work-conserving-policy, when $\ell_i X_i > n$, the drift of normalized work is approximately fixed, equal to $-(n - \sum_{i=1}^I \ell_i \bar{x}_i)$; when $\ell_i X_i \leq n$, all jobs enter service and the drift equals $-\sum_{i=1}^I \ell_i(X_i - \bar{x}_i)$. In the latter case, given the same amount of normalized work, because $X_i \geq 0$, the drift is smaller if more jobs in the system have larger service rates, or equivalently have larger server needs in the setting of Experiment 2. Therefore, intuitively, because SNF and SNF-NP give the maximal-need jobs the lowest priority, there are more jobs with large service rates under SNF and SNF-NP than under other policies, which results in a more negative drift of normalized work when $\ell_i X_i \leq n$. Consequently, the normalized work under SNF and SNF-NP can reach lower values during the periods when $\ell_i X_i \leq n$ and thus have a smaller steady-state expectation. Validating this heuristic argument is left for future work.

To relate the expected workload back to mean waiting times, recall that by Little's law, the mean waiting time is proportional to the total queue length, $\mathbb{E}[\sum_{i=1}^I Q_i]$. Heuristically, for a priority policy that gives type i_0 the lowest priority, all jobs except the type i_0 jobs are in light traffic and their queue lengths are nearly zero. As a result,

$$\mathbb{E} \left[\sum_{i=1}^I Q_i \right] \approx \frac{\mu_{i_0}}{\ell_{i_0}} \mathbb{E} \left[\sum_{i=1}^I \frac{\ell_i}{\mu_i} Q_i \right].$$

In the case when the expected workloads $\mathbb{E} \left[\sum_{i=1}^I \frac{\ell_i}{\mu_i} Q_i \right]$ do not vary much across policies, this formula suggests that one should give the least priority

to the jobs with the *largest expected size*, $\frac{\ell_i}{\mu_i}$. However, when the expected workloads heavily depend on the policy, the benefit of having a small expected workload could be dominant and thus make a different priority order prevail — like giving the least priority to the jobs with the *largest server needs* in the setting of Experiment 2.

Bibliography

- [1] Samuli Aalto, Urtzi Ayesta, and Rhonda Righter. On the Gittins index in the $M/G/1$ queue. *Queueing Syst.*, 63(1-4):437–458, 2009.
- [2] Samuli Aalto, Urtzi Ayesta, and Rhonda Righter. Properties of the Gittins index with application to optimal scheduling. *Probab. Eng. Inf. Sci.*, 25(3):269–288, 2011.
- [3] Samuli Aalto, Pasi Lassila, and Ianire Taboada. Whittle index approach to opportunistic scheduling with partial channel information. *Perform. Eval.*, 136:102052, 2019.
- [4] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek Gordon Murray, Benoit Steiner, Paul A. Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: A system for large-scale machine learning. In *Proc. USENIX Conf. Operating Systems Design and Implementation (OSDI)*, pages 265–283, Berkeley, CA, 2016.
- [5] Larisa Afanaseva, Elena Bashtova, and Svetlana Grishunina. Stability analysis of a multi-server model with simultaneous service and a regenerative input flow. *Methodol. Comp. Appl. Probab.*, pages 1–17, 2019.
- [6] Reza Aghajani and Kavita Ramanan. The hydrodynamic limit of a randomized load balancing network. *Adv. Appl. Probab.*, 29(4):pp. 2114–2174, 2019.
- [7] Reza Aghajani and Kavita Ramanan. The limit of stationary distributions of many-server queues in the Halfin–Whitt regime. *Math. Oper. Res.*, 45(3):1016–1055, 2020.

- [8] Sebastian Allmeier and Nicolas Gast. Mean field and refined mean field approximations for heterogeneous systems: It works! *Proc. ACM Meas. Anal. Comput. Syst.*, 6(1), 2022.
- [9] Sebastian Allmeier and Nicolas Gast. Accuracy of the graphon mean field approximation for interacting particle systems. *Stoch. Syst.*, 15(4): 273–290, 2025.
- [10] Eitan Altman. *Constrained Markov Decision Processes*. Stochastic Modeling Series. Chapman & Hall/CRC, Boca Raton, FL, 1999.
- [11] P. S. Ansell, K. D. Glazebrook, J. Niño-Mora, and M. O’Keeffe. Whittle’s index policy for a multi-class queueing system with convex holding costs. *Math. Oper. Res.*, 57(1):21–39, 2003.
- [12] Ruicheng Ao, Gan Luo, David Simchi-Levi, and Xinshang Wang. Optimizing LLM inference: Fluid-guided online scheduling with memory constraints. *arXiv:2504.11320 [cs.LG]*, 2026.
- [13] T. W. Archibald, D. P. Black, and K. D. Glazebrook. Indexability and index heuristics for a simple class of inventory routing problems. *Oper. Res.*, 57(2):314–326, 2009.
- [14] Jesus R. Artalejo, Antonis Economou, and Maria J. Lopez-Herrero. Analysis of a multiserver queue with setup times. *Queueing Syst.*, 51(1-2):53–76, 2005.
- [15] Ed Arthurs and Joseph S. Kaufman. Sizing a message store subject to blocking criteria. In M. Arató, A. Butrimenko, and Erol Gelenbe, editors, *Proc. Int. Symp. Computer Performance, Modeling, Measurements and Evaluation (IFIP Performance)*, pages 547–564, Amsterdam, Netherland, 1979.
- [16] Søren Asmussen. *Applied Probability and Queues*. Springer-Verlag New York, 2003.
- [17] Søren Asmussen and Peter W. Glynn. *Steady-State Simulation*, pages 96–125. Springer New York, New York, NY, 2007.
- [18] Baris Ata and Itai Gurvich. On optimality gaps in the Halfin-Whitt regime. *Ann. Appl. Probab.*, 22(1):407–455, 2012.
- [19] Rami Atar. A diffusion regime with nondegenerate slowdown. *Oper. Res.*, 60(2):490–500, 2012.

- [20] Rami Atar and Nir Solomon. Asymptotically optimal interruptible service policies for scheduling jobs in a diffusion regime with nondegenerate slowdown. *Queueing Syst.*, 69(3):217–235, 2011.
- [21] Rami Atar, Avi Mandelbaum, and Martin I. Reiman. Scheduling a multi class queue with many exponential servers: Asymptotic optimality in heavy traffic. *Ann. Appl. Probab.*, 14(3):1084–1134, 2004.
- [22] Konstantin Avrachenkov and Vivek S. Borkar. A learning algorithm for the Whittle index policy for scheduling web crawlers. In *Proc. Ann. Allerton Conf. Communication, Control and Computing*, pages 1001–1006, 2019.
- [23] Konstantin Avrachenkov, Urtzi Ayesta, Josu Doncel, and Peter Jacko. Congestion control of TCP flows in Internet routers by means of index policy. *Computer Networks*, 57(17):3463–3478, 2013.
- [24] Konstantin E. Avrachenkov and Vivek S. Borkar. Whittle index based Q-learning for restless bandits with average reward. *Automatica*, 139: 110186, 2022.
- [25] Nikhil Ayyadevara, Rajni Dabas, Arindam Khan, and K. V. N. Sreenivas. Near-Optimal Algorithms for Stochastic Online Bin Packing. In *Proc. Int. Conf. Automata, Languages and Programming (ICALP)*, volume 229, pages 12:1–12:20, 2022.
- [26] François Baccelli and Pierre Brémaud. *Elements of Queueing Theory*. Springer-Verlag Berlin Heidelberg, 2nd edition, 2003.
- [27] Sayan Banerjee, Amarjit Budhiraja, and Amber L. Puha. Heavy traffic scaling limits for shortest remaining processing time queues with heavy tailed processing time distributions. *Ann. Appl. Probab.*, 32(4):2587–2651, 2022.
- [28] Noman Bashir, Nan Deng, Krzysztof Rządca, David Irwin, Sree Kodak, and Rohit Jnagal. Take it to the limit: Peak prediction-driven resource overcommitment in datacenters. In *Proc. European Conf. Computer Systems (EuroSys)*, page 556–573, Online Event, United Kingdom, 2021.
- [29] N. G. Bean, R. J. Gibbens, and S. Zachary. Asymptotic analysis of single resource loss systems in heavy traffic, with applications to integrated networks. *Adv. Appl. Probab.*, 27(1):273–292, 1995.
- [30] Nabil Benameur, Slim Ben Fredj, Frank Delcoigne, Sara Oueslati-Boulahia, and James W. Roberts. Integrated admission control for

- streaming and elastic traffic. In *Int. Workshop Quality of Future Internet Services (QofIS)*, volume 2156, pages 69–81, Berlin, Heidelberg, 2001.
- [31] Dimitris Bertsimas and José Niño Mora. Conservation laws, extended polymatroids and multiarmed bandit problems; a polyhedral approach to indexable systems. *Math. Oper. Res.*, 21(2):257–306, 1996.
- [32] Dimitris Bertsimas and José Niño Mora. Restless bandits, linear programming relaxations, and a primal-dual index heuristic. *Oper. Res.*, 48(1):80–90, 2000.
- [33] Dimitris Bertsimas, David Gamarnik, and John N. Tsitsiklis. Performance of multiclass Markovian queueing networks via piecewise linear Lyapunov functions. *Ann. Appl. Probab.*, 11(4):1384–1428, 11 2001.
- [34] Wolfgang Bischof. Analysis of $M/G/1$ -queues with setup times and vacations under six different service disciplines. *Queueing Syst.*, 39(4):265–301, 2001.
- [35] Arpita Biswas, Gaurav Aggarwal, Pradeep Varakantham, and Milind Tambe. Learning index policies for restless bandits with application to maternal healthcare. In *Proc. Int. Conf. Autonomous Agents and Multiagent Syst.*, page 1467–1468, 2021.
- [36] Arpita Biswas, Gaurav Aggarwal, Pradeep Varakantham, and Milind Tambe. Learning index policies for restless bandits with application to maternal healthcare. In *Proc. Int. Jt. Conf. Artificial Intelligence (IJCAI)*, pages 1467–1468, 2021.
- [37] Gunter Bolch, Stefan Greiner, Hermann de Meer, and Kishor S. Trivedi. *Queueing networks and Markov chains: modeling and performance evaluation with computer science applications*. John Wiley & Sons, 2nd edition, 2006.
- [38] A. A. Borovkov. Some limit theorems in the theory of mass service, II multiple channels systems. *Theory Probab. Appl.*, 10(3):375–400, 1965.
- [39] Sem Borst, Avi Mandelbaum, and Martin I. Reiman. Dimensioning large call centers. *Oper. Res.*, 52(1):17–34, 2004.
- [40] Craig Boutilier and Tyler Lu. Budget allocation using weakly coupled, constrained Markov decision processes. In *Conf. Uncertainty in Artificial Intelligence (UAI)*, pages 52–61, 2016.

- [41] Onno J. Boxma and Wim P. Groenendijk. Pseudo-conservation laws in cyclic-service systems. *J. Appl. Probab.*, 24(4):949–964, 1987.
- [42] Maury Bramson, Yi Lu, and Balaji Prabhakar. Asymptotic independence of queues under randomized load balancing. *Queueing Syst.*, 71(3):247–292, 2012.
- [43] Alfred Brauer. Limits for the characteristic roots of a matrix. IV: Applications to stochastic matrices. *Duke Math. J.*, 19:75–91, 1952.
- [44] Anton Braverman. Stein’s method for steady-state diffusion approximations. *arXiv:1704.08398 [math.PR]*, 2017.
- [45] Anton Braverman. Steady-state analysis of the join-the-shortest-queue model in the Halfin–Whitt regime. *Math. Oper. Res.*, 45(3):1069–1103, 2020.
- [46] Anton Braverman. The prelimit generator comparison approach of Stein’s method. *Stoch. Syst.*, 12(2):181–204, 2022.
- [47] Anton Braverman. The join-the-shortest-queue system in the Halfin–Whitt regime: Rates of convergence to the diffusion limit. *Stoch. Syst.*, 13(1):1–39, 2023.
- [48] Anton Braverman and J. G. Dai. High order steady-state diffusion approximation of the Erlang-C system. *arXiv:1602.02866 [math.PR]*, 2016.
- [49] Anton Braverman and JG Dai. Stein’s method for steady-state diffusion approximations of $M/Ph/n + M$ systems. *Ann. Appl. Probab.*, 27:550–581, 2017.
- [50] Anton Braverman and Ziv Scully. Diffusion approximation error for queueing systems with general primitives. *arXiv:2407.12716 [math.PR]*, 2025.
- [51] Anton Braverman, J. G. Dai, and Jiekun Feng. Stein’s method for steady-state diffusion approximations: an introduction through the Erlang-A and Erlang-C models. *Stoch. Syst.*, 6(2):301–366, 2017.
- [52] Anton Braverman, J. G. Dai, and Masakiyo Miyazawa. Heavy traffic approximation for the stationary distribution of a generalized Jackson network: The BAR approach. *Stoch. Syst.*, 7(1):143–196, 2017.
- [53] Anton Braverman, J. G. Dai, and Xiao Fang. High-order steady-state diffusion approximations. *Oper. Res.*, 72(2):604–616, 2024.

- [54] Anton Braverman, J. G. Dai, and Masakiyo Miyazawa. The BAR approach for multiclass queueing networks with SBP service policies. *Stoch. Syst.*, 15(1):1–49, 2025.
- [55] Percy H. Brill and Linda Green. Queues in which customers receive simultaneous service from a random number of servers: A system point approach. *Manage. Sci.*, 30(1):51–68, 1984.
- [56] David B. Brown and James E. Smith. Index policies and performance bounds for dynamic selection problems. *Manage. Sci.*, 66(7):3029–3050, 2020.
- [57] David B. Brown and Jingwei Zhang. Dynamic programs with shared resources and signals: Dynamic fluid policies and asymptotic optimality. *Oper. Res.*, 70(5):3015–3033, 2022.
- [58] David B. Brown and Jingwei Zhang. Fluid policies, reoptimization, and performance guarantees in dynamic resource allocation. *Oper. Res.*, 73(2):1029–1045, 2025.
- [59] Niv Buchbinder, Yaron Fairstein, Konstantina Mellou, Ishai Menache, and Joseph (Seffi) Naor. Online virtual machine allocation with lifetime and load predictions. *ACM SIGMETRICS Perform. Evaluation Rev.*, 49(1):9–10, 2021.
- [60] Cheng-Shang Chang, Joy A. Thomas, and Shaw-Hwa Kiang. On the stability of open networks: A unified approach by stochastic dominance. *Queueing Syst.*, 15(1):239–260, 1994.
- [61] Mohan L. Chaudhry and James G. C. Templeton. *A First Course in Bulk Queues*. Wiley, New York, NY, 1983.
- [62] Zixi Chen, Yinyu Ye, and Zijie Zhou. Adaptively robust LLM inference optimization under prediction uncertainty. *arXiv:2508.14544 [cs.LG]*, 2025.
- [63] Gautam Choudhury. On a batch arrival Poisson queue with a random setup time and vacation period. *Computers & Operations Research*, 25(12):1013–1026, 1998.
- [64] E. G. Coffman, Jr., M. R. Garey, and D. S. Johnson. Dynamic bin packing. *SIAM J. Comput.*, 12(2):227–258, 1983.
- [65] Coastas Courcoubetis and Richard Weber. Stability of on-line bin packing with random arrivals and long-run-average constraints. *Probab. Eng.*

- Inf. Sci.*, 4(4):447–460, 1990.
- [66] C. Courcoubetis and R. R. Weber. Necessary and sufficient conditions for stability of a bin-packing system. *J. Appl. Probab.*, 23(4):989–999, 1986.
 - [67] Mark E. Crovella and Azer Bestavros. Self-similarity in World Wide Web traffic: Evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, Dec./1997.
 - [68] Janos Csirik, David S. Johnson, Claire Kenyon, James B. Orlin, Peter W. Shor, and Richard R. Weber. On the sum-of-squares algorithm for bin packing. *J. ACM*, 53(1):1–65, 2006.
 - [69] J. G. Dai. On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Ann. Appl. Probab.*, 5(1):49 – 77, 1995.
 - [70] J. G. Dai and Dongyan Huo. Asymptotic product-form steady-state for multiclass queueing networks with SBP service policies in multi-scale heavy traffic. *arXiv:2403.04090 [math.PR]*, 2024.
 - [71] J. G. Dai, Shuangchi He, and Tolga Tezcan. Many-server diffusion limits for G/Ph/n+GI queues. *Ann. Appl. Probab.*, 20(5):1854 – 1890, 2010.
 - [72] J. G. Dai, Jin Guang, and Yaosheng Xu. Steady-state convergence of the continuous-time JSQ system with general distributions in heavy traffic. *ACM SIGMETRICS Perform. Evaluation Rev.*, 52(2):39–41, 2024.
 - [73] J. G. Dai, Peter W. Glynn, and Yaosheng Xu. Asymptotic product-form steady state for generalized Jackson networks in multiscale heavy traffic. *Oper. Res.*, 2026.
 - [74] J.G. Dai and S.P. Meyn. Stability and convergence of moments for multiclass queueing networks via fluid limit models. *IEEE Trans. Autom. Control*, 40(11):1889–1904, 1995.
 - [75] Jim G Dai, AB Dieker, and Xuefeng Gao. Validity of heavy-traffic steady-state approximations in many-server queues with abandonment. *Queueing Syst.*, 78:1–29, 2014.
 - [76] Abel Dasylva and R. Srikant. Bounds on the performance of admission control and routing policies for general topology networks with multiple

- call classes. In *Proc. IEEE Int. Conf. Computer Communications (INFOCOM)*, volume 2, pages 505–512, Los Alamitos, CA, 1999.
- [77] Andrew Daw and Jamol Pender. On the distributions of infinite server queues with batch arrivals. *Queueing Syst.*, 91(3–4):367–401, apr 2019.
 - [78] Andrew Daw, Robert C Hampshire, and Jamol Pender. How to staff when customers arrive in batches. *arXiv:1907.12650 [cs.PF]*, 2019.
 - [79] Andrew Daw, Brian Fralix, and Jamol Pender. Non-stationary queues with batch arrivals. *arXiv:2008.00625 [math.PR]*, 2020.
 - [80] Christina Delimitrou and Christos Kozyrakis. Quasar: Resource-efficient and QoS-aware cluster management. In *Proc. Int. Conf. Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, page 127–144, Salt Lake City, UT, 2014.
 - [81] Lijun Ding and Yudong Chen. Leave-one-out approach for matrix completion: Primal and dual analysis. *IEEE Trans. Inf. Theory*, 66(11):7274–7301, 2020.
 - [82] Bharat T. Doshi. A note on stochastic decomposition in a $GI/G/1$ queue with vacations or set-up times. *J. Appl. Probab.*, 22(2):419–428, 1985.
 - [83] Bharat T. Doshi. Queueing systems with vacations — a survey. *Queueing Syst.*, 1(1):29–66, 1986.
 - [84] Bharat T. Doshi. Generalizations of the stochastic decomposition results for single server queues with vacations. *Communications in Statistics. Stochastic Models*, 6(2):307–333, 1990.
 - [85] Peter J. Downey. Bounding synchronization overhead for parallel iteration. *ORSA J. Comput.*, 3(4):288–298, 1991.
 - [86] Rick Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5 edition, 2019.
 - [87] Josh C. D’Aeth, Shubhechyya Ghosal, Fiona Grimm, David Haw, Esma Koca, Krystal Lau, Huikang Liu, Stefano Moret, Dheeya Rizmie, Peter C. Smith, Giovanni Forchini, Marisa Miraldo, and Wolfram Wiesemann. Optimal hospital care scheduling during the SARS-CoV-2 pandemic. *Manage. Sci.*, 69(10):5923–5947, 2023.

- [88] Ibrahim El Shar and Daniel Jiang. Weakly coupled deep Q-networks. *Conf. Neural Information Processing Systems (NeurIPS)*, 36, 2023.
- [89] Atilla Eryilmaz and R. Srikant. Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Syst.*, 72(3-4): 311–359, 2012.
- [90] Patrick Eschenfeldt and David Gamarnik. Join the shortest queue with many servers. the heavy-traffic asymptotics. *Math. Oper. Res.*, 43(3): 867–886, 2018.
- [91] D. Filippopoulos and H. Karatza. A two-class parallel system with general service times of the parallel class. *Journal of Computer and System Sciences*, 74(6):942–964, 2008.
- [92] Apache Software Foundation. Apache Mesos: Containerizers.
- [93] Apache Software Foundation. Apache Mesos: Oversubscription.
- [94] Daniel Freund and Siddhartha Banerjee. Good prophets know when the end is near. Available at SSRN: <https://ssrn.com/abstract=3479189>, 2019.
- [95] S. W. Fuhrmann. A note on the $M/G/1$ queue with server vacations. *Oper. Res.*, 32(6):1368–1373, 1984.
- [96] S. W. Fuhrmann and Robert B. Cooper. Stochastic decompositions in the $M/G/1$ queue with generalized vacations. *Oper. Res.*, 33(5): 1117–1129, 1985.
- [97] David Gamarnik and David A. Goldberg. Steady-state GI/GI/n queue in the Halfin–Whitt regime. *Ann. Appl. Probab.*, 23(6):2382–2419, 2013.
- [98] David Gamarnik and Petar Momčilović. Steady-state analysis of a multiserver queue in the Halfin–Whitt regime. *Adv. Appl. Probab.*, 40(2):548–577, 2008.
- [99] David Gamarnik and Alexander L. Stolyar. Multiclass multiserver queueing system in the Halfin–Whitt heavy traffic regime: asymptotics of the stationary distribution. *Queueing Syst.*, 71(1):25–51, 2012.
- [100] Anshul Gandhi and Mor Harchol-Balter. $M/G/k$ with exponential setup. Technical Report CMU-CS-09-166, Carnegie Mellon University, Pittsburgh, PA, 2009.
- [101] Anshul Gandhi and Mor Harchol-Balter. $M/G/k$ with staggered setup. *Oper. Res. Lett.*, 41(4):317–320, 2013.

- [102] Anshul Gandhi, Mor Harchol-Balter, and Ivo Adan. Server farms with setup costs. *Perform. Eval.*, 67(11):1123–1138, 2010.
- [103] Anshul Gandhi, Sherwin Doroudi, Mor Harchol-Balter, and Alan Scheller-Wolf. Exact analysis of the M/M/k/setup class of Markov chains via recursive renewal reward. *Queueing Syst.*, 77(2):177–209, 2014.
- [104] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manuf. Serv. Oper. Manag.*, 4(3):208–227, 2002.
- [105] Nicolas Gast, Bruno Gaujal, and Chen Yan. Exponential convergence rate for the asymptotic optimality of Whittle index policy. *arXiv:2012.09064 [cs.PF]*, 2020.
- [106] Nicolas Gast, Bruno Gaujal, and Kimang Khun. Testing indexability and computing Whittle and Gittins index in subcubic time. *Math. Methods Oper. Res.*, 97(3):391–436, 2023.
- [107] Nicolas Gast, Bruno Gaujal, and Chen Yan. Exponential asymptotic optimality of Whittle index policy. *Queueing Syst.*, 104(1):107–150, 2023.
- [108] Nicolas Gast, Bruno Gaujal, and Chen Yan. Linear program-based policies for restless bandits: Necessary and sufficient conditions for (exponentially fast) asymptotic optimality. *Math. Oper. Res.*, 49(4):2468–2491, 2024.
- [109] Nicolas Gast, Bruno Gaujal, and Chen Yan. Reoptimization nearly solves weakly coupled Markov decision processes. *arXiv:2211.01961 [math.OC]*, 2024.
- [110] Javad Ghaderi, Yuan Zhong, and R. Srikant. Asymptotic optimality of BestFit for stochastic bin packing. *ACM SIGMETRICS Perform. Evaluation Rev.*, 42(2):64–66, 2014.
- [111] Abheek Ghosh, Dheeraj Nagaraj, Manish Jain, and Milind Tambe. Indexability is not enough for Whittle: Improved, near-optimal algorithms for restless bandits. In *Proc. Int. Conf. Autonomous Agents and Multiagent Syst.*, page 1294–1302, Richland, SC, 2023.
- [112] John C. Gittins. Bandit processes and dynamic allocation indices. *J. Roy. Stat. Soc. B Met.*, 41(2):148–177, 1979.

- [113] John C. Gittins and D. M. Jones. A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics*, pages 241–266. North-Holland, Amsterdam, 1974.
- [114] John C. Gittins, K. D. Glazebrook, and Richard Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.
- [115] K. D. Glazebrook, H. M. Mitchell, and P. S. Ansell. Index policies for the maintenance of a collection of machines by a set of repairmen. *Eur. J. Oper. Res.*, 165(1):267–284, 2005.
- [116] Kevin D. Glazebrook. An analysis of Klimov’s problem with parallel servers. *Mathematical Methods of Operations Research*, 58(1):1–28, 2003.
- [117] Kevin D. Glazebrook and José Niño-Mora. Parallel scheduling of multiclass $M/M/m$ queues: Approximate and heavy-traffic optimization of achievable performance. *Oper. Res.*, 49(4):609–623, 2001.
- [118] Peter W. Glynn and Assaf Zeevi. *Bounding Stationary Expectations of Markov Processes*, page 195–214. Institute of Mathematical Statistics, 2008.
- [119] Caner Gocmen, Thodoris Lykouris, Deeksha Sinha, and Wentao Weng. Scheduling in queueing systems with uncertain and evolving holding costs. *arXiv:2505.21331 [cs.DS]*, 2025.
- [120] Alan J Goldman and Albert W Tucker. Theory of linear programming. *Linear inequalities and related systems*, 38:53–97, 1956.
- [121] Diego Goldsztajn and Konstantin Avrachenkov. Asymptotically optimal policies for weakly coupled Markov decision processes. *arXiv:2406.04751 [math.OC]*, 2024.
- [122] Google. Google Kubernetes Engine. accessed Jan. 2023.
- [123] Google Cloud. Virtual machine instances.
- [124] Google Cloud. Overcommitting CPUs on sole-tenant VMs.
- [125] Isaac Groszof, Ziv Scully, and Mor Harchol-Balter. SRPT for multiserver systems. *Perform. Eval.*, 127–128:154–175, 2018.
- [126] Isaac Groszof, Ziv Scully, and Mor Harchol-Balter. Load balancing guardrails: Keeping your heavy traffic on the road to low response times. *Proc. ACM Meas. Anal. Comput. Syst.*, 3(2), 2019.

- [127] Isaac Grosof, Mor Harchol-Balter, and Alan Scheller-Wolf. Stability for Two-class Multiserver-job Systems. *arXiv:2010.00631 [cs.PF]*, 2020.
- [128] Isaac Grosof, Mor Harchol-Balter, and Alan Scheller-Wolf. WCFS: a new framework for analyzing multiserver systems. *Queueing Syst.*, 102(1):143–174, 2022.
- [129] Isaac Grosof, Ziv Scully, Mor Harchol-Balter, and Alan Scheller-Wolf. Optimal scheduling in the multiserver-job model under heavy traffic. *Proc. ACM Meas. Anal. Comput. Syst.*, 6(3), dec 2022.
- [130] Isaac Grosof, Yige Hong, Mor Harchol-Balter, and Alan Scheller-Wolf. The RESET and MARC techniques, with application to multiserver-job analysis. *Perform. Eval.*, 162:102378, 2023.
- [131] Jin Guang, Xinyun Chen, and J. G. Dai. Uniform moment bounds for generalized Jackson networks in multiscale heavy traffic. *Math. Oper. Res.*, 51(1):668–685, 2026.
- [132] Varun Gupta and Ana Radovanović. Interior-point-based online stochastic bin packing. *Oper. Res.*, 68(5):1474–1492, 2020.
- [133] Itai Gurvich, Junfei Huang, and Avishai Mandelbaum. Excursion-based universal approximations for the Erlang-A queue in steady-state. *Math. Oper. Res.*, 39(2):325–373, 2014.
- [134] Bruce Hajek. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Adv. Appl. Probab.*, 14(3):502–525, 1982.
- [135] Shlomo Halfin and Ward Whitt. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.*, 29(3):567–588, 1981.
- [136] Mor Harchol-Balter. *Network Analysis without Exponentiality Assumptions*. PhD thesis, University of California, Berkeley, Berkeley, CA, 1996.
- [137] Mor Harchol-Balter. *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge University Press, New York, NY, 1st edition, 2013.
- [138] Mor Harchol-Balter. Open problems in queueing theory inspired by datacenter computing. *Queueing Syst.*, 97(1):3–37, 2021.
- [139] Mor Harchol-Balter and Allen B. Downey. Exploiting process lifetime distributions for dynamic load balancing. *ACM Transactions on Computer Systems*, 15(3):253–285, 1997.

- [140] J. Michael Harrison and Marcel J. López. Heavy traffic resource pooling in parallel-server systems. *Queueing Syst.*, 33(4):339–368, 1999.
- [141] J. Michael Harrison and Assaf Zeevi. Dynamic scheduling of a multiclass queue in the Halfin-Whitt heavy traffic regime. *Oper. Res.*, 52(2):1–31, 2004.
- [142] Jeffrey Thomas Hawkins. *A Lagrangian decomposition approach to weakly coupled dynamic optimization problems and its applications*. PhD thesis, Massachusetts Institute of Technology, 2003.
- [143] Qi-Ming He and Elizabeth Jewkes. Flow time in the $MAP/G/1$ queue with customer batching and setup times. *Communications in Statistics. Stochastic Models*, 11(4):691–711, 1995.
- [144] D. J. Hodge and K. D. Glazebrook. On the asymptotic optimality of greedy index heuristics for multi-action restless bandits. *Adv. Appl. Probab.*, 47(3):652–667, 2015.
- [145] Per Hokstad. Relations for the workload of the $GI/G/s$ queue. *Adv. Appl. Probab.*, 17(4):887–904, 1985.
- [146] Yige Hong and Ziv Scully. Performance of the Gittins policy in the $G/G/1$ and $G/G/k$, with and without setup times. *Perform. Eval.*, 163: 102377, 2024.
- [147] Yige Hong and Weina Wang. Sharp waiting-time bounds for multiserver jobs. In *ACM Int. Symp. Mobile Ad Hoc Networking and Computing (MobiHoc)*, pages 161–170, New York, NY, 2022.
- [148] Yige Hong and Weina Wang. Sharp waiting-time bounds for multiserver jobs. *Stoch. Syst.*, 14(4):455–478, 2024.
- [149] Yige Hong, Qiaomin Xie, Yudong Chen, and Weina Wang. Restless bandits with average reward: Breaking the uniform global attractor assumption. In *Advances in Neural Information Processing Systems*, volume 36, pages 12810–12844, 2023.
- [150] Yige Hong, Qiaomin Xie, and Weina Wang. Near-optimal stochastic bin-packing in large service systems with time-varying item sizes. *Proc. ACM Meas. Anal. Comput. Syst.*, 7(3), dec 2023.
- [151] Yige Hong, Qiaomin Xie, Yudong Chen, and Weina Wang. Achieving exponential asymptotic optimality in average-reward restless bandits without global attractor assumption. *arXiv:2405.17882 [cs.LG]*, 2024.

- [152] Yige Hong, Qiaomin Xie, Yudong Chen, and Weina Wang. Unichain and aperiodicity are sufficient for asymptotic optimality of average-reward restless bandits.
- [153] Yige Hong, Qiaomin Xie, Yudong Chen, and Weina Wang. Unichain and aperiodicity are sufficient for asymptotic optimality of average-reward restless bandits. *Math. Oper. Res.*, 2025.
- [154] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2013.
- [155] Weici Hu and Peter Frazier. An asymptotically optimal index policy for finite-horizon restless bandits. *arXiv:1707.00205 [math.OC]*, 2017.
- [156] P. J. Hunt and T. G. Kurtz. Large loss networks. *Stoch. Proc. Appl.*, 53(2):363 – 378, 1994.
- [157] P. J. Hunt and C. N. Laws. Optimization via trunk reservation in single resource loss systems under heavy traffic. *Ann. Appl. Probab.*, 7(4): 1058–1079, 1997.
- [158] Daniela Hurtado-Lange and Siva Theja Maguluri. Transform methods for heavy-traffic analysis. *Stoch. Syst.*, 10(4):275–309, 2020.
- [159] Donald L. Iglehart and Ward Whitt. Multiple channel queues in heavy traffic. I. *Adv. Appl. Probab.*, 2(1):150–177, 1970.
- [160] Patrick Jaillet, Jiashuo Jiang, Konstantina Mellou, Marco Molinaro, Chara Podimata, and Zijie Zhou. Online scheduling for LLM inference with KV cache constraints. *arXiv:2502.07115 [cs.LG]*, 2026.
- [161] A. J. E. M. Janssen, J. S. H. van Leeuwen, and B. Zwart. Corrected asymptotics for a multi-server queue in the Halfin-Whitt regime. *Queueing Syst.*, 58(4):261–301, 2008.
- [162] A. J. E. M. Janssen, J. S. H. van Leeuwen, and Bert Zwart. Refining square-root safety staffing by expanding Erlang C. *Oper. Res.*, 59(6): 1512–1522, 2011.
- [163] Predrag Jelenković, Avishai Mandelbaum, and Petar Momčilović. Heavy traffic limits for queues with many deterministic servers. *Queueing Syst.*, 47(1):53–69, 2004.
- [164] Xinghu Jin, Guodong Pang, Lihu Xu, and Xin Xu. An approximation to the invariant measure of the limiting diffusion of G/Ph/n + GI queues in the Halfin-Whitt regime and related asymptotics. *Math. Oper. Res.*,

- 50(2):783–812, 2025.
- [165] Offer Kella and Wolfgang Stadjé. Superposition of renewal processes and an application to multi-server queues. *Statistics & Probability Letters*, 76(17):1914–1924, 2006.
 - [166] Jack Kiefer and Jacob Wolfowitz. On the theory of queues with many servers. *Transactions of the American Mathematical Society*, 78(1):1–18, 1955.
 - [167] Jackson A Killian, Arpita Biswas, Sanket Shah, and Milind Tambe. Q-learning Lagrange policies for multi-action restless bandits. In *Proc. Ann. ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD)*, pages 871–881, 2021.
 - [168] Jackson A Killian, Lily Xu, Arpita Biswas, and Milind Tambe. Restless and uncertain: Robust policies for restless bandits via deep multi-agent reinforcement learning. In *Conf. Uncertainty in Artificial Intelligence (UAI)*, pages 990–1000, 2022.
 - [169] J. F. C. Kingman. Inequalities in the theory of queues. *J. Roy. Stat. Soc. B Met.*, 32(1):102–110, 12 1970.
 - [170] John F C Kingman. On queues in heavy traffic. *J. Roy. Stat. Soc. B Met.*, 24(2):383–392, 1962.
 - [171] John F C Kingman. Some inequalities for the queue GI/G/1. *Biometrika*, 49(3-4):315–324, 1962.
 - [172] Leonard Kleinrock. *Queueing Systems*. John Wiley & Son, 1975.
 - [173] Julian Köllerström. Heavy traffic theory for queues with several servers. I. *J. Appl. Probab.*, 11(3):544–552, 1974.
 - [174] Julian Köllerström. Heavy traffic theory for queues with several servers. II. *J. Appl. Probab.*, 16(2):393–401, 1979.
 - [175] Peter A. Lachenbruch and M. Ray Mickey. Estimation of error rates in discriminant analysis. *Technometrics*, 10(1):1–11, 1968.
 - [176] Maialen Larrañaga, Urtzi Ayesta, and Ina Maria Verloop. Dynamic control of birth-and-death restless bandits: Application to resource-allocation problems. *IEEE/ACM Trans. Netw.*, 24(6):3812–3825, 2016.
 - [177] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

- [178] Huan Li and Yixin Zhu. A new approach to the $G/G/1$ queue with generalized setup time and exhaustive service. *J. Appl. Probab.*, 31(4): 1083–1097, 1994.
- [179] Jingwen Li and Jihong Ou. Characterizing the idle-period distribution of $GI/G/1$ queues. *J. Appl. Probab.*, 32(1):247–255, 1995.
- [180] Yuan Li and David A. Goldberg. Simple and explicit bounds for multiserver queues with $1/(1 - \rho)$ scaling. *Math. Oper. Res.*, 50(2): 813–837, 2025.
- [181] Yusen Li, Xueyan Tang, and Wentong Cai. On dynamic bin packing for resource allocation in the cloud. In *Proc. Ann. ACM Symp. Parallelism in Algorithms and Architectures (SPAA)*, page 2–11, Prague, Czech Republic, 2014.
- [182] Zhouzi Li, Mor Harchol-Balter, and Benjamin Berg. Mean field optimal core allocation across malleable jobs. *arXiv:2602.01411 [cs.DC]*, 2026.
- [183] Minghong Lin, Adam Wierman, and Bert Zwart. Heavy-traffic analysis of mean response time under shortest remaining processing time. *Perform. Eval.*, 68(10):955–966, 2011.
- [184] Sung-Han Lin, Marco Paolieri, Cheng-Fu Chou, and Leana Golubchik. A model-based approach to streamlining distributed training for asynchronous SGD. In *IEEE Int. Symp. Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS)*, pages 306–318, Los Alamitos, CA, 2018.
- [185] John D. C. Little. Little’s law as viewed on its 50th anniversary. *Oper. Res.*, 59(3):536–549, 2011.
- [186] Xin Liu. *Steady State Analysis of Load Balancing Algorithms in the Heavy Traffic Regime*. PhD thesis, Arizona State University, 2019.
- [187] Xin Liu and Lei Ying. Steady-state analysis of load-balancing algorithms in the sub-Halfin–Whitt regime. *J. Appl. Probab.*, 57(2):578–596, 2020.
- [188] Xin Liu and Lei Ying. Universal scaling of distributed queues under load balancing in the super-Halfin–Whitt regime. *IEEE/ACM Trans. Netw.*, 30(1):190–201, 2022.
- [189] Xin Liu, Kang Gong, and Lei Ying. Steady-state analysis of load balancing with Coxian-2 distributed service times. *Nav. Res. Log.*, 69(1):57–75, 2022.

- [190] David Lo, Liquan Cheng, Rama Govindaraju, Parthasarathy Ranganathan, and Christos Kozyrakis. Heracles: Improving resource efficiency at scale. In *Prof. ACM/IEEE Ann. Int. Symp. Computer Architecture (ISCA)*, pages 450–462, 2015.
- [191] Gary Lorden. On excess over the boundary. *Ann. Math. Statist.*, 41(2): 520–527, 1970.
- [192] Richard Loulou. Multi-channel queues in heavy traffic. *J. Appl. Probab.*, 10(4):769–777, 1973.
- [193] Robert M. Loynes. The stability of a queue with non-independent inter-arrival and service times. *Mathematical Proceedings of the Cambridge Philosophical Society*, 58(3):497–520, 1962.
- [194] Siva Theja Maguluri and R. Srikant. Scheduling jobs with unknown duration in clouds. *IEEE/ACM Trans. Netw.*, 22(6):1938–1951, 2014.
- [195] Siva Theja Maguluri and R. Srikant. Heavy traffic queue length behavior in a switch under the MaxWeight algorithm. *Stoch. Syst.*, 6(1):211–250, 2016.
- [196] Siva Theja Maguluri, R. Srikant, and Lei Ying. Stochastic models of load balancing and scheduling in cloud computing clusters. In *Proc. IEEE Int. Conf. Computer Communications (INFOCOM)*, pages 702–710, Los Alamitos, CA, 2012.
- [197] Siva Theja Maguluri, R. Srikant, and Lei Ying. Heavy traffic optimal resource allocation algorithms for cloud computing clusters. *Perform. Eval.*, 81:20–39, 2014.
- [198] Avi Mandelbaum, William A. Massey, and Martin I. Reiman. Strong approximations for Markovian service networks. *Queueing Syst.*, 30(1): 149–201, 1998.
- [199] Avishai Mandelbaum and Petar Momčilović. Queues with many servers: The virtual waiting-time process in the QED regime. *Math. Oper. Res.*, 33(3):561–586, 2008.
- [200] Shie Mannor and John N. Tsitsiklis. On the empirical state-action frequencies in Markov decision processes under general policies. *Math. Oper. Res.*, 30(3):545–561, 2005.
- [201] Judith J McDonald and Pietro Paparella. A short and elementary proof of Brauer’s theorem. *Teach. Math.*, 24:85–86, 2021.

- [202] H. P. McKean. A class of Markov processes associated with nonlinear parabolic equations. *Proceedings of the National Academy of Sciences*, 56(6):1907–1911, 1966.
- [203] Agassi Melikov. Computation and optimization methods for multire-source queues. *Cybern. Syst. Anal.*, 32(6):821–836, 1996.
- [204] Konstantina Mellou, Marco Molinaro, and Rudy Zhou. The power of migrations in dynamic bin packing. *arXiv:2408.13178 [cs.DS]*, 2024.
- [205] Sean Meyn. *Control Techniques for Complex Networks*. Cambridge University Press, USA, 1st edition, 2007.
- [206] Sean Meyn, Richard L. Tweedie, and Peter W. Glynn. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2nd edition, 2009.
- [207] Sean P. Meyn and R. L. Tweedie. Stability of Markovian processes III: Foster–Lyapunov criteria for continuous-time processes. *Adv. Appl. Probab.*, 25(3):518–548, 1993.
- [208] Rupert G. Miller. A contribution to the theory of bulk queues. *J. Roy. Statist. Soc. Ser. B*, 21(2):320–337, 1959.
- [209] Michael Mitzenmacher and Rana Shahout. Queueing, predictions, and LLMs: Challenges and open problems. *arXiv:2503.07545 [cs.AI]*, 2025.
- [210] Masakiyo Miyazawa. Decomposition formulas for single server queues with vacations : A unified approach by the rate conservation law. *Commun. Stat. Stoch. Models*, 10(2):389–413, 1994.
- [211] Masakiyo Miyazawa. Rate conservation laws: A survey. *Queueing Syst.*, 15(1):1–58, 1994.
- [212] Masakiyo Miyazawa. Diffusion approximation for stationary analysis of queues and their networks: A review. *J. Operat. Res. Soc. Japan*, 58(1):104–148, 2015.
- [213] MASAO Mori. Some bounds for queues. *J. Operat. Res. Soc. Japan*, 18:152–181, 1975.
- [214] Evsey Morozov and Alexander S. Rumyantsev. Stability analysis of a MAP/M/s cluster model by matrix-analytic method. In *European Workshop Computer Performance Engineering (EPEW)*, volume 9951, pages 63–76, Cham, Switzerland, 2016.
- [215] Evsey Morozov and Bart Steyaert. *Stability Analysis of Regenerative Queueing Models: Mathematical Methods and Applications*. Springer,

Cham, Switzerland, 2021.

- [216] Khaled Nakhleh and I-Hong Hou. DeepTOP: Deep threshold-optimal policy for MDPs and RMABs. In *Conf. Neural Information Processing Systems (NeurIPS)*, pages 28734–28746, 2022.
- [217] Khaled Nakhleh, Santosh Ganji, Ping-Chun Hsieh, I-Hong Hou, and Srinivas Shakkottai. NeurWIN: Neural Whittle index network for restless bandits via deep RL. In *Conf. Neural Information Processing Systems (NeurIPS)*, pages 828–839, 2021.
- [218] José Niño-Mora. Markovian restless bandits and index policies: A review. *Mathematics*, 11(7):1639, 2023.
- [219] Diletta Olliaro, Marco Ajmone Marsan, Simonetta Balsamo, and Andrea Marin. The saturated multiserver job queuing model with two classes of jobs: Exact and approximate results. *Perform. Eval.*, 162:102370, 2023.
- [220] Christos H. Papadimitriou and John N. Tsitsiklis. The complexity of optimal queuing network control. *Math. Oper. Res.*, 24(2):293–305, 1999.
- [221] Kihong Park and Walter Willinger, editors. *Self-Similar Network Traffic and Performance Evaluation*. Wiley, New York, NY, 2000.
- [222] Jamol Pender and Tuan Phung-Duc. A law of large numbers for M/M/c/Delayoff-Setup queues with nonstationary arrivals. In *Analytical and Stochastic Modelling Techniques and Applications*, volume 9845, pages 253–268. Springer, Cham, Switzerland, 2016.
- [223] David L. Peterson. Data center I/O patterns and power laws. In *22nd International Computer Measurement Group Conference*, pages 1034–1045, San Diego, CA, 1996.
- [224] Leonid Ponomarenko, Che Soong Kim, and Agassi Melikov. *Performance analysis and optimization of multi-traffic on communication networks*. Springer Science & Business Media, 2010.
- [225] Konstantinos Psychas and Javad Ghaderi. On non-preemptive VM scheduling in the cloud. *ACM SIGMETRICS Perform. Evaluation Rev.*, 46(1):67–69, jun 2018.
- [226] Konstantinos Psychas and Javad Ghaderi. Scheduling jobs with random resource requirements in computing clusters. In *Proc. IEEE Int.*

- Conf. Computer Communications (INFOCOM)*, pages 2269–2277, Los Alamitos, CA, 2019.
- [227] Konstantinos Psychas and Javad Ghaderi. High-throughput bin packing: Scheduling jobs with random resource demands in clusters. *IEEE/ACM Trans. Netw.*, 29(1):220–233, 2021.
 - [228] Konstantinos Psychas and Javad Ghaderi. A theory of auto-scaling for resource reservation in cloud services. *Stoch. Syst.*, 12(3):227–252, 2022.
 - [229] A. A. Puhalskii and M. I. Reiman. The multiclass GI/PH/N queue in the Halfin-Whitt regime. *Adv. Appl. Probab.*, 32(2):564–595, 2000.
 - [230] Martin L Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, 2005.
 - [231] Prakirt Raj Jhunjhunwala, Daniela Hurtado-Lange, and Siva Theja Maguluri. Exponential tail bounds on queues: A confluence of non-asymptotic heavy traffic and large deviations. *ACM SIGMETRICS Perform. Evaluation Rev.*, 51(4):18–19, 2024.
 - [232] Josh Reed. The G/GI/N queue in the Halfin-Whitt regime. *Ann. Appl. Probab.*, 19(6):2211 – 2269, 2009.
 - [233] Martin I. Reiman. Some diffusion approximations with state space collapse. In *Modelling and Performance Evaluation Methodology*, pages 207–240, Berlin, Heidelberg, 1984.
 - [234] Charles Reiss, Alexey Tumanov, Gregory R. Ganger, Randy H. Katz, and Michael A. Kozuch. Heterogeneity and dynamicity of clouds at scale: Google trace analysis. In *Proc. ACM Symp. Cloud Computing (SoCC)*, San Jose, CA, 2012.
 - [235] Rhonda Richter and J. George Shanthikumar. Scheduling multiclass single server queueing systems to stochastically maximize the number of successful departures. *Probab. Eng. Inf. Sci.*, 3(3):323–333, 1989.
 - [236] Rhonda Richter, J. George Shanthikumar, and Genji Yamazaki. On extremal service disciplines in single-stage queueing systems. *J. Appl. Probab.*, 27(2):409–416, 1990.
 - [237] Francisco Robledo, Vivek Borkar, Urtzi Ayesta, and Konstantin Avrachenkov. QWI: Q-learning with Whittle index. *ACM SIGMETRICS Perform. Evaluation Rev.*, 49(2):47–50, 2022.

- [238] Alexander Rumyantsev and Evsey Morozov. Stability criterion of a multiserver model with simultaneous service. *Ann. Oper. Res.*, 252(1): 29–39, 2017.
- [239] Alexander Rumyantsev, Robert Basmadjian, Sergey Astafiev, and Alexander Golovin. Three-level modeling of a speed-scaling super-computer. *Annals of Operations Research*, 2022.
- [240] Krzysztof Rządca, Paweł Findeisen, Jacek Swiderski, Przemysław Zych, Przemysław Broniek, Jarek Kusmirek, Paweł Nowak, Beata Strack, Piotr Witusowski, Steven Hand, and John Wilkes. Autopilot: Workload autoscaling at Google. In *Proc. European Conf. Computer Systems (EuroSys)*, Heraklion, Greece, 2020.
- [241] Linus E. Schrage. A proof of the optimality of the shortest remaining processing time discipline. *Oper. Res.*, 16(3):687–690, 1968.
- [242] Ziv Scully. *A New Toolbox for Scheduling Theory*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 2022.
- [243] Ziv Scully and Mor Harchol-Balter. The Gittins policy in the M/G/1 queue. In *19th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt 2021)*, pages 248–255, Philadelphia, PA, 2021.
- [244] Ziv Scully, Mor Harchol-Balter, and Alan Scheller-Wolf. SOAP: One clean analysis of all age-based scheduling policies. *Proc. ACM Meas. Anal. Comput. Syst.*, 2(1), 2018.
- [245] Ziv Scully, Isaac Grosf, and Mor Harchol-Balter. The Gittins policy is nearly optimal in the M/G/ k under extremely general conditions. *Proc. ACM Meas. Anal. Comput. Syst.*, 4(3), 2020.
- [246] Ziv Scully, Mor Harchol-Balter, and Alan Scheller-Wolf. Simple near-optimal scheduling for the M/G/1. *Proc. ACM Meas. Anal. Comput. Syst.*, 4(1), 2020.
- [247] Ziv Scully, Lucas van Kreveld, Onno J. Boxma, Jan-Pieter Dorsman, and Adam Wierman. Characterizing policies with optimal response time tails under heavy-tailed job sizes. *Proc. ACM Meas. Anal. Comput. Syst.*, 4(2), 2020.
- [248] Ziv Scully, Isaac Grosf, and Mor Harchol-Balter. Optimal multiserver scheduling with unknown job sizes in heavy traffic. *Perform. Eval.*, 145, 2021.

- [249] Ziv Scully, Isaac Grosof, and Michael Mitzenmacher. Uniform bounds for scheduling with job size estimates. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, Berkeley, CA, 2022.
- [250] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *J. Mach. Learn. Res.*, 11:2635–2670, 2010.
- [251] Karl Sigman. One-dependent regenerative processes and queues in continuous time. *Math. Oper. Res.*, 15(1):175–189, 1990.
- [252] Alexander L. Stolyar. An infinite server system with general packing constraints. *Oper. Res.*, 61(5):1200–1217, 2013.
- [253] Alexander L. Stolyar. Large-scale heterogeneous service systems with general packing constraints. *Adv. Appl. Probab.*, 49:61–83, 2017.
- [254] Alexander L Stolyar and Yuan Zhong. A large-scale service system with packing constraints: Minimizing the number of occupied servers. *ACM SIGMETRICS Perform. Evaluation Rev.*, 41(1):41–52, 2013.
- [255] Alexander L. Stolyar and Yuan Zhong. Asymptotic optimality of a greedy randomized algorithm in a large-scale service system with general packing constraints. *Queueing Syst.*, 79:117–143, 2015.
- [256] Alexander L. Stolyar and Yuan Zhong. A service system with packing constraints: Greedy randomized algorithm achieving sublinear in scale optimality gap. *Stoch. Syst.*, 11:83–111, 2021.
- [257] Oleg M. Tikhonenko. Generalized Erlang problem for service systems with finite total capacity. *Probl. Inf. Transm.*, 41(3):243–253, 2005.
- [258] Muhammad Tirmazi, Adam Barker, Nan Deng, Md E. Haque, Zhi-jing Gene Qin, Steven Hand, Mor Harchol-Balter, and John Wilkes. Borg: the next generation. In *Proc. European Conf. Computer Systems (EuroSys)*, pages 30:1–30:14, New York, NY, 2020.
- [259] John N. Tsitsiklis. A short proof of the Gittins index theorem. *Ann. Appl. Probab.*, 4(1):194 – 199, 1994.
- [260] Shweta Upadhyaya. Queueing systems with vacation: An overview. *International Journal of Mathematics in Operational Research*, 9(2): 167, 2016.
- [261] Mark van der Boor, Martin Zubeldia, and Sem Borst. Zero-wait load balancing with sparse messaging. *Oper. Res. Lett.*, 48(3):368–375, 2020.

- [262] Nico M. van Dijk. Blocking of finite source inputs which require simultaneous servers with general think and holding times. *Oper. Res. Lett.*, 8(1):45 – 52, 1989.
- [263] Rocco van Vreumingen. Queueing systems with non-standard service policies and server vacations. Master’s thesis, University of Amsterdam, Amsterdam, The Netherlands, 2019.
- [264] P. Varaiya, J. Walrand, and C. Buyukkoc. Extensions of the multiarmed bandit problem: The discounted case. *IEEE Trans. Autom. Control*, 30(5):426–439, 1985.
- [265] Michael H. Veatch and Lawrence M. Wein. Scheduling a make-to-stock queue: Index policies and hedging points. *Oper. Res.*, 44(4):634–647, 1996.
- [266] I. M. Verloop. Asymptotically optimal priority policies for indexable and nonindexable restless bandits. *Ann. Appl. Probab.*, 26(4):1947–1995, 2016.
- [267] Abhishek Verma, Luis Pedrosa, Madhukar Korupolu, David Oppenheimer, Eric Tune, and John Wilkes. Large-scale cluster management at Google with Borg. In *Proc. European Conf. Computer Systems (EuroSys)*, page 18, Bordeaux, France, 2015.
- [268] Sofía S Villar. Indexability and optimal index policies for a class of reinitialising restless bandits. *Probab. Eng. Inf. Sci.*, 30(1):1–23, 2016.
- [269] Sofía S. Villar, Jack Bowden, and James Wason. Multi-armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges. *Statistical Science*, 30(2):199 – 215, 2015.
- [270] Weina Wang, Mor Harchol-Balter, Haotian Jiang, Alan Scheller-Wolf, and R. Srikant. Delay asymptotics and bounds for multitask parallel jobs. *Queueing Syst.*, 91(3):207–239, 2019.
- [271] Weina Wang, Qiaomin Xie, and Mor Harchol-Balter. Zero queueing for multi-server jobs. *Proc. ACM Meas. Anal. Comput. Syst.*, 5(1), 2021.
- [272] Weina Wang, Siva Theja Maguluri, R. Srikant, and Lei Ying. Heavy-traffic insensitive bounds for weighted proportionally fair bandwidth sharing policies. *Math. Oper. Res.*, 47(4):2691–2720, 2022.
- [273] Richard Weber. On the Gittins index for multiarmed bandits. *Ann. Appl. Probab.*, 2(4):1024 – 1033, 1992.

- [274] Richard R. Weber and Gideon Weiss. On an index policy for restless bandits. *J. Appl. Probab.*, 27(3):637–648, 1990.
- [275] Peter D. Welch. On a generalized M/G/1 queuing process in which the first customer of each busy period receives exceptional service. *Oper. Res.*, 12(5):736–752, 1964.
- [276] Wentao Weng and Weina Wang. Achieving zero asymptotic queueing delay for parallel jobs. *Proc. ACM Meas. Anal. Comput. Syst.*, 4(3), 2020.
- [277] Wentao Weng, Xingyu Zhou, and R. Srikant. Optimal load balancing with locality constraints. *Proc. ACM Meas. Anal. Comput. Syst.*, 4(3), nov 2020.
- [278] Ward Whitt. Existence of limiting distributions in the $GI/GI/s$ queue. *Math. Oper. Res.*, 7(1):88–94, 1982.
- [279] Ward Whitt. Blocking when service is required from several facilities simultaneously. *AT&T Tech. J.*, 64(8):1807 – 1856, 1985.
- [280] Ward Whitt. Understanding the efficiency of multi-server service systems. *Manage. Sci.*, 38(5):708–723, 1992.
- [281] Ward Whitt. *Stochastic-Process Limits*. Springer Series in Operations Research and Financial Engineering. Springer, New York, NY, 2002 edition, 2002.
- [282] Peter Whittle. Multi-armed bandits and the Gittins index. *J. Roy. Stat. Soc. B Met.*, 42(2):143–149, 12 1980.
- [283] Peter Whittle. Restless bandits: activity allocation in a changing world. *J. Appl. Probab.*, 25:287 – 298, 1988.
- [284] Adam Wierman, Mor Harchol-Balter, and Takayuki Osogami. Nearly insensitive bounds on SMART scheduling. *ACM SIGMETRICS Performance Evaluation Review*, 33(1):205–216, 2005.
- [285] John Wilkes. Google cluster-usage traces v3.
- [286] Jalani K. Williams, Mor Harchol-Balter, and Weina Wang. The M/M/k with deterministic setup times. *Proc. ACM Meas. Anal. Comput. Syst.*, 6(3):1–45, 2022.
- [287] Ronald W. Wolff and Chia-Li Wang. Idle period approximations and bounds for the GI/G/1 queue. *Adv. Appl. Probab.*, 35(3):773–792, 2003.

- [288] D. Worthington. Reflections on queue modelling from the last 50 years. *J. Oper. Res. Soc.*, 60(1):S83–S92, 2009.
- [289] Tianhao Wu, Matthew Zurek, Weina Wang, and Qiaomin Xie. Lyapunov-based sample complexity analysis for weakly-coupled MDPs. In *Proc. Conf. Learning Theory (COLT)*, 2026.
- [290] Tianhao Wu, Matthew Zurek, Weina Wang, and Qiaomin Xie. Lyapunov-based sample complexity analysis for weakly-coupled MDPs (extended abstract). In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pages 6852–6857, 2026.
- [291] Qiaomin Xie, Xiaobo Dong, Yi Lu, and Rayadurgam Srikant. Power of d choices for large-scale bin packing: A loss model. *ACM SIGMETRICS Perform. Evaluation Rev.*, 43(1):321–334, jun 2015.
- [292] Guojun Xiong and Jian Li. Finite-time analysis of Whittle index based Q-learning for restless multi-armed bandits with neural network function approximation. In *Conf. Neural Information Processing Systems (NeurIPS)*, pages 29048–29073, 2023.
- [293] Guojun Xiong, Shufan Wang, and Jian Li. Learning infinite-horizon average-reward restless multi-action bandits via index awareness. In *Conf. Neural Information Processing Systems (NeurIPS)*, pages 17911–17925, 2022.
- [294] Chen Yan. An optimal-control approach to infinite-horizon restless bandits: Achieving asymptotic optimality with minimal assumptions. In *Proc. IEEE Conf. Decision and Control (CDC)*, pages 6665–6672, 2024.
- [295] Zhe Yu, Yunjian Xu, and Lang Tong. Deadline scheduling as restless bandits. *IEEE Trans. Autom. Control*, 63(8):2343–2358, 2018.
- [296] Gabriel Zayas-Cabán, Stefanus Jasin, and Guihua Wang. An asymptotically optimal heuristic for general nonstationary finite-horizon restless multi-armed, multi-action bandits. *Adv. Appl. Probab.*, 51(3):745–772, 2019.
- [297] Jingwei Zhang. Leveraging nondegeneracy in dynamic resource allocation. *Available at SSRN 4920428*, 2024.
- [298] Xiangcheng Zhang, Yige Hong, and Weina Wang. Projection-based Lyapunov method for fully heterogeneous weakly-coupled MDPs.

- [299] Xiangcheng Zhang, Yige Hong, and Weina Wang. Projection-based Lyapunov method for fully heterogeneous weakly-coupled MDPs. In *Advances in Neural Information Processing Systems*, volume 38, pages 91054–91076, 2025.
- [300] Xiangyu Zhang and Peter I. Frazier. Restless bandits with many arms: Beating the central limit theorem. *arXiv:2107.11911 [math.OC]*, 2021.
- [301] Xiangyu Zhang and Peter I. Frazier. Near-optimality for infinite-horizon restless bandits with many arms. *arXiv:2203.15853 [cs.LG]*, 2022.
- [302] Zhisheng Zhao and Debankur Mukherjee. Optimal rate-matrix pruning for heterogeneous systems. *ACM SIGMETRICS Perform. Evaluation Rev.*, 51(4):26–27, 2024.
- [303] Zhisheng Zhao, Debankur Mukherjee, and Ruoyu Wu. Exploiting data locality to improve performance of heterogeneous server clusters. *Stoch. Syst.*, 14(3):229–272, 2024.
- [304] Jiahong Zhou, Shunhui Mao, Guoliang Yang, Bo Tang, Qianlong Xie, Lebin Lin, Xingxing Wang, and Dong Wang. RL-MPCA: A reinforcement learning based multi-phase computation allocation approach for recommender systems. In *Proceedings of the ACM Web Conference 2023*, pages 3214–3224, 2023.
- [305] Martin Zubeldia. Delay-optimal policies in partial fork-join systems with redundancy and random slowdowns. *Proc. ACM SIGMETRICS Int. Conf. Measurement and Modeling of Computer Systems*, 4(1), 2020.
- [306] Noa Zychlinski, Carri W. Chan, and Jing Dong. Managing queues with different resource requirements. *Oper. Res.*, 71(4):1387–1413, 2023.