Dynamic Multi-agent Autonomous Systems for Societal Transformation

Chinmay Maheshwari

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2025-138
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2025/EECS-2025-138.html

June 18, 2025

Acknowledgement

Thanks to UC Berkeley for the academic ecosystem.

Dynamic Multi-agent Autonomous Systems for Societal Transformation

By

Chinmay Maheshwari

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Shankar Sastry, Chair
Professor Claire Tomlin
Professor Chris Shannon
Professor Manxi Wu

Summer 2025

Dynamic Multi-agent Autonomous Systems for Societal Transformation

Abstract

Dynamic Multi-agent Autonomous Systems for Societal Transformation

by

Chinmay Maheshwari

Doctor of Philosophy in Engineering - Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Shankar Sastry, Chair

Autonomous AI technologies are increasingly embedded in critical societal systems, including robotics, transportation, logistics, and energy—where they enable large-scale, data-driven decision-making. While recent advances have enabled autonomous agents to perform effectively in isolated or structured environments, a fundamental open challenge is to integrate such agents into dynamic, uncertain, and resource-constrained multi-agent environments where they must learn and interact strategically with other autonomous systems and with humans. The emerging outcomes not only impact the individual utility but also impacts societal efficiency, equity and safety.

This dissertation addresses the design and analysis of intelligent autonomous agents in such multi-agent societal settings. It is motivated by two central questions: (1) How can we design learning and decision-making algorithms that allow autonomous agents to act rationally and strategically in the presence of other agents? (2) How can we ensure that the collective outcomes of such agent interactions align with broader societal goals such as efficiency, equity, and safety?

To answer these questions, the dissertation introduces new theoretical, algorithmic, and computational frameworks for multi-agent learning, decision-making, and design of multi-agent interactions in societal systems. These contributions are organized into four parts, each grounded in application domains that highlight key challenges and propose novel solutions.

Part I focuses on learning in general-sum Markov games, which model multi-agent interactions in uncertain, dynamic environments. Unlike classical control or reinforcement learning settings that assume either fully cooperative or fully adversarial interactions, many real-world systems exhibit a mix of cooperative and competitive behavior. To address this, we propose a new theoretical framework of *Markov near-potential games*, which approximates the underlying multi-agent interaction using a potential game. We leverage this framework to design and analyze multi-agent learning algorithms. Specifically, we use it to design real-

time, high-performance strategies for autonomous multi-car racing that outperform several existing baselines. Additionally, we use the framework to characterize the long-run outcomes of interactions between decentralized reinforcement learning algorithms, with a focus on actor-critic methods.

Part II examines strategic learning under competition induced due to shared resource and infrastructure constraints, including settings with congestion. The focus is on domains such as transportation networks and two-sided matching markets, where agents compete over scarce, congestible resources. This part introduces learning dynamics that achieve desirable performance guarantees—such as low regret and equilibrium convergence—even when agents adapt based on local observations and uncertain feedback.

Part III shifts from agent-level optimization to designing mechanisms to align strategic agent behavior with societal objectives. A key challenge here is that agents may respond strategically to deployed mechanisms, leading to distribution shifts, while designers often lack access to private agent preferences. This part proposes data-driven methods for design of societal mechanisms that remain robust to strategic behavior and result in socially beneficial outcome. We highlight applications in design of congestion pricing on road networks and design of data-driven online services.

Part IV explores market design for the emerging Advanced Air Mobility (AAM)—a future mobility paradigm involving UAVs and air taxis operating in low-altitude urban airspace. Given the decentralized and adaptive nature of AAM systems, traditional centralized air traffic control methods are inadequate. This part introduces market-based mechanisms for allocating trajectories to UAVs with potentially heterogeneous preferences that ensure safety, fairness, and efficiency.

Overall, the dissertation offers new theoretical insights, algorithmic tools, and practical mechanisms for ensuring that future autonomous systems are not only efficient in maximizing individual utility, but also result in socially efficient, equitable and safe outcomes.

To my father, Sh. Naresh Kahalya,
my mother, Smt. Sheela Kahalya, and
my sister, Geetanshi Maheshwari

Contents

List of Figures

List of Tables

Acknowledgments

The past six years as a PhD student at Berkeley have been a truly transformative experience, both personally and professionally.

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Shankar Sastry, for taking me under his wing and providing invaluable guidance over the years. His profound wisdom—often delivered in memorable one-liners—has inspired me to think deeply and creatively, and has shaped both my personal and professional growth. After every meeting, I left with renewed clarity and motivation to tackle the challenges ahead. Beyond his academic insight, Professor Sastry has made the research journey truly enjoyable. His fascinating stories about every corner of Berkeley have added color and context to my time here, making each conversation both enlightening and entertaining. More than a mentor, Professor Sastry has been a true role model to me. As I look ahead, I'll always aspire to emulate his intellectual depth, leadership, and his infectious enthusiasm and passion for research. Thank you, Shankar for making this journey so meaningful.

I am also grateful to my dissertation committee members—Professor Shankar Sastry, Professor Claire Tomlin, Professor Chris Shannon, and Professor Manxi Wu—for serving on my committee and for their invaluable feedback in helping bring this dissertation to life.

First, I would like to thank Professor Claire Tomlin. Beyond serving on my committee, she has been a constant source of encouragement and mentorship during my time at Berkeley. She has always been welcoming and generous with her time, whether in discussing research ideas or offering feedback. Her insightful questions and thoughtful comments have pushed me to refine my thinking and communicate my work more clearly. I am especially grateful for the detailed suggestions she provided on my presentations, which greatly improved their clarity and impact.

I would also like to thank Professor Chris Shannon for teaching one of my favorite courses at Berkeley. I have deeply appreciated our research discussions after class or in office hours, which broadened my perspective and deepened my understanding of several key ideas in economics.

Finally, I would like to thank Professor Manxi Wu. Many of the chapters in this dissertation stem from our collaborations over the past four years. I have greatly valued our work together and the many afternoons we spent brainstorming ideas and working through proofs at the whiteboard. I have learned immensely from her thoughtful approach to research and problem-solving. She has been incredibly generous with her time and provided detailed feedback on both my paper drafts and presentations, for which I am sincerely thankful.

In past six years, I have been fortunate to have many other mentors along the way. First, I would like to thank Professor Lillian Ratliff. I began collaborating with her on my first research project at Berkeley and learned a great deal from her about how to think deeply about research problems. She has always been supportive in discussing several interesting ideas in past years. I am very grateful to her for her valuable feedback on my talks. I am also thankful to other faculty members, Professor Anil Aswani, Professor Eric Mazumdar,

Professor Alex Bayen, and Professor Hamsa Balakrishnan, who have been very generous with their time and provided invaluable feedback during my PhD journey.

I have been extremely fortunate to collaborate with and learn from several outstanding researchers. The work in this thesis would not be what it is without them. I would like to thank Professor Shankar Sastry, Professor Manxi Wu, Professor Lillian Ratliff, Professor Eric Mazumdar, Professor Anil Aswani, Professor Hamsa Balakrishnan, Professor Sanjit Seshia, Professor Xin Guo, Professor Pramod Khargonekar, Professor Deepan Muthirayan, Dvij Kalaria, Victoria Tuck, Maria Mendoza, Addison Kalanther, Daniel Bostwick, Pan-Yang Su, Chih-Yuan Chiu, Druv Pai, Xinyu Li, Sukanya Kudva, Victor Qin, James Cheng, Jiarui Yang, and Kshitij Kulkarni for the engaging discussions, fruitful collaborations, and technical insights that enriched my understanding.

Outside of research, I have had the privilege of being a part of the teaching staff for several courses. I want to thank Professor Shankar Sastry, Professor Venkat Anantharam, Professor Gireeja Ranade, Professor Manxi Wu, and Professor Anil Aswani for selecting me to be part of their teaching teams. I learned tremendously about course design—from developing lecture material, homework, discussion sets, and exams to handling administrative responsibilities. Most importantly, I learned from these wonderful faculty members how to create a safe and inspiring learning environment that encourages students to think independently and creatively. I also want to thank my fellow semiautonomous co-organizers, Chris, Frank, Pan-Yang, Sampada, and Victoria, for helping organize one of the most intellectually stimulating seminar series I have been a part of.

I would like to thank the many friends I've met, both within and outside of research, for the countless fun moments and memorable experiences we've shared.

I also want to thank the administrative staff of the EECS department at UC Berkeley for helping me smoothly navigate the logistical aspects of my graduate journey.

Before coming to UC Berkeley, I had the great fortune of attending IIT Bombay for my undergraduate studies. I am deeply thankful to my undergraduate mentors, Professor Debasish Chatterjee and Professor Sukumar Srikant, for introducing me to the fascinating world of control theory. Their guidance and mentorship were instrumental in helping me appreciate the importance of mathematical rigor in the research I pursue today. I remain indebted to them for their support and encouragement.

Finally, I am deeply grateful to my family for being a constant pillar of support. I would like to thank them for their blessings, support, and encouragement throughout this journey. No words can do justice in expressing my gratitude to my father, Sh. Naresh Kahalya; my mother, Smt. Sheela Kahalya; and my sister, Geetanshi Maheshwari, for their unconditional love, values, and belief in me. This dissertation is dedicated to them.

Last but not least, I thank God for everything.

Chapter 1

Overview

Autonomous AI technologies are fundamentally reshaping critical societal systems—including robotics, mobility, logistics, and energy—by enabling intelligent, data-driven decision-making at unprecedented scale. While significant progress has been made in developing autonomous agents that operate effectively in isolation or structured environments, a central challenge remains: integrating these agents into complex societal systems. In such settings, agents must make real-time decisions while interacting strategically with other autonomous systems and humans, all within uncertain, dynamic, and resource-constrained environments. Crucially, these agents often operate without access to others' private information, and must continually adapt to the evolving behavior of both autonomous agents and humans.

Such interactions raise a range of theoretical, algorithmic, computational, and societal challenges. These include developing new frameworks for modeling and analyzing strategic behavior in societal systems. Furthermore, we need efficient algorithms that support real-time learning and decision-making in multi-agent environments. Additionally, without careful societal design, interactions among autonomous agents can lead to undesirable societal outcomes—including inefficiencies in shared resource usage, inequities in access or outcomes, or systemic risks to safety and reliability. Addressing these challenges is essential to ensure that the next generation of autonomous systems are both efficient and socially responsible.

This dissertation addresses these challenges by focusing on two fundamental questions:

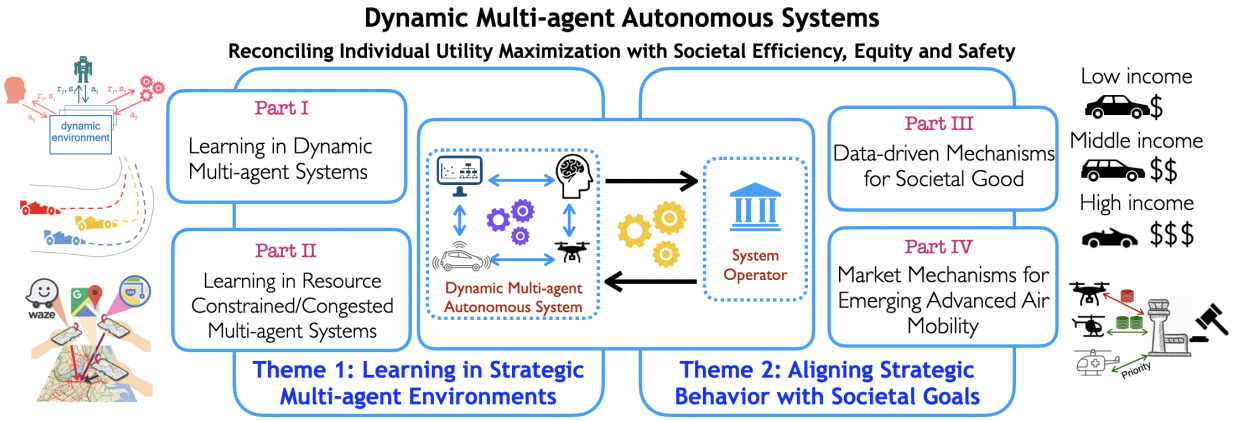
Q1 How can we design learning and decision-making algorithms for strategic autonomous agents operating in societal settings?

Q2 How can we ensure that their interactions align with broader societal goals such as efficiency, equity, and safety?

To answer these questions, this dissertation introduces new theoretical and algorithmic frameworks and demonstrates their effectiveness through a variety of application-driven case studies.

The dissertation is organized into four parts. **Parts I** and **II** address **Q1**, focusing on the incorporating the strategic nature of multi-agent interaction in design and analysis of learning and decision-making algorithms. **Parts III** and **IV** address **Q2**, developing adaptive incentive mechanisms and market-based coordination tools that guide agent behavior toward socially beneficial outcomes. These contributions are illustrated through a range of applications, including high-performance multi-robot system, efficient and equitable urban mobility (both in air and road networks), and the design of online services.

The following sections provide a detailed overview of each part, highlighting the theoretical insights and practical implications of the proposed approaches.



1.1 Part I: Multi-agent Learning in Dynamic Environments

Modern autonomous systems increasingly operate in environments shared with other autonomous agents. For instance, consider autonomous robo-taxis navigating busy urban areas, autonomous drones or racing cars competing in high-speed events, or online service providers leveraging data to facilitate new online services. In such settings, agents must not only adapt to uncertain dynamics, but also anticipate and respond to the strategic behavior of others. This interdependence fundamentally alters the nature of decision-making: outcomes depend not only on an agent's own actions but also on the evolving strategies of other agents. As a result, classical single-agent learning and control methods fall short, motivating the need for a rigorous understanding of multi-agent learning in dynamic, uncertain environments.

This part (Chapters 2–5) develops new theoretical and algorithmic frameworks for strategic decision-making and learning in multi-agent autonomous systems operating in uncertain and dynamic environments. The analysis is grounded in the framework of *Markov games*, which model interactions among strategic agents over long time horizons under stochastic dynamics. A central objective is to understand how independently learning agents can operate in a rational manner in multi-agent environment. This objective is often characterized by ensuring converge to some approximate *Nash equilibria*—a popular solution concept in game

theory that characterizes stable outcomes where no agent has an incentive to unilaterally deviate. Unlike classical formulations that assume either fully cooperative (team-optimal) or fully adversarial (zero-sum) objectives, many real-world multi-agent systems involve mixed cooperative and competitive (aka "general-sum") interactions. This part addresses the algorithmic and theoretical challenges inherent in such settings.

Chapter 2 introduces Markov α-potential games, a novel theoretical framework for design and analysis of multi-agent learning algorithms in general-sum Markov games. The core idea is to approximate the long-run utility difference resulting from a unilateral policy deviation by a scalar function, termed the α-potential function. We show that such an approximation exists for any general-sum Markov game, and that optimizing the α-potential yields an approximate Nash equilibrium, where α quantifies the deviation from an exact equilibrium. This framework encompasses several structured game classes—including Markov congestion games and perturbed team games—that are not captured by dominant frameworks such as team games or zero-sum games. The chapter presents a computational approach based on semi-infinite linear programming to estimate the α parameter for a given game. Additionally, it introduces learning dynamics such as projected gradient ascent and sequential best-response updates, and provides both theoretical and empirical evidence of their convergence to approximate equilibria, even without explicit knowledge of the α-potential function. This chapter is based on the reference [171].

Building on this foundation, Chapter 3 applies the Markov α-potential games framework to the domain of autonomous racing, a benchmark for real-time, strategic decision making in competitive environments. The chapter formalizes multi-agent racing as a general-sum Markov game and introduces a two-phase solution approach: (1) an offline phase that approximates the α-potential function from simulated game data, and (2) an online phase that performs real-time planning by maximizing this learned potential. Empirical results in three-car racing demonstrate the efficacy of this approach by beating several existing baselines. Moreover, winning strategies generate strategic maneuvers like overtaking and blocking. This chapter is based on the reference [277].

Chapters 4–5 focus on decentralized learning in multi-agent environments, where agents must act based on local observations without **any** access to information about other agents. We study decentralized actor-critic style algorithms. In particular, the algorithm has three key features: (1) each agent updates its long-term utility estimates using Temporal Difference (TD) learning, and its policy updates using inertial one-stage best response; (2) each agent updates its policy slower than its long-term utility estimates, resulting in a timescale-separated dynamical system; and (3) each agent relies solely on bandit feedback for the most recent state-action pair, leading to asynchronous updates of different components of the policy and utility estimates.

Leveraging advances in two-timescale stochastic approximation theory, we decouple the convergence analysis of the critic and policy updates. The analysis reduces to showing that the critic updates converge to the true value functions for fixed policies, followed by analyzing the convergence of policy updates assuming the critic has already converged to the corresponding true value function. The convergence of the critic updates is established

using the contraction property of the Temporal Difference (TD) operator.

In Chapter 4, we establish the convergence of policy updates under the structural assumption that the underlying game is a Markov potential game—a special case of Markov α-potential games with $\alpha = 0$. In particular, we show that the potential function serves as a Lyapunov function for the policy updates. This chapter is based on the reference [275].

Chapter 5 extends this analysis to general-sum Markov games. Central to our analysis is the framework of Markov near-potential functions (MNPFs), a generalization of Markov α-potential functions introduced in Chapter 2. We show that for any general-sum Markov game, the MNPF acts as an approximate Lyapunov function for the policy updates—one that increases unless the policies converge to an (approximate) Nash equilibrium set—thus enabling us to characterize the set of convergent policies. This chapter is based on the reference [276].

Taken together, these chapters highlight the critical role of tools from game theory, machine learning, dynamical systems, and control theory in developing the theoretical and algorithmic foundations for multi-agent learning in dynamic environments.

1.2 Part II: Multi-agent Learning in Resource-Constrained/Congested Environments

In many existing and emerging AI-driven societal systems, agents operate in environments with limited resources and shared infrastructure, which add important challenges to strategic decision-making. Resource constraints naturally create competition, while congestion effects lead to costs that depend on the collective actions of many agents. These factors require agents to learn and adapt to the competition and congestion caused by other agents in the system.

This part, comprising Chapters 6–8, investigates multi-agent learning specifically in two key domains characterized by such constraints: two-sided matching markets and transportation networks. In matching markets, agents learn to compete for scarce resources (e.g., job openings, college placements), while in transportation systems, agents interact over shared, congestible network resources where travel costs depend on aggregate usage. Across both domains, we develop learning algorithms with provable performance guarantees—such as regret bounds and convergence to equilibrium—that enable agents to effectively navigate competition and congestion in uncertain environments.

Chapter 6 focuses on online learning in two-sided matching markets, which model multi-agent interactions in large-scale marketplaces such as Amazon Mechanical Turk. These markets involve two types of participants: *agents* and *firms*. While we consider that firms have fixed and known preferences over agents, agents must learn their preferences over firms through repeated interactions, all while competing with other agents for successful matches. The central challenge is to design decentralized algorithms that enable agents to learn their preferences while simultaneously competing for limited resources, since each firm can match

with only one agent at a time. We propose a class of decentralized, communication- and coordination-free algorithms in which agents make decisions solely based on their individual histories of interaction, without any knowledge of firms' or other agents' preferences. Our approach decouples the statistical learning of preferences (from noisy observations) from the strategic competition for firms. Under mild structural assumptions on the preferences of agents and firms, we show that agents incur at most logarithmic regret over the time horizon. This chapter is based on the reference [274].

Chapter 7 addresses multi-agent learning in matching markets with time-varying preferences, where agents' preferences are unknown and may change over time. Unlike Chapter 6, which assumes fixed but unknown preferences, this chapter considers a setting in which one side of the market (firms) has known preferences, while the other side (agents) must adapt to unknown, *time-varying* preferences. We propose a centralized algorithm that enables agents to learn and track their evolving preferences over time. We prove that agents achieve uniform sub-linear regret that scales with the number of preference changes. Remarkably, this matches the best-known regret bounds in the single-agent setting, up to constant factors, despite the added complexity of strategic competition. This chapter is based on the reference [309].

Chapter 8 studies routing in congested environments, focusing on arc-based traffic assignment models (TAMs) in transportation community. In this framework, travelers make sequential routing decisions based on observed congestion. We develop a *Condensed DAG* (CoDAG) representation of the network graph that enables agents to learn to route on the network in presence of the evolving congestion patterns due to actions of other agents. We define and analyze the *Condensed DAG equilibrium*—a unique equilibrium flow computable via a strictly convex optimization program. Additionally, we propose a natural learning dynamics that allows agents to select arcs in the transportation network based on past observed congestion. This dynamics is proven to converge to a neighborhood of the Condensed DAG equilibrium. To our knowledge, this is the first framework that jointly analyzes learning and equilibrium behavior in arc-based TAMs. This chapter is based on the reference [278].

Collectively, these chapters advance the theoretical understanding of learning and adaptation in multi-agent systems with resource constraints, providing new algorithmic and analytical tools with potential applicability beyond matching markets and transportation systems.

1.3 Part III: Data-driven Mechanisms for Societal Good

Interactions among self-interested agents with misaligned objectives often lead to undesirable societal outcomes. In transportation systems, for instance, individual route optimization leads to congestion and increased emissions. In automated decision-making systems—such as loan approvals—users may manipulate input features if they understand the model's structure, compromising both accuracy and fairness. These examples underscore the need

for a system designer to implement mechanisms that align individual incentives with social objectives. The widespread availability of behavioral and system-level data has enabled the use of data-driven approaches to design such mechanisms. However, doing so requires addressing two fundamental challenges that lie beyond the scope of traditional data science and machine learning paradigms. First, agents are strategic and may adapt their behavior in response to deployed mechanisms, resulting in distribution shifts and model misspecification. Second, system designers often have limited or no access to agent-specific private information due to privacy constraints.

This part of the dissertation, comprising Chapters 9–12, investigates the design of data-driven mechanisms that address these challenges. The first two chapters study scenario when the agents know their preferences while the latter two allow online adaptation of strategies of agents as they learn and the mechanism steers such adaptive agents toward socially desirable outcomes under limited knowledge of private information of agents.

Chapter 9 studies the design of congestion pricing mechanisms using behavioral and socioeconomic data from modern transportation networks. Travelers aim to minimize individual travel costs, often resulting in system-level inefficiencies such as congestion and pollution. Congestion pricing is a standard intervention, but its regressive impact on low-income users raises equity concerns. This chapter proposes a new class of pricing mechanisms that improve both efficiency and equity by minimizing total congestion while reducing cost disparities across income groups. Using data from the San Francisco Bay Area Freeway Network, we develop an equilibrium model that captures heterogeneous traveler sensitivities to tolls. Machine learning and optimization techniques are employed to estimate behavioral parameters across income segments. Our analysis demonstrates that failure to account for demographic and geographic heterogeneity can exacerbate congestion and socioeconomic disparities. This chapter is based on the reference [269].

Chapter 10 addresses the design of robust machine learning classifiers under the distribution shifts induced to the strategic response of users. In settings such as credit scoring or spam detection, users may manipulate features in response to deployed classifiers, undermining predictive performance. Additionally, the data-generating process needs to be robust against adversarial manipulation or unmodeled user behavior. We model this as a distributionally robust strategic classification problem. These problems are generally intractable and hard due to the lack of access to agents' private utility functions. To overcome these challenges, we reformulate the problem as a finite-dimensional convex-concave min-max optimization and introduce a gradient-free learning algorithm that computes a robust classifier using only observed user responses. This approach enables robustness to both strategic behavior and distribution shifts without requiring knowledge of agents' private preferences. Additionally, we also derive finite time convergence of proposed algorithm to the solution of min-max optimization problem. This chapter is based on the reference [268].

The focus then shifts to online, adaptive incentive mechanisms where the agents themselves are learning and adapting their strategies with time. In Chapter 11, we develop an efficient algorithm for computing the incentive mechanism when the agents are updating their strategies over time (a la **Part I-II**) and the operator has no knowledge of the agents'

utility functions, strategy spaces, or learning algorithms. By designing a novel gradient estimator based solely on observed actions of agents, our algorithm allows the operator to adapt incentives in real-time. We establish convergence guarantees to a stationary point of societal objective and demonstrate the effectiveness of this approach in large-scale transportation network problems, where it enables incentive design without requiring access to user demand information. This chapter is based on the reference [271].

A key shortcoming of gradient-based method presented in Chapter 11 is that sometimes the critical points of the resulting dynamics may not be desirable in terms of social cost function. We overcome this shortcoming by proposing a novel adaptive incentive mechanism in Chapter 12. The mechanism updates incentives based on each player's externality – the difference between their marginal cost and the system operator's marginal cost–on a slower timescale relative to the agents' learning dynamics. This two-timescale approach is agnostic to the specific learning dynamics of the agents and ensures that any fixed point of the mechanism corresponds to a socially optimal Nash equilibrium. We provide sufficient conditions for asymptotic convergence of the mechanism and validate the mechanism in both atomic and non-atomic game settings, with applications to aggregative and routing games. This chapter is based on the reference [272].

Together, these chapters present new methods to design and analyze data-driven incentive mechanisms, demonstrating how both offline and online data can be harnessed to align strategic agents' behavior with societal objectives such as efficiency, equity, and robustness.

1.4 Part IV: Market Mechanisms for Emerging Advanced Air Mobility

As autonomous technologies increasingly permeate the physical world, they are poised to enable a wide range of new services. There is a tremendous opportunity to ensure that multi-agent interactions within these systems are socially responsible from the outset. One such emerging service is Advanced Air Mobility (AAM), which encompasses the use of unmanned aerial vehicles (UAVs), air taxis, and novel cargo and passenger transport solutions. By leveraging previously underutilized airspace, AAM has the potential to transform urban transportation. Recent studies project that the air mobility market could exceed USD 50 billion by 2035, highlighting its substantial growth potential.

Despite the widespread optimism surrounding AAM, the design of regulatory and operational frameworks remains an open challenge. While concepts from traditional air traffic management can provide a foundation, they often fall short in addressing the dynamic, decentralized, and adaptive nature of AAM operations. In recognition of this, the Federal Aviation Administration (FAA) is actively developing a clean-slate congestion management framework aimed at ensuring efficiency, fairness, and safety. Market-based congestion management mechanisms have been proposed as promising tools for coordinating AAM operations. However, despite their conceptual appeal, practical progress in developing and de-

ploying such mechanisms has been limited—underscoring the need for new, interdisciplinary approaches that integrate tools from engineering, economics, and algorithmic design.

This part (Chapters 13–14) proposes two market-based mechanisms for allocating airspace infrastructure resources, with the goal of ensuring socially beneficial outcome in the presence of strategic agents and real-time operational considerations. We model the airspace as a time-extended network of contiguous regions, each subject to capacity constraints on entry, transit, and exit. Vehicle requests, expressed as desired time-trajectories, are represented as paths in this time-extended graph. Leveraging the structure of this graph-based model, we develop two distinct market mechanisms to allocate conflict-free paths to AAM vehicles.

In Chapter 13, we address the problem of designing a market mechanism for coordinating the movement of AAM vehicles operated by multiple flight operators. Each operator manages a heterogeneous fleet and holds private valuations over airspace resources. The model incorporates operational constraints on arrival, departure, and parking at vertiports. We propose a centralized mechanism that elicits bids from operators representing their private preferences over the desired airspace resources. Based on these bids, the system designer allocates resources and determines payments. The proposed mechanism guarantees four key properties. First, it ensures efficiency by maximizing the total reported value of flight operators over the allocated resources. Second, it ensures safety by distributing congestion and guaranteeing that all allocations respect the relevant airspace and vertiport capacity constraints. Third, it ensures that truthful reporting of preferences is in each operator's best interest. Fourth, to improve computational scalability, we develop a mixed-integer linear programming formulation that exploits the underlying network flow structure of the problem. This chapter is based on the reference [397].

In Chapter 14, we introduce a novel market mechanism for allocating airspace resources in scenarios where the private valuations of operators are not accessible to the central system designer. To address this challenge while preserving operator privacy, we assign each vehicle a fixed budget of artificial currency, referred to as air-credits, and anonymously post prices for traversing edges of the time-extended network. The goal is to compute a competitive equilibrium such that all capacity constraints are satisfied, each vehicle receives an allocation that is optimal given the posted prices and its budget, and no payment is charged for underutilized resources. Since such competitive equilibria with integral allocations may not always exist, we establish sufficient conditions under which a fractional competitive equilibrium exists and can be computed efficiently. Building on this theoretical foundation, we propose a distributed two-step algorithm: the first step computes a fractional competitive equilibrium, and the second derives an integral, feasible allocation from this fractional solution.

The mechanism developed in Chapter 14 offers several advantages. It encourages fair allocation by allowing the system designer to regulate operator budgets to promote equitable access to airspace resources. It ensures safe allocation by satisfying all capacity constraints across the network. It is privacy-preserving and distributed in nature, relying only on limited information from the operators, and does not require access to their private utility functions. Furthermore, it supports flexible and adaptive routing by encouraging operators to modify their travel plans—such as departure times or routes—in response to network conditions

and posted prices, thereby facilitating congestion mitigation. The efficacy of this approach is demonstrated using real-world drone delivery data from Airbus operations in Toulouse. This chapter is based on the reference [273].

Part I

Multi-agent Learning in Dynamic Environments

Chapter 2

Markov α-Potential Games: A New Framework for Multi-agent Reinforcement Learning

Designing non-cooperative multi-agent systems interacting within a shared dynamic environment is a central challenge in many existing and emerging autonomy applications, including autonomous driving, smart grid management, and e-commerce. Markov game, proposed in [382], provide a mathematical framework for studying such interactions [447]. A primary objective in these systems is for agents to reach a *Nash equilibrium*, where no agent benefits from changing its strategy unilaterally. However, designing algorithms for approximating or computing Nash equilibrium are generally intractable [329], unless certain structure of underlying multi-agent interactions are exploited. There is a rich line of literature on equilibrium computation and approximation algorithms for Nash equilibrium in Markov zero-sum games (see [367] and references therein), Markov team games (see [36] and references therein), symmetric Markov games (see [440]), and in particular, Markov potential games (see [275, 450, 238, 313] and references therein) and its generalization to weakly acyclic games (see [17, 439] and references therein).

In this chapter, we propose the *Markov α-potential game* framework, where changes in an agent's long-run utility from unilateral policy deviations are captured by an "*α-potential function*" and a parameter α (Definition 2.2.3). We establish that **any** finite-state, finite-action Markov game is a Markov α-potential game for some $\alpha \geqslant 0$, and there exists an α-potential function (Theorem 2.2.1). Furthermore, we show that any optimizer of an α-potential function, if it exists, is an α-stationary Nash equilibrium (Proposition 2.2.1).

Markov α-potential games generalize the framework of Markov potential games (MPGs). MPGs, originally proposed in [410] and [238], correspond to the special case of $\alpha = 0$ and extend a rich body of literature on static potential games (or static congestion games) [306]. The MPG structure has enabled learning algorithms with convergence guarantees to Nash equilibrium (e.g., [275, 121]). However, two main challenges remain: (1) the lack of real-world examples that can be provably shown to be MPGs, and (2) the difficulty of certifying games as

MPGs and constructing potential functions, except in special cases (e.g., state-independent transitions or identical payoffs [313, 238]). Our α-potential game framework addresses both challenges: it shows that any finite-state, finite-action Markov game is a Markov α-potential game and provides a semi-infinite linear programming approach to certify MPGs (Section 2.4).

Our Markov α-potential games framework extends the static near-potential games, proposed in [81], to Markov games. Unlike static games, where the nearest potential function always exists, the existence of an α-potential function requires additional analysis (Theorem 2.2.1). Moreover, while finding the nearest static potential function involves finite-dimensional linear programming, computing the α and its potential function requires solving a semi-infinite linear programming problem, as the α-potential function spans both state and policy spaces, the latter being uncountable.

We derive explicit upper bounds on the parameter α for two classes of relevant games. First, we consider *Markov Congestion Games (MCGs)*, where each stage game is a congestion game (proposed in [351]) and the state transition depends on agents' aggregate resource utilization. This is equivalent to Markov games where each stage is a static potential game, as static congestion games and static potential games are equivalent [306]. This class models applications like dynamic routing, communication networks, and robotic interactions [111, 398, 210]. We show that the upper bound on α for MCGs scales linearly with the state and resource set sizes, and inversely with the number of agents (Proposition 2.3.2). Second, we consider *Perturbed Markov Team Games (PMTGs)*, which generalize Markov team games by allowing deviations of individual utility from the team objective. We provide an upper bound for PMTGs that scales with the magnitude of these deviations (Proposition 2.3.3). For both MCGs and PMTGs, we calculate an upper bound on α by using a *specific* candidate α-potential function to compute an analytical upper bound on α. However, this upper bound can be loose. In such cases, the semi-infinite linear programming method described in Section 2.4 can be used to obtain tighter numerical estimates of α.

We propose two algorithms to approximate stationary Nash equilibrium in Markov α-potential games that are based only on the utility function of agents and not on the knowledge of $\alpha-$potential function. We study the Nash-regret of both algorithms and characterize its dependence on α (Theorems 2.5.1 and 2.5.2). First, we analyze the *projected gradient-ascent algorithm* (Algorithm 1), originally proposed in [121] for MPGs, in the context of Markov α-potential games by bounding the path length of policy updates using changes in the α-potential function and α. Following our proof technique, the analysis of many existing algorithms for MPGs can be extended similarly to Markov α-potential games. Second, we propose a *new algorithm* called the *sequential maximum improvement algorithm* (Algorithm 2) and derive its Nash-regret. The main technical novelty in the analysis is to bound the maximum improvement of a "smoothed" Q-functions with respect to change in policies (aka "path length of policies"), which in turn is bounded by cumulative change in α-potential function (Lemma 2.5.5). For $\alpha = 0$, this algorithm and its analysis are independently relevant to MPGs. We numerically validate these algorithms on examples of MCGs and PMTGs.

Additional Related Works

This chapter on Markov α-potential games is related to the literature on weakly acyclic Markov games, proposed in [17]. Weakly acyclic Markov games extend weakly acyclic static games to Markov games, encompassing MPGs as a special case. Unlike MPGs, weakly acyclic Markov games do not require the existence of an exact potential function, instead retain many key properties of potential games, such as the existence of pure equilibria and finite strict best-response paths. Just as MPGs, most games are not weakly acyclic, and determining whether a game is weakly acyclic remains an open problem. On one hand, the introduction of a Markov α-potential games allows for design and analysis of algorithms as a game diverges from a MPG. On the other hand, if a game is weakly acyclic, it is an α-potential game with the value of α not necessarily zero. Exploring the connection between these two approaches and how they might be used together to analyze general Markov games is an interesting and open direction for future research.

Our Algorithm 1 for Markov α-potential games is connected with a substantial body of work on learning approximate Nash equilibria (NEs) in MPGs (see [275, 121, 284, 391, 144, 400, 449]). The first global convergence result for the policy gradient method in MPGs was established in [238]. Additionally, these algorithms have been studied in both discounted infinite horizon settings [121, 144] and finite horizon episodic settings [284, 391]. Other methods, such as natural policy gradient [144, 400, 449] and best-response based methods [275], have also been explored.

Our Algorithm 2 is reminiscent of the "Nash-CA" algorithm developed for MPGs in [391], which requires each player to sequentially compute the best response policy using an RL algorithm in each iteration; in contrast, our algorithm only computes a smoothed *one-step* optimal deviation. One-step optimal deviation based algorithms has also been studied for MPGs [275] [83]. Additionally, incorporating smoothness for better performance in Markov games is also studied in [86, 133, 297].

Finally, a recent work [111] introduces an approximation algorithm for MCGs and investigates the Nash-regret. Their results and approach are tailored exclusively for congestion games, whereas our work focuses on a broader framework of Markov α-potential games.

Notations

For any $n \in \mathbb{N}$, $[n] := \{1, 2, 3, ..., n\}$. For a finite set X, $\Delta(X)$ denotes the set of probability distributions over X. For any function $f : X \to \mathbb{R}$, the L_∞-norm is defined by $\|f\|_\infty = \max_{x \in X} |f(x)|$, the L_1-norm is $\|f\|_1 = \sum_{x \in X} |f(x)|$, and the L_2-norm is $\|f\| = \sqrt{\sum_{x \in X} |f(x)|^2}$.

Organization

The rest of the chapter is organized as follows. Section 2.1 introduces the setup of Markov games; Section 2.2 introduces the framework of Markov α-potential game and establishes the

existence of associated α-potential function, with examples of Markov α-potential games analyzed in Section 2.3; an optimization framework in the form of semi-infinite linear programming for finding the upper bound of α is in Section 2.4; Section 2.5 presents two algorithms, projected gradient-ascent and sequential maximum improvement, and their Nash-regret analysis; and numerical examples are in Section 2.6.

The proofs of technical lemmas and propositions, unless otherwise specified, are deferred to Appendix A.

2.1 Preliminaries on Markov Games

Let us first recall the mathematical setup of Markov games. An N-player Markov game $\mathcal{G}$ is characterized by $\langle I, S, (A_i)_{i\in I}, (u_i)_{i\in I}, P, \gamma\rangle$, where

1. $I = (1, 2, \cdots N)$ is the finite set of $N \in \mathbb{N}$ players,
2. S is the finite set of states,
3. A_i is the finite set of actions of player $i \in I$ and $A := \times_{i\in I} A_i$ is the set of joint actions of all players,
4. $u_i : S \times A \to \mathbb{R}$ is the one-stage payoff function of player $i \in I$,
5. $P = (P(s'|s, a))_{s,s'\in S, a\in A}$ is the probability transition kernel such that the probability of transitioning to state $s' \in S$ given the current state $s \in S$ and action profile $a \in A$ is given by $P(s'|s, a)$, and
6. $\gamma \in [0, 1)$ is the discount factor.

The game proceeds in discrete time steps. At each time step $k = 0, 1, 2, \cdots$, the state of the game is $s^k \in S$, the action taken by player $i \in I$ is $a_i^k \in A_i$, and the joint action of all players is $a^k = (a_i^k)_{i\in I} \in A$. Once players select their actions, each player $i \in I$ observes her one-stage payoff $u_i(s^k, a^k) \in \mathbb{R}$, and the system transits to state s^{k+1}, where $s^{k+1} \sim P(\cdot|s^k, a^k)$.

In this study, we assume that the action taken by any player is based on a randomized stationary Markov policy, as in the Markov games literature [149, 117, 238, 121, 440]. That is, for any player $i \in I$, the action selected at time step k is $a_i^k \sim \pi_i(\cdot|s^k)$, and the joint policy of all players is $\pi = (\pi_i)_{i\in I} \in \Pi := \times_{i\in I}\Pi_i$, with $\Pi_i := \{\pi_i : S \to \Delta(A_i)\}$. The joint policy of all players except player i is denoted as $\pi_{-i} = (\pi_j)_{j\in I\setminus\{i\}} \in \Pi_{-i} := \times_{j\in I\setminus\{i\}}\Pi_j$. Given $\pi \in \Pi$, the probability of the system transiting from s to s' is denoted as $P^\pi(s'|s) := \mathbb{E}_{a\sim\pi}[P(s'|s, a)]$. The accumulated reward (a.k.a. the *utility function*) for player i, given the initial state $s \in S$ and the joint policy $\pi \in \Pi$, is

$$V_i(s, \pi) := \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k u_i\left(s^k, a^k\right) \mid s^0 = s\right], \tag{2.1}$$

where $\gamma \in [0,1)$ is the discount factor, $a^k \sim \pi\left(s^k\right)$, and $s^{k+1} \sim P\left(\cdot|s^k, a^k\right)$. Denote also $V_i(\mu,\pi) := \mathbb{E}_{s\sim\mu}[V_i(s,\pi)]$, if the initial state follows a distribution $\mu \in \Delta(S)$. Additionally, define the discounted state visitation distribution as

$$d_\mu^\pi(s) := (1-\gamma)\sum_{k=0}^{\infty}\gamma^k P(s^k = s|s^0 \sim \mu). \tag{2.2}$$

To analyze this game, we will adopt the solution concept of ϵ-stationary Nash equilibrium (NE).

Definition 2.1.1. *(ϵ-stationary Nash equilibrium). For any $\epsilon \geqslant 0$, a policy profile $\pi^* = (\pi_i^*, \pi_{-i}^*)$ is an* ϵ-stationary Nash equilibrium *of the Markov game $\mathcal{G}$ if for any $i \in I$, any $\pi_i \in \Pi_i$, and any $\mu \in \Delta(S)$,*

$$V_i(\mu, \pi_i^*, \pi_{-i}^*) \geqslant V_i(\mu, \pi_i, \pi_{-i}^*) - \epsilon.$$

When $\epsilon = 0$, it is simply called a *stationary NE*. A stationary NE always exists in any Markov game with finite states and actions [149].

2.2 Markov α-potential games

In this section, we introduce the framework of Markov α-potential games. We show that any Markov game can be analyzed under this framework. First, we introduce some preliminaries. We define a metric $\mathbf{d}$ on Π as follows: for any $\pi, \tilde{\pi} \in \Pi$,

$$\begin{aligned}\mathbf{d}_i\left(\pi_i, \tilde{\pi}_i\right) &:= \max_{s\in S, a_i \in A_i} \left|\pi_i\left(a_i \mid s\right) - \tilde{\pi}_i\left(a_i \mid s\right)\right|, \quad \forall i \in I,\\ \mathbf{d}(\pi, \tilde{\pi}) &:= \max_{i\in I} \mathbf{d}_i\left(\pi_i, \tilde{\pi}_i\right).\end{aligned} \tag{2.3}$$

Evidently, the sets of policies $\{\Pi_i\}_{i\in I}$ are compact in the topology induced by the metrics $\{\mathbf{d}_i\}_{i\in I}$, Π is compact in the topology induced by $\mathbf{d}$, and the utility functions are continuous with respect to π under the metric $\mathbf{d}$ [440]. Next, we introduce the notion of maximum pairwise distance between a Markov game and a real-valued function defined on $S \times \Pi$.

Definition 2.2.1. *(Maximum pairwise distance). Given any Markov game $\mathcal{G}$ and a function $\Psi : S \times \Pi \to \mathbb{R}$, the* maximum pairwise distance $\widehat{\mathbf{d}}$ *between Ψ and $\mathcal{G}$ is defined as*

$$\begin{aligned}&\widehat{\mathbf{d}}(\Psi, \mathcal{G})\\ &:= \sup_{\substack{s\in S, i\in I,\\ \pi_i, \pi_i' \in \Pi_i,\\ \pi_{-i}\in\Pi_{-i}}} \left|\Psi\left(s, \pi_i', \pi_{-i}\right) - \Psi\left(s, \pi_i, \pi_{-i}\right) - \left(V_i\left(s, \pi_i', \pi_{-i}\right) - V_i\left(s, \pi_i, \pi_{-i}\right)\right)\right|.\end{aligned}$$

Definition 2.2.1 generalizes the concept of maximum pairwise distance from [81, Definition 2.3], extending it from static games (action profiles) to Markov games, where the distance is measured over policies that map states to action distributions. Next, we introduce the notion of a game elasticity parameter, which is useful for defining Markov α-potential games. Intuitively, this parameter captures the smallest value of the maximum pairwise distance between any function in a set $\mathcal{F}_{\mathcal{G}}$ (to be defined shortly) and $\mathcal{G}$.

Definition 2.2.2. *(Game elasticity parameter). Given any game $\mathcal{G}$, its* game elasticity parameter α *is defined as*

$$\alpha := \inf_{\Psi \in \mathcal{F}^{\mathcal{G}}} \widehat{\mathbf{d}}(\Psi, \mathcal{G}), \tag{2.4}$$

where

$$\mathcal{F}^{\mathcal{G}} := \{\Psi : S \times \Pi \to \mathbb{R} \text{ } s.t. \text{ } \|\Psi\|_\infty \leqslant (2N)/(1-\gamma) \max_{i \in I} \|u_i\|_\infty\}$$

is a class of bounded uniformly equi-continuous function on Π [1]

Our choice of the specific value of the upper bound on functions in $\mathcal{F}^{\mathcal{G}}$ is useful for the proof of Proposition 4.1.

Clearly $\alpha < \infty$ as one can take $\Psi = 0$ in (2.4) to ensure $\alpha \leqslant 2\|V_i\|_\infty < \infty$.

Furthermore, the game elasticity parameter depends on variety of game parameters, including the number of players, the action and state sets, the utility function values, the Markov state transition dynamics, and the discount factor.

Next, we define Markov α-potential games.

Definition 2.2.3. *(Markov α-potential game). A Markov game $\mathcal{G}$ is a* Markov α-potential game *if α is the game elasticity parameter. Furthermore, any $\Phi \in \mathcal{F}^{\mathcal{G}}$ such that $\widehat{\mathbf{d}}(\Phi, \mathcal{G}) = \alpha$ is called an* α-potential function *of $\mathcal{G}$.*

Next, we present a useful property due to Definition 2.2.3.

Corollary 2.2.1. *Let $\mathcal{G}$ be a Markov α-potential game with α-potential function Φ. Then $\forall s \in S, \pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$,*

$$|V_i(s, \pi_i, \pi_{-i}) - V_i(s, \pi_i', \pi_{-i}) - (\Phi(s, \pi_i, \pi_{-i}) - \Phi(s, \pi_i', \pi_{-i}))| \leqslant \alpha. \tag{2.5}$$

Next, we show existence of an α-potential function.

Theorem 2.2.1. *(Existence of α-potential function). For any Markov game $\mathcal{G}$, there exists $\Phi \in \mathcal{F}^{\mathcal{G}}$ such that*

$$\widehat{\mathbf{d}}(\Phi, \mathcal{G}) = \inf_{\Psi \in \mathcal{F}^{\mathcal{G}}} \widehat{\mathbf{d}}(\Psi, \mathcal{G}).$$

[1] A set $\mathcal{F}$ of functions $f : S \times \Pi \to \mathbb{R}$ is called *uniformly equi-continuous on* Π, if there exists $\delta_{\mathcal{F}} : \mathbb{R}_+ \to \mathbb{R}_+$ such that for every $\epsilon > 0$, $|f(s, \pi) - f(s, \pi')| \leqslant \epsilon$ for all $f \in \mathcal{F}, s \in S$, $\pi, \pi' \in \Pi$ such that $\mathbf{d}(\pi, \pi') \leqslant \delta_{\mathcal{F}}(\epsilon)$.

Proof. Define a mapping

$$\begin{aligned}&\mathcal{F}^{\mathcal{G}} \times \Pi \times \Pi \ni (\Psi, \pi, \pi') \mapsto h(\Psi, \pi, \pi') \\ &\quad := \max_{s \in S, i \in I} \left| \Psi\left(s, \pi_i', \pi_{-i}\right) - \Psi\left(s, \pi_i, \pi_{-i}\right) - \left(V_i\left(s, \pi_i', \pi_{-i}\right) - V_i\left(s, \pi_i, \pi_{-i}\right)\right) \right| \in \mathbb{R}.\end{aligned}$$

Note that such h is continuous under the standard topology induced by sup-norm on $\mathcal{F}^{\mathcal{G}} \times \Pi \times \Pi$. By Berge's maximum theorem, $g(\Psi) := \max_{\pi, \pi' \in \Pi} h(\Psi, \pi, \pi')$ is continuous with respect to Ψ. Since $\mathcal{F}^{\mathcal{G}}$ is uniformly bounded and uniformly equi-continuous, Arzelà–Ascoli theorem implies that $\mathcal{F}^{\mathcal{G}}$ is relatively compact in $\mathcal{C}^{\Pi}$, where $\mathcal{C}^{\Pi} := \{f : S \times \Pi \to \mathbb{R} \mid \forall s \in S, f(s, \cdot) \text{ is a continuous function}\}$ [357]. Finally, by extreme-value theorem [357], there exists a function $\Phi \in \mathcal{F}^{\mathcal{G}}$ such that $\widehat{\mathbf{d}}(\Phi, \mathcal{G}) = \inf_{\Psi \in \mathcal{F}^{\mathcal{G}}} \widehat{\mathbf{d}}(\Psi, \mathcal{G})$. □

Corollary 2.2.1 and Theorem 2.2.1 jointly show that for any Markov game $\mathcal{G}$, an α-potential function exists such that the gap between the change in the utility function of any agent due to a unilateral change in its policy and the change in α-potential function is at most α. Next, we show that any optimizer of α-potential function with respect to policy π yields an α-Nash equilibrium (NE) of game $\mathcal{G}$.

Proposition 2.2.1. *Given a Markov α-potential game $\mathcal{G}$ with an α-potential function Φ, for any $\epsilon > 0$, if there exists a $\pi^* \in \Pi$ such that for every $s \in S$, $\Phi(s, \pi^*) + \epsilon \geqslant \sup_{\pi \in \Pi} \Phi(s, \pi)$, then $\pi^* \in \Pi$ is an $(\alpha + \epsilon)$-stationary NE of $\mathcal{G}$.*

Remark 2.2.1. *Note that Proposition 2.2.1 holds for any function $\Psi \in \mathcal{F}^{\mathcal{G}}$ that yields an upper bound for α. That is, given a Markov α-potential game $\mathcal{G}$ and a function Ψ satisfying*

$$\begin{aligned}&|V_i(s, \pi_i, \pi_{-i}) - V_i(s, \pi_i', \pi_{-i}) - (\Psi(s, \pi_i, \pi_{-i}) - \Psi(s, \pi_i', \pi_{-i}))| \leqslant \bar{\alpha}, \\ &\forall s \in S, \pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i},\end{aligned}$$

for some $\bar{\alpha} \in [\alpha, \infty)$, then for any $\pi^* \in \Pi$ such that $\Psi(s, \pi^*) + \epsilon \geqslant \sup_{\pi \in \Pi} \Psi(s, \pi)$ for any $s \in S$, π^* is an $(\bar{\alpha} + \epsilon)$-stationary NE of $\mathcal{G}$.

2.3 Examples of Markov α-potential game

In this section, we present three important classes of games, Markov potential games, Markov congestion games, and perturbed Markov team games, which can be analytically analyzed within the framework of Markov α-potential games.

Markov potential game

A game is a Markov potential game if there exists an auxiliary function (a.k.a. potential function) such that when a player unilaterally deviates from her policy, the change of the potential function is equal to the change of her utility function.

Definition 2.3.1 (Markov potential games [238]). *A Markov game $\mathcal{G}$ is a* Markov potential game *(MPG) if there exists a potential function $\Phi : S \times \Pi \to \mathbb{R}$ such that for any $i \in I$, $s \in S$, $\pi_i, \pi_i' \in \Pi_i$, and $\pi_{-i} \in \Pi_{-i}$, $\Phi(s, \pi_i', \pi_{-i}) - \Phi(s, \pi_i, \pi_{-i}) = V_i(s, \pi_i', \pi_{-i}) - V_i(s, \pi_i, \pi_{-i})$.*

Proposition 2.3.1. *An MPG is a Markov α-potential game with $\alpha = 0$.*

Markov congestion game

The Markov congestion game (MCG) $\mathcal{G}_{\mathsf{mcg}}$ is a dynamic counterpart to the static congestion game introduced by [306], involving a finite number of players using a finite set of resources. Each stage of $\mathcal{G}_{\mathsf{mcg}}$ is a static congestion game with a state-dependent reward function for each resource, and the state transition depends on the aggregated usage of each resource by the players. Specifically, let the finite set of resources in the one-stage congestion game be denoted as E. The action $a_i \in A_i \subseteq 2^E$ of each player $i \in I$ represents the set of resources chosen by player i. Here, the action set A_i is the set of all resource combinations that are feasible for player i. The total usage demand of all players is 1, and each player's demand is assumed to be $1/N$.

Given an action profile $a = (a_i)_{i \in I}$, the aggregated usage demand of each resource $e \in E$ is given by

$$w_e(a) = \frac{1}{N} \sum_{i \in I} \mathbb{1}(e \in a_i). \tag{2.6}$$

In each state s, the reward for using resource e is denoted as $(1/N) \cdot c_e(s, w_e(a))$. Thus, the one-stage payoff for player $i \in I$ in state $s \in S$, given the joint action profile $a \in A$, is $u_i(s, a) = (1/N) \cdot \sum_{e \in a_i} c_e(s, w_e(a))$. The state transition probability, denoted as $P(s'|s, w)$, depends on the aggregate usage vector $w = (w_e)_{e \in E}$, which is induced by the players' action profile as in (2.6). The set of all feasible aggregate usage demands is denoted by W.

The next proposition shows that, under a regularity condition on the state transition probability, $\mathcal{G}_{\mathsf{mcg}}$ is a Markov α-potential game such that the upper bound of α scales linearly with respect to the Lipschitz constant ζ, the size of state space $|S|$, resource set $|E|$, and decreases as N increases.

Proposition 2.3.2. *If there exists some $\zeta > 0$ such that for any $s, s' \in S, w, w' \in W$, $|P(s'|s, w) - P(s'|s, w')| \leqslant \zeta \|w - w'\|_1$, then the congestion game $\mathcal{G}_{\mathsf{mcg}}$ is a Markov α-potential game with $\alpha \leqslant 2\zeta\gamma|S||E| \sup_{s,\pi} \Psi(s, \pi)/(N(1-\gamma))$, where*

$$\Psi(\mu, \pi) := \frac{1}{N} \mathbb{E}_{\mu,\pi} \left[\sum_{k=0}^{\infty} \gamma^k \left(\sum_{e \in E} \sum_{j=1}^{w_e^k N} c_e\Big(s^k, \frac{j}{N}\Big) \right) \right], \tag{2.7}$$

such that $s^0 \sim \mu$, the aggregate usage vector $w^k = (w_e^k)_{e \in E}$ is induced by $a^k \sim \pi(s^k)$, and $s^k \sim P(\cdot|s^{k-1}, w^{k-1})$.

Perturbed Markov team game

A Markov game is called a perturbed Markov team game (PMTG) $\mathcal{G}_{\mathsf{pmtg}}$ if the payoff function for each player $i \in I$ can be decomposed as $u_i(s,a) = r(s,a) + \xi_i(s,a)$. Here, $r(s,a)$ represents the common interest of the team, and $\xi_i(s,a)$ represents player i's heterogeneous preference, such that $\|\xi_i\|_{L_\infty} \leqslant \kappa$, where $\kappa \geqslant 0$ measures each player's deviation from the team's common interest. As $\kappa \to 0$, $\mathcal{G}_{\mathsf{pmtg}}$ becomes a Markov team game, which is an MPG [238].

The next proposition shows that a $\mathcal{G}_{\mathsf{pmtg}}$ is a Markov α-potential game, and the upper bound of α decreases as the magnitude of the payoff perturbation κ decreases.

Proposition 2.3.3. *A perturbed Markov team game $\mathcal{G}_{\mathsf{pmtg}}$ is a Markov α-potential game with $\alpha \leqslant \frac{2\kappa}{(1-\gamma)^2}$.*

2.4 Finding an upper bound of α

The analysis of MCG and PMTG in Section 2.3 utilizes a specific form of the Markov α-potential function to obtain an upper bound on α. In this section, we provide an optimization-based procedure to find an upper bound on α by also computing the α-potential function.

Our approach is based on changing the feasible set of the optimization problem in (2.4) to $\tilde{\mathcal{F}}^{\mathcal{G}}$, defined as follows:

$$\tilde{\mathcal{F}}^{\mathcal{G}} := \Bigg\{ \Psi(s,\pi) = \sum_{s' \in S, a' \in A} d^s(s',a';\pi)\phi(s',a'), \forall s \in S, \pi \in \Pi \Big| \\ \exists\ \phi : S \times A \to \mathbb{R} \text{ s.t. } \|\phi\|_\infty \leqslant N \max_{i \in I} \|u_i\|_\infty \Bigg\}, \tag{2.8}$$

where, for any $s \in S$, $d^s(\cdot;\pi) : S \times A \to \mathbb{R}$ is the *state-action occupancy measure* induced due to π, defined as follows:

$$d^s(s',a';\pi) := \pi(a'|s') \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k \mathbb{1}(s^k = s') \Big| s^0 = s \right],$$

where $a^k \sim \pi\left(s^k\right)$, and $s^{k+1} \sim P\left(\cdot|s^k, a^k\right)$. Intuitively, for any $\Psi \in \tilde{\mathcal{F}}^{\mathcal{G}}$, there exists $\phi : S \times A \to \mathbb{R}$ such that $\Psi(s,\pi)$ represents the long-horizon discounted value of a Markov decision process with state transition P, starting from state s, using policy $\pi \in \Pi$, and one-step utility ϕ.

Proposition 2.4.1. *For any Markov α-potential game $\mathcal{G}$, $\tilde{\mathcal{F}}^{\mathcal{G}} \subseteq \mathcal{F}^{\mathcal{G}}$. That is, $\bar{\alpha} \geqslant \alpha$ with*

$$\bar{\alpha} := \inf_{\Psi \in \tilde{\mathcal{F}}^{\mathcal{G}}} \widehat{\mathbf{d}}(\Psi, \mathcal{G}). \tag{2.9}$$

Using Remark 2.2.1, we can conclude that any optimizer of $\bar{\Psi}$, where $\widehat{\mathbf{d}}(\bar{\Psi}, \mathcal{G}) = \bar{\alpha}$, can be used to find a $\bar{\alpha}$-stationary NE for the game $\mathcal{G}$.

Next, we provide an optimization based method to compute $\bar{\alpha}$. Note that (2.9) can be reformulated as follows:

$$\begin{aligned} \min_{\substack{y\in\mathbb{R} \\ \phi: S\times A\to\mathbb{R}}} \quad & y && (2.10)\\ \text{s.t.} \quad & \Big| \sum_{s',a'} \big(d^s(s',a';\pi_i,\pi_{-i}) - d^s(s',a';\pi_i',\pi_{-i})\big)\cdot(\phi - u_i)(s',a') \Big| \leqslant y, && (\text{C1})\\ & \forall s\in S,\ \forall i\in I,\ \forall \pi_i,\pi_i'\in\Pi_i,\ \forall \pi_{-i}\in\Pi_{-i}, \\ & |\phi(s,a)| \leqslant N \max_{i\in I} \|u_i\|_\infty, \quad \forall s\in S, a\in A. \end{aligned}$$

Here, we use

$$V_i(s,\pi) = \sum_{s'\in S, a'\in A} d^s(s',a';\pi) u_i(s',a'),$$

and

$$\Psi(s,\pi) = \sum_{s'\in S, a'\in A} d^s(s',a';\pi)\phi(s',a'),$$

for some $\phi : S\times A\to\mathbb{R}$. Note that (2.10) is a semi-infinite linear program where the objective is a linear function with an uncountable number of linear constraints. Particularly, in (C1) there is one linear constraint corresponding to each tuple $(s,i,\pi_i,\pi_i',\pi_{-i})$. Moreover, the coefficients of each linear constraint in (C1) are composed of state-action occupancy measures which are computed by solving a Bellman equation. There are a number of algorithmic approaches to solve semi-infinite linear programming problems [402, 180]. In Appendix A, we adopt an algorithm from [402] to solve (2.10) and find (an upper bound of) α. Figure 2.1 illustrates how α varies with different discount factors γ in a PMTG. Note that the growth of the numerical estimate of α is much more benign than the analytical characterization obtained in Proposition 2.3.3.

2.5 Approximation algorithms and Nash-regret analysis

In this section, we present two equilibrium approximation algorithms for Markov α-potential games: the *projected gradient-ascent algorithm*, proposed in [121] for MPGs, and the *sequential maximum improvement algorithm*, where each player's strategy is updated based on a one-stage smoothed best response. We also derive non-asymptotic convergence rates for these algorithms in terms of Nash-regret, defined as

$$\text{Nash-regret}(T) := \frac{1}{T}\sum_{t=1}^{T} \max_{i\in I} R_i^{(t)},$$

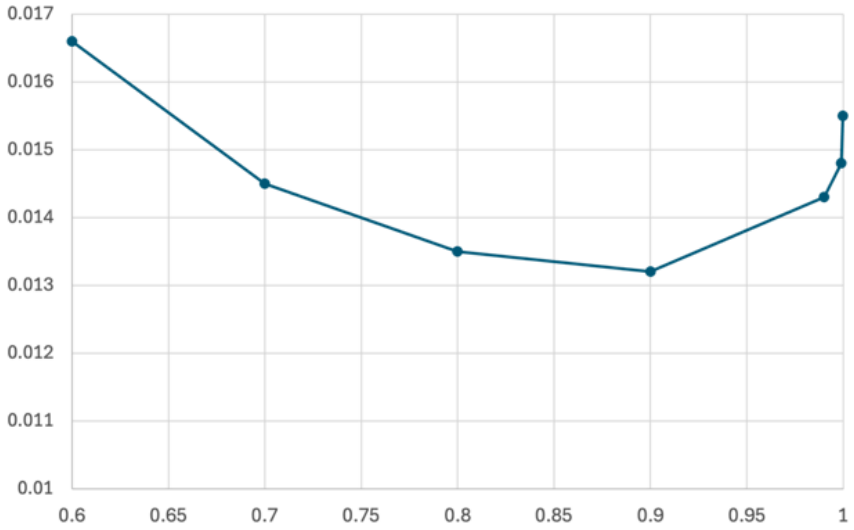


Figure 2.1: Variation of α with the discount factor in the perturbed Markov team game with $N = 3$ and perturbation parameter $\kappa = 0.1$. The setup of this game is same as that in Section 2.6 with $\lambda_1 = \lambda_3 = 0.8, \lambda_2 = \lambda_4 = 0.2$.

where

$$R_i^{(t)} := \max_{\pi_i' \in \Pi_i} V_i\left(\mu, \pi_i', \pi_{-i}^{(t)}\right) - V_i\left(\mu, \pi^{(t)}\right),$$

and $\pi^{(t)}$ denotes the t-th iterate. Note that Nash-regret is always non-negative; if

$$\text{Nash-regret}(T) \leqslant \epsilon$$

for some $\epsilon > 0$, then there exists $t^\dagger$ such that $\pi^{(t^\dagger)}$ is an ϵ-approximate NE.

Projected gradient-ascent algorithm

First, we define some useful notations. Given a joint policy $\pi \in \Pi$, define player i's *Q-function* as

$$Q_i^\pi(s, a_i) = \mathbb{E}_{a_{-i} \sim \pi_{-i}(s)}\Big[u_i(s, a_i, a_{-i}) + \gamma \sum_{s' \in S} P(s'|s, a_i, a_{-i}) V_i(s', \pi)\Big],$$

and denote $\mathbf{Q}_i^\pi(s) = (Q_i^\pi(s, a_i))_{a_i \in A_i}$. Let κ_μ denote the maximum distribution mismatch of π relative to μ, and let $\tilde{\kappa}_\mu$ denote the minimax value of the distribution mismatch of π relative to μ. That is,

$$\kappa_\mu := \sup_{\pi \in \Pi} \left\| d_\mu^\pi / \mu \right\|_\infty, \quad \tilde{\kappa}_\mu := \inf_{\nu \in \Delta(S)} \sup_{\pi \in \Pi} \|d_\mu^\pi / \nu\|_\infty, \tag{2.11}$$

where d_μ^π is defined in (2.2), and the division d_μ^π/ν is evaluated in a component-wise manner. The algorithm iterates for T steps. We abuse the notation to use $Q_i^{(t)}$ to denote $Q_i^{\pi^{(t)}}$, and $\mathbf{Q}_i^{(t)}$ to denote $\mathbf{Q}_i^{\pi^{(t)}}$. In every step $t \in [T-1]$, each player $i \in I$ updates her policy following a projected gradient-ascent algorithm as in (2.12).

Algorithm 1 Projected Gradient-Ascent Algorithm

Input: Step size η, for every $i \in I, a_i \in A_i, s \in S$, set $\pi_i^{(0)}(a_i|s) = 1/|A_i|$.
for $t = 0, 1, 2, ..., T-1$ **do**
For every $i \in I, s \in S$, update the policies as follows

$$\pi_i^{(t+1)}(s) = \texttt{Proj}_{\Pi_i}\left(\pi_i^{(t)}(s) + \eta \mathbf{Q}_i^{(t)}(s)\right), \tag{2.12}$$

where $\texttt{Proj}_{\Pi_i}$ denotes the orthogonal projection on Π_i.
end for

Remark 2.5.1. *Algorithm 1 is not the standard policy gradient algorithm. The standard policy gradient (cf. [238]) is given by*

$$\frac{\partial V_i^\pi(\rho)}{\partial \pi_i\left(a_i \mid s\right)} = 1/(1-\gamma) \cdot d_\rho^\pi(s) Q_i^\pi\left(s, a_i\right).$$

The RHS in the this equation scales with the state visitation frequency $d_\rho^\pi(s)$, which results in slow learning rate for states with low visitation frequencies under the current policy. To address this issue, [121] proposed to remove the term $d_\rho^\pi(s)/(1-\gamma)$ from the standard policy gradient update, which accelerates the learning for states with low visitation probabilities. We adopted the convention of [121] to call it "policy gradient-ascent algorithm".

Theorem 2.5.1. *Given a Markov α-potential game with an α-potential function Φ and an initial state distribution μ, the policy updates generated from Algorithm 1 satisfies*

(i) $\textit{Nash-regret}(T) \leqslant \mathcal{O}\left(\frac{\sqrt{\tilde{\kappa}_\mu \bar{A} N}}{(1-\gamma)^{\frac{9}{4}}}\left(\frac{C_\Phi}{T} + N^2\alpha\right)^{\frac{1}{4}}\right)$ *with* $\eta = \frac{(1-\gamma)^{2.5}\sqrt{C_\Phi + N^2\alpha T}}{2N\bar{A}\sqrt{T}}$;

(ii) $\textit{Nash-regret}(T) \leqslant \mathcal{O}\left(\sqrt{\frac{\min\left(\kappa_\mu, |S|\right)^4 N\bar{A}}{(1-\gamma)^6}}\left(\frac{C_\Phi}{T} + N^2\alpha\right)^{\frac{1}{2}}\right)$ *with* $\eta = \frac{(1-\gamma)^4}{8\min\left(\kappa_\mu, |S|\right)^3 N\bar{A}}$,

where $\bar{A} \coloneqq \max_{i\in I}|A_i|$, κ_μ and $\tilde{\kappa}_\mu$ are defined in (2.11), and $C_\Phi > 0$ is a constant satisfying $|\Phi(\mu,\pi) - \Phi(\mu,\pi')| \leqslant C_\Phi$ for any $\pi, \pi' \in \Pi, \mu \in \Delta(S)$.

We emphasize that the Nash-regret bounds in Theorem 2.5.1 (also Theorem 2.5.2 in the next section) will hold even without knowing the exact form of Φ and the game elasticity parameter α. It is sufficient to have an upper bound $\bar{\alpha}$ for α and an associated function Ψ

for which this upper bound holds. In the special case of $\alpha = 0$, the Nash-regret bound in Theorem 2.5.1 recovers the Nash-regret bound from [121] for MPG.

The proof of Theorem 2.5.1 is inspired by [121] for the Nash-regret analysis of MPGs. First, we state multi-player performance difference lemma (Lemma 2.5.1), which enables bounding the Nash-regret of an algorithm by summing the norms of policy updates, denoted as $\|\pi_i^{(t+1)} - \pi_i^{(t)}\|$. The main modification for our analysis is to bound the sum of these policy update differences by the game elasticity parameter α and the change in the α-potential function Φ (Lemma 2.5.2).

Lemma 2.5.1 (Performance difference (Lemma 1 in [121])). *For any $i \in I$, $\mu \in \Delta(S)$, $\pi_i', \pi_i \in \Pi_i$, and $\pi_{-i} \in \Pi_{-i}$,*

$$\begin{aligned} &V_i(\mu, \pi_i', \pi_{-i}) - V_i(\mu, \pi_i, \pi_{-i}) \\ &= \frac{1}{1-\gamma} \sum_{s,a_i} d_\mu^{\pi_i', \pi_{-i}}(s) \cdot \left(\pi_i'(a_i|s) - \pi_i(a_i|s)\right) Q_i^{\pi_i, \pi_{-i}}(s, a_i). \end{aligned}$$

Lemma 2.5.2 (Policy improvement). *For Markov α-potential game (2.4) with any state distribution $\nu \in \Delta(S)$, the α-potential function $\Phi(\nu, \pi)$ at two consecutive policies $\pi^{(t+1)}$ and $\pi^{(t)}$ in Algorithm 1 satisfies*

$$\begin{aligned} &(i)\, \Phi(\nu, \pi^{(t+1)}) - \Phi(\nu, \pi^{(t)}) + N^2\alpha \geqslant -\frac{4\eta^2 \bar{A}^2 N^2}{(1-\gamma)^5} \\ &\quad + \frac{1}{2\eta(1-\gamma)} \sum_{i \in I, s \in S} d_\nu^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s) \left\| \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\|^2; \\ &(ii)\, \Phi(\nu, \pi^{(t+1)}) - \Phi(\nu, \pi^{(t)}) + N^2\alpha \geqslant \\ &\quad \frac{1}{2\eta(1-\gamma)} \left(1 - \frac{4\eta\kappa_\nu^3 \bar{A} N}{(1-\gamma)^4}\right) \sum_{i \in I, s \in S} d_\nu^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}(s)} \cdot \left\| \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\|^2. \end{aligned}$$

Proof of Theorem 2.5.1

Using the variational characterization of projection operation in (2.12), we note that for any $\pi_i' \in \Pi_i$,

$$\left\langle \pi_i'(s) - \pi_i^{(t+1)}(s), \eta \mathbf{Q}_i^{(t)}(s) - \pi_i^{(t+1)}(s) + \pi_i^{(t)}(s) \right\rangle_{A_i} \leqslant 0.$$

Therefore, for any $\pi_i' \in \Pi_i$,

$$\begin{aligned} &\left\langle \pi_i'(s) - \pi_i^{(t)}(s), \mathbf{Q}_i^{(t)}(s) \right\rangle_{A_i} \\ &= \left\langle \pi_i'(s) - \pi_i^{(t+1)}(s), \mathbf{Q}_i^{(t)}(s) \right\rangle_{A_i} + \left\langle \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s), \mathbf{Q}_i^{(t)}(s) \right\rangle_{A_i} \\ &\leqslant \frac{1}{\eta} \left\langle \pi_i'(s) - \pi_i^{(t+1)}(s), \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\rangle_{A_i} + \left\langle \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s), \mathbf{Q}_i^{(t)}(s) \right\rangle_{A_i}. \end{aligned}$$

Note that for any two probability distributions p_1 and p_2,

$$\|p_1 - p_2\| \leqslant \|p_1 - p_2\|_1 \leqslant 2.$$

Therefore,

$$\begin{aligned}
&\left\langle \pi_i'(s) - \pi_i^{(t)}(s), \mathbf{Q}_i^{(t)}(s) \right\rangle_{A_i} \\
&\leqslant \frac{2}{\eta} \left\| \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\| + \left\| \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\| \left\| \mathbf{Q}_i^{(t)}(s) \right\| \\
&\leqslant \frac{3}{\eta} \left\| \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\|, \qquad (2.13)
\end{aligned}$$

where the last inequality is due to $\left\| \mathbf{Q}_i^{(t)}(s) \right\| \leqslant \frac{\sqrt{A}}{1-\gamma}$ and $\eta \leqslant \frac{1-\gamma}{\sqrt{A}}$. Hence, by Lemma 2.5.1 and (2.13),

$$\begin{aligned}
T \cdot \text{Nash-regret}(T) &= \sum_{t=1}^{T} \max_{i \in I, \pi_i'} V_i(\mu, \pi_i', \pi_{-i}^{(t)}) - V_i(\mu, \pi^{(t)}) \\
&= \sum_{t=1}^{T} \max_{\pi_i'} \sum_{s, a_i} \frac{d_\mu^{\pi_i', \pi_{-i}^{(t)}}(s)}{1-\gamma} (\pi_i'(a_i|s) - \pi_i^{(t)}(a_i|s)) \mathbf{Q}_i^{(t)}(s, a_i) \\
&\leqslant \frac{3}{\eta(1-\gamma)} \sum_{t=1}^{T} \sum_{s} d_\mu^{\pi_i', \pi_{-i}^{(t)}}(s) \left\| \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\|,
\end{aligned}$$

where in the second line we slightly abuse the notation i to represent $\arg\max_i$ and in the last line we slightly abuse the notation π_i' to represent $\arg\max_{\pi_i'}$. Now, continuing the above calculation with an arbitrary $\nu \in \Delta(S)$ and using

$$\frac{d_\mu^{\pi_i', \pi_{-i}^{(t)}}(s)}{d_\nu^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s)} \leqslant \frac{d_\mu^{\pi_i', \pi_{-i}^{(t)}}(s)}{(1-\gamma)\nu(s)} \leqslant \frac{\sup_{\pi \in \Pi} \left\| d_\mu^\pi / \nu \right\|_\infty}{1-\gamma}$$

to get:

$$\begin{aligned}
&T \cdot \text{Nash-regret}(T) \\
&\leqslant \frac{3\sqrt{\sup_{\pi \in \Pi} \left\| d_\mu^\pi / \nu \right\|_\infty}}{\eta(1-\gamma)^{\frac{3}{2}}} \sum_{t=1}^{T} \sum_{s} \sqrt{d_\mu^{\pi_i', \pi_{-i}^{(t)}}(s) d_\nu^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s)} \cdot \left\| \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\| \\
&\leqslant \frac{3\sqrt{\sup_{\pi \in \Pi} \left\| d_\mu^\pi / \nu \right\|_\infty}}{\eta(1-\gamma)^{\frac{3}{2}}} \sqrt{\sum_{t=1}^{T} \sum_{s} d_\mu^{\pi_i', \pi_{-i}^{(t)}}(s)} \qquad (2.14) \\
&\qquad \cdot \sqrt{\sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{s} d_\nu^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}(s)} \left\| \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\|^2}, \qquad (2.15)
\end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality and replacing $\arg\max_i$ by the sum over all players. There are two choices to proceed beyond (2.14):
1) Fix $\epsilon > 0$. Take $\nu_\epsilon^* \in \Delta(S)$ such that

$$\sup_{\pi\in\Pi} \left\| d_\mu^\pi / \nu_\epsilon^* \right\|_\infty - \epsilon \leqslant \inf_{\nu\in\Delta(S)} \sup_{\pi\in\Pi} \left\| d_\mu^\pi / \nu_\epsilon^* \right\|_\infty .$$

Then apply Lemma 2.5.2 (i) and the fact $|\Phi(\nu,\pi) - \Phi(\nu,\pi')| \leqslant C_\Phi$ for any $\pi, \pi' \in \Pi, \nu \in \Delta(S)$ to get

$$\text{Nash-regret}(T) \leqslant \frac{3}{T}\left(\frac{2(\widetilde{\kappa}_\mu + \epsilon)T(C_\Phi + N^2\alpha \cdot T)}{\eta(1-\gamma)^2} + \frac{8(\widetilde{\kappa}_\mu + \epsilon)\eta T^2 \bar{A}^2 N^2}{(1-\gamma)^7} \right)^{\frac{1}{2}}.$$

By letting ϵ to 0 and taking step size $\eta = \frac{(1-\gamma)^{2.5}\sqrt{C_\Phi + N^2\alpha T}}{2N\bar{A}\sqrt{T}}$, we have

$$\text{Nash-regret}(T) \leqslant \frac{3 \cdot 2^{\frac{3}{2}}\sqrt{\widetilde{\kappa}_\mu \bar{A} N}}{(1-\gamma)^{\frac{9}{4}}} \left(\frac{C_\Phi}{T} + N^2\alpha \right)^{\frac{1}{4}}.$$

2) We can also proceed (2.14) with Lemma 2.5.2 (ii) and $\eta \leqslant \frac{(1-\gamma)^4}{8\kappa_\nu^3 N\bar{A}}$ to get

$$\text{Nash-regret}(T) \leqslant 6\sqrt{\frac{\sup_{\pi\in\Pi} \left\| \frac{d_\mu^\pi}{\nu} \right\|_\infty (C_\Phi + N^2\alpha \cdot T)}{\eta T(1-\gamma)^2}}.$$

We next discuss two special choices of ν for proving our bound. First, if $\nu = \mu$, then $\eta \leqslant \frac{(1-\gamma)^4}{8\kappa_\mu^3 N\bar{A}}$. By letting $\eta = \frac{(1-\gamma)^4}{8\kappa_\mu^3 N\bar{A}}$, the last square root term can be bounded by

$$\mathcal{O}\left(\sqrt{\frac{\kappa_\mu^4 N\bar{A}(C_\Phi + N^2\alpha \cdot T)}{T(1-\gamma)^6}} \right).$$

Second, if $\nu = \frac{1}{|S|}\mathbf{1}$, the uniform distribution over S, then $\kappa_\nu \leqslant \frac{1}{S}$, which allows a valid choice $\eta = \frac{(1-\gamma)^4}{8|S|^3 N\bar{A}} \leqslant \frac{(1-\gamma)^4}{8\kappa_\nu^3 N\bar{A}}$. Hence, we can bound the last square root term by

$$\mathcal{O}\left(\sqrt{\frac{|S|^4 N\bar{A}(C_\Phi + N^2\alpha \cdot T)}{T(1-\gamma)^6}} \right).$$

Since ν is arbitrary, combining these two special choices completes the proof.

Sequential maximum improvement algorithm

Let us first fix some notations. Associated with any Markov game $\mathcal{G}$, we define *smoothed* (or regularized) Markov game $\tilde{\mathcal{G}}$, where the expected one-stage payoff of each player i with state s under the joint policy π is $\tilde{u}_i(s,\pi) = \mathbb{E}_{a\sim\pi(s)}[u_i(s,a)] - \tau\sum_{j\in I}\nu_j(s,\pi_j)$, where

$$\nu_j(s,\pi_j) \coloneqq \sum_{a_j\in A_j}\pi_j(a_j|s)\log(\pi_j(a_j|s))$$

is the entropy function, and $\tau > 0$ denotes the regularization parameter. With the smoothed one-stage payoffs, the expected total discounted infinite horizon payoff of player i under policy π is given by

$$\tilde{V}_i(s,\pi) = \mathbb{E}_\pi\Big[\sum_{k=0}^{\infty}\gamma^k\Big(u_i(s^k,a^k) - \tau\sum_{j\in I}\nu_j(s^k,\pi_j)\Big)|s^0 = s\Big], \tag{2.16}$$

for every $s \in S$. The *smoothed* (or entropy-regularized) Q-function is given by

$$\begin{aligned}&\tilde{Q}_i^\pi(s,a_i)\\ &= \sum_{a_{-i}\in A_{-i}}\pi_{-i}(a_{-i}|s)\Big(u_i(s,a_i,a_{-i}) - \tau\sum_{j\in I_N}\nu_j(s,\pi_j) + \gamma\sum_{s'\in S}P(s'|s,a)\tilde{V}_i(s',\pi)\Big).\end{aligned} \tag{2.17}$$

Algorithm 2 has two main components: first, it computes the optimal one-stage policy update, by using the smoothed Q-function. Here the vector of smoothed Q-functions for all $a_i \in A_i$ is denoted by $\tilde{\mathbf{Q}}_i^\pi(s) = (\tilde{Q}_i^\pi(s,a_i))_{a_i\in A_i}$. Second, it selects the player who achieves the maximum improvement in the current state to adopt her one-stage policy update, with the policy for the remaining players and the remaining states unchanged. More specifically, the algorithm iterates for T time steps. In every time step $t \in [T-1]$, based on the current policy profile $\pi^{(t)}$, we abuse the notation to use $\tilde{Q}_i^{(t)}$ to denote $\tilde{Q}_i^{\pi^{(t)}}$ and $\tilde{\mathbf{Q}}_i^{(t)}$ to denote $\tilde{\mathbf{Q}}_i^{\pi^{(t)}}$. The expected smoothed Q-function of player i is computed as $\tilde{Q}_i^{(t)}(s,\pi_i) = \sum_{a_i\in A_i}\pi_i(a_i|s)\tilde{Q}_i^{(t)}(s,a_i)$ for all $s \in S$ and all $i \in I$. Then, each player computes her one-stage best response strategy by maximizing the smoothed Q-function: for every $i \in I, a_i \in A_i, s \in S$,

$$\begin{aligned}\mathrm{BR}_i^{(t)}(a_i|s) &= \left(\arg\max_{\pi_i'\in\Pi_i}\Big(\tilde{Q}_i^{(t)}(s,\pi_i') - \tau\nu_i(s,\pi_i')\Big)\right)_{a_i}\\ &= \frac{\exp(\tilde{Q}_i^{(t)}(s,a_i)/\tau)}{\sum_{a_i'\in A_i}\exp(\tilde{Q}_i^{(t)}(s,a_i')/\tau)},\end{aligned} \tag{2.18}$$

and its maximum improvement of smoothed Q-function value in comparison to current policy is

$$\Omega_i^{(t)}(s) = \max_{\pi_i'\in\Pi_i}\Big(\tilde{Q}_i^{(t)}(s,\pi_i') - \tau\nu_i(s,\pi_i')\Big) - \Big(\tilde{Q}_i^{(t)}(s,\pi_i^{(t)}) - \tau\nu_i(s,\pi_i^{(t)})\Big), \quad \forall s \in S. \tag{2.19}$$

Note that computing $\mathbf{\Omega}_i^{(t)}$ is straightforward as the maximization in (2.19) is attained at $\mathrm{BR}_i^{(t)}(s)$ (cf. (2.18)).

If the maximum improvement $\mathbf{\Omega}_i^{(t)}(s) \leqslant 0$ for all $i \in I$ and all $s \in S$, then the algorithm terminates and returns the current policy profile $\pi^{(t)}$. Otherwise, the algorithm chooses a tuple of player and state $(\bar{i}^{(t)}, \bar{s}^{(t)})$ associated with the maximum improvement value $\mathbf{\Omega}_i^{(t)}(s)$, and updates the policy of player $\bar{i}^{(t)}$ in state $\bar{s}^{(t)}$ with her one-stage best response strategy[2]. The policies of all other players and other states remain unchanged.

Remark 2.5.2. *Using entropy regularization in* (2.18) *has several advantages: (i) unlike Algorithm 1, it avoids projection over simplex which can be costly in large-scale problems; (ii) it ensures that the optimizer is unique.*

Remark 2.5.3. *Algorithm 2 is reminiscent of the "Nash-CA" algorithm*[3] *proposed in [391], which requires each player to sequentially compute the best response policy using an RL algorithm in each iteration, while keeping the strategies of other players fixed. Such sequential best response algorithms are known to ensure finite improvement in the potential function value in potential games [306], which ensures convergence. Meanwhile, Algorithm 2 does not compute the best response strategy in the updates. Instead, it only computes a smoothed* one-step *optimal deviation, as per* (2.18)*, for the current state. The policies for the remaining states and other players are unchanged. The analysis of such one-step deviation-based dynamics is non-trivial and requires new techniques, as highlighted in the next section.*

Remark 2.5.4. *While Algorithm 1 can be run independently by each player in a decentralized fashion, Algorithm 2 is centralized as players do not update their policies simultaneously. Comparing Nash regret in Theorems 2.5.1 and 2.5.2, it is evident that the coordination in Algorithm 2 ensures better scaling of regret with respect to the number of players.*

Theorem 2.5.2. *Consider a Markov α-potential game with an α-potential function Φ and initial state distribution μ such that $\bar{\mu} := \min_{s\in S} \mu(s) > 0$. Denote $\bar{A} := \max_{i\in I} |A_i|$ and $C := \max_{i\in I} \|u_i\|_\infty$. Then the policy updates generated from Algorithm 2 with parameter*

$$\tau = \frac{1}{N}\left(\log(\bar{A}) + \frac{\log(\bar{A})}{\sqrt{\alpha + \frac{C_\Phi}{T}}} \sqrt{\frac{2\log(\bar{A})}{(1-\gamma)}} \sqrt{\frac{N}{T}} + \frac{2\sqrt{\bar{\mu}}(1-\gamma)\log(\bar{A})}{8C\sqrt{\bar{A}}\sqrt{\alpha + \frac{C_\Phi}{T}}} \right)^{-1} \tag{2.22}$$

has the Nash-regret(T) bounded by

$$\mathcal{O}\left(\frac{\sqrt{N^{3/2}\bar{A}}\log(\bar{A})}{(1-\gamma)^{5/2}\sqrt{\bar{\mu}}} \max\left\{ \left(\alpha + \frac{C_\Phi}{T}\right)^{\frac{1}{2}}, \left(\alpha + \frac{C_\Phi}{T}\right)^{\frac{1}{4}} \right\} \right),$$

[2]Any tie-breaking rule can be used here if the maximum improvement is achieved by more than one tuple.

[3]Unlike here, the Nash-CA Algorithm in [391] was proposed in the context of finite horizon Markov potential games.

Algorithm 2 Sequential Maximum Improvement Algorithm

Input: Smoothness parameter τ, for every $i \in I, a_i \in A_i, s \in S$, set $\pi_i^{(0)}(a_i|s) = 1/|A_i|$.
for $t = 0, 1, 2, ..., T-1$ **do**
Compute the maximum improvement of smoothed Q-function $\{\Omega_i^{(t)}(s)\}_{i\in I, s\in S}$ as in (2.19).
if $\Omega_i^{(t)}(s) \leqslant 0$ for all $i \in I$ and all $s \in S$ **then**
return $\pi^{(t)}$.
else
Choose the tuple $(\bar{i}^{(t)}, \bar{s}^{(t)})$ with the maximum improvement

$$(\bar{i}^{(t)}, \bar{s}^{(t)}) \in \arg\max_{i\in I, s\in S} \Omega_i^{(t)}(s), \tag{2.20}$$

and update policy

$$\begin{aligned} &\pi_{\bar{i}^{(t)}}^{(t+1)}(a|\bar{s}^{(t)}) = \mathrm{BR}_{\bar{i}^{(t)}}^{(t)}(a|\bar{s}^{(t)}),\ \forall a \in A_{\bar{i}^{(t)}}, \\ &\pi_i^{(t+1)}(s) = \pi_i^{(t)}(s)\ \forall (i,s) \neq (\bar{i}^{(t)}, \bar{s}^{(t)}). \end{aligned} \tag{2.21}$$

end if
end for

where $C_\Phi > 0$ is a constant satisfying $|\Phi(\mu, \pi) - \Phi(\mu, \pi')| \leqslant C_\Phi$ for any $\pi, \pi' \in \Pi, \mu \in \Delta(S)$.

In the special case of $\alpha = 0$, Theorem 2.5.2 provides a Nash-regret bound of Algorithm 2 for the case of MPGs.

To prove Theorem 2.5.2, we first develop a smoothed version of the multi-agent performance difference lemma (Lemma 2.5.3). This lemma bounds the difference in the smoothed value function $\tilde{V}_i$ by the changes in policy π_i, which is further bounded by the maximum improvements $\Omega_i^{(t)}$. Lemma 2.5.4 bounds the discrepancy between the value function V_i and the smoothed value function $\tilde{V}_i$. Lemma 2.5.3 and 2.5.4 together implies that the Nash-regret of Algorithm 2 is bounded by $\Omega_i^{(t)}$ (2.19). Finally, Lemma 2.5.5 establishes $\Omega_i^{(t)}$ can be bounded by policy updates, which in turn, are bounded by α and the difference in the α-potential function Φ.

Lemma 2.5.3 (Smoothed performance difference)**.** *For any $i \in I$, $\mu \in \Delta(S)$, $\pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$,*

$$\begin{aligned} &\tilde{V}_i(\mu, \pi) - \tilde{V}_i(\mu, \pi') \\ &\qquad = \frac{1}{1-\gamma} \sum_{s'\in S} d_\mu^\pi(s') \Big((\pi_i(s') - \pi_i'(s'))^\top \cdot \tilde{\mathbf{Q}}_i^{\pi'}(s') + \tau\nu_i(s', \pi_i') - \tau\nu_i(s', \pi_i) \Big), \end{aligned}$$

where $\pi = (\pi_i, \pi_{-i})$, *and* $\pi' = (\pi'_i, \pi_{-i})$.

Lemma 2.5.4. *For any* $i \in I, \mu \in \Delta(S), \pi_i, \pi'_i \in \Pi_i, \pi_{-i} \in \Pi_{-i}$,

$$\left|V_i(\mu, \pi_i, \pi_{-i}) - V_i(\mu, \pi'_i, \pi_{-i}) - (\tilde{V}_i(\mu, \pi_i, \pi_{-i}) - \tilde{V}_i(\mu, \pi'_i, \pi_{-i}))\right| \leqslant \frac{2\tau N \log(\bar{A})}{1-\gamma}.$$

Lemma 2.5.5. *The following inequalities hold:*

(1)

$$\Omega^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)}) \leqslant \frac{4C\sqrt{\bar{A}}(1+\tau N \log(\bar{A}))}{1-\gamma} \|\pi^{(t+1)}_{\bar{i}(t)}(\bar{s}^{(t)}) - \pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)})\|_2, \quad \textit{for any } t \in [T]. \tag{2.23}$$

(2)

$$\begin{aligned}&\sum_{t=0}^{T-1} \|\pi^{(t+1)}_{\bar{i}(t)}(\bar{s}^{(t)}) - \pi^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)})\|_2^2 \\ &\leqslant \frac{2}{\tau\bar{\mu}}\Big(|\Phi(\mu, \pi^{(T)}) - \Phi(\mu, \pi^{(0)})| + \alpha T + \frac{2\tau N \log(\bar{A})}{1-\gamma}\Big).\end{aligned} \tag{2.24}$$

Proof of Theorem 2.5.2

First, we bound the instantaneous regret $R_i^{(t)}$ for any arbitrary player $i \in I$ at time $t \in [T]$. Recall that

$$R_i^{(t)} = V_i(\mu, \pi_i^\dagger, \pi_{-i}^{(t)}) - V_i(\mu, \pi^{(t)}),$$

where $\pi_i^\dagger \in \arg\max_{\pi'_i \in \Pi_i} V_i(\mu, \pi'_i, \pi_{-i}^{(t)})$. By Lemma 2.5.4,

$$R_i^{(t)} \leqslant \tilde{V}_i(\mu, \pi_i^\dagger, \pi_{-i}^{(t)}) - \tilde{V}_i(\mu, \pi^{(t)}) + \frac{2\tau N \log(\bar{A})}{(1-\gamma)}.$$

Next, note that for any $i \in I, \mu \in \Delta(S)$, by Lemma 2.5.3,

$$\begin{aligned}&\tilde{V}_i(\mu, \pi_i^\dagger, \pi_{-i}^{(t)}) - \tilde{V}_i(\mu, \pi_i^{(t)}, \pi_{-i}^{(t)}) \\ \leqslant &\frac{1}{1-\gamma}\sum_{s\in S} d_\mu^{\pi_i^\dagger, \pi_{-i}^{(t)}}(s)\Bigg(\tau(\nu_i(s, \pi_i^{(t)}) - \nu_i(s, \pi'_i)) \\ &\quad + \max_{\pi'_i} \sum_{a_i \in A_i} \left(\left(\pi'_i(a_i|s) - \pi_i^{(t)}(a_i|s)\right)\tilde{Q}_i^{(t)}(s, a_i)\right)\Bigg) \\ \overset{(a)}{=} &\frac{1}{1-\gamma}\sum_{s\in S} d_\mu^{\pi_i^\dagger, \pi_{-i}^{(t)}}(s)\Omega_i^{(t)}(s) \\ \overset{(b)}{\leqslant} &\frac{1}{1-\gamma}\sum_{s\in S} d_\mu^{\pi_i^\dagger, \pi_{-i}^{(t)}}(s)\Omega^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)}) = \frac{1}{1-\gamma}\left(\Omega^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)})\right),\end{aligned}$$

where (a) is by (2.19), (b) holds since $\Omega_i^{(t)}(s) \leqslant \Omega_{\bar{i}(t)}^{(t)}(\bar{s}^{(t)})$ for all $i \in I, s \in S$. To summarize,

$$R_i^{(t)} \leqslant \frac{1}{1-\gamma}\left(\Omega_{\bar{i}(t)}^{(t)}(\bar{s}^{(t)}) + 2\tau N \log(\bar{A})\right).$$

Then by Lemma 2.5.5 (1),

$$\begin{aligned}
\text{Nash-regret}(T) &\leqslant \frac{1}{T(1-\gamma)} \sum_{t\in[T]} \left(\Omega_{\bar{i}(t)}^{(t)}(\bar{s}^{(t)}) + 2\tau N \log(\bar{A})\right) \\
&\leqslant \frac{2\tau N \log(\bar{A})}{(1-\gamma)} + \frac{4C\sqrt{\bar{A}}(1+\tau N \log(\bar{A}))}{T(1-\gamma)^2} \sum_{t\in[T]} \left\| \pi_{\bar{i}(t)}^{(t+1)}(\bar{s}^{(t)}) - \pi_{\bar{i}(t)}^{(t)}(\bar{s}^{(t)}) \right\|_2 \\
&\leqslant \frac{2\tau N \log(\bar{A})}{(1-\gamma)} + \frac{4C\sqrt{\bar{A}}(1+\tau N \log(\bar{A}))}{\sqrt{T}(1-\gamma)^2} \Big(\sum_{t\in[T]} \left\| \pi_{\bar{i}(t)}^{(t+1)}(\bar{s}^{(t)}) - \pi_{\bar{i}(t)}^{(t)}(\bar{s}^{(t)}) \right\|_2^2 \Big)^{\frac{1}{2}}, \qquad (2.25)
\end{aligned}$$

where the last inequality follows from Cauchy-Schwarz inequality. For ease of exposition, define $D_1 \coloneqq \frac{8C\sqrt{\bar{A}}}{\sqrt{\bar{\mu}}(1-\gamma)^2}$, $D_2 \coloneqq \sqrt{\alpha + \frac{C_\Phi}{T}}$, and $D_3 \coloneqq \sqrt{\frac{2\log(\bar{A})}{(1-\gamma)}}$. Then by Lemma 2.5.5 (2),

$$\begin{aligned}
(2.25) &\leqslant \frac{D_1(1+\tau N \log(\bar{A}))}{\sqrt{\tau}} \sqrt{D_2^2 + \frac{\tau N}{T} D_3^2} + \tau N D_3^2 \\
&\leqslant \frac{D_1(1+\tau N \log(\bar{A}))}{\sqrt{\tau}} \left(D_2 + \sqrt{\frac{\tau N}{T}} D_3 \right) + \tau N D_3^2,
\end{aligned}$$

where the last inequality follows from the fact that for any two positive scalars x, y, $\sqrt{x+y} \leqslant \sqrt{x} + \sqrt{y}$. Setting τ as per (2.22) ensures that $\tau < \sqrt{\tau}$ as $\tau \leqslant 1$. Thus,

$$\text{Nash-regret}(T) \leqslant \frac{D_1 D_2}{\sqrt{\tau}} + \frac{D_1 D_3 \sqrt{N}}{\sqrt{T}} + \sqrt{\tau} N \left(D_1 D_2 \log(\bar{A}) + D_1 D_3 \log(\bar{A}) \sqrt{\frac{N}{T}} + D_3^2 \right).$$

Plugging in the value of τ as per (2.22) we obtain,

$$\begin{aligned}
\text{Nash-regret}(T) &\leqslant \sqrt{N} \left(D_1^2 D_2^2 \log(\bar{A}) + D_1^2 D_2 D_3 \log(\bar{A}) \sqrt{\frac{N}{T}} + D_1 D_2 D_3^2 \right)^{\frac{1}{2}} + \frac{D_1 D_3 \sqrt{N}}{\sqrt{T}} \\
&\leqslant D_1 D_2 \sqrt{N} \sqrt{\log(\bar{A})} + D_1 \sqrt{D_2 D_3 \log(\bar{A})} \frac{N^{\frac{3}{4}}}{T^{\frac{1}{4}}} + \sqrt{D_1 D_2} D_3 \sqrt{N} + \frac{D_1 D_3 \sqrt{N}}{\sqrt{T}}.
\end{aligned}$$

Note that $D_3 \geqslant 1$ and additionally, we assume that $D_1 \geqslant 1$ (choose large enough C that ensures this). Then,

$$\begin{aligned}
&\text{Nash-regret}(T) \\
&\leqslant D_1 D_2 D_3 \sqrt{N} \sqrt{\log(\bar{A})} + D_1 D_3 \sqrt{D_2 \log(|\bar{A}|)} \frac{N^{\frac{3}{4}}}{T^{\frac{1}{4}}} + \sqrt{D_2} D_1 D_3 \sqrt{N} + \frac{D_1 D_3 \sqrt{N}}{\sqrt{T}} \\
&\leqslant D_1 D_3 \sqrt{N \log(\bar{A})} \left(D_2 + \sqrt{D_2} \left(1 + \left(\frac{N}{T} \right)^{\frac{1}{4}} \right) + \sqrt{\frac{1}{T}} \right) \\
&\leqslant D_1 D_3 \sqrt{\log(\bar{A})} N^{\frac{3}{4}} \mathcal{O}(\max\{D_2, \sqrt{D_2}\}).
\end{aligned}$$

The proof is finished by plugging in D_1, D_2 and D_3.

2.6 Numerical Results

This section studies the empirical performance of Algorithms 1 and 2 for Markov congestion game (MCG) and perturbed Markov team game (PMTG) discussed in Section 2.2. Although Section 2.5 focuses on model-based algorithms, in our numerical study both Algorithm 1 and Algorithm 2 are implemented in a model-free manner, where the Q-functions are estimated from samples [121, 238]. Below are the details for the setup of the experiments.

MCG: Consider MCG with $N = 8$ players, where there are $|E| = 4$ facilities A, B, C, D that each player can select from, i.e., $|A_i| = 4$. For each facility j, there is an associated state s_j: *normal* ($s_j = 0$) or *congested* ($s_j = 1$) state, and the state of the game is $s = (s_j)_{j \in E}$. The reward for each player being at facility k is equal to w_k^{safe} times the number of players at $k = A, B, C, D$. We set $w_A^{\text{safe}} = 1 < w_B^{\text{safe}} = 2 < w_C^{\text{safe}} = 4 < w_D^{\text{safe}} = 6$, i.e., facility D is most preferable by all players. However, if more than $N/2$ players find themselves in the same facility, then this facility transits to the *congested* state, where the reward for each player is reduced by a large constant $c = -100$. To return to the *normal* state, the facility should contain no more than $N/4$ players.

PMTG: Consider a game where each player votes for approving or disapproving a project, which is only conducted if a majority of players vote for approval. The state of excitement about the project changes between different rounds depending on the number of players approving it. Mathematically, consider a game with $N = 16$ players, where there are two actions per player: *approve* ($a_i = 1$) or *disapprove* ($a_i = 0$). There can be two states of the project: *high* ($s = 1$) and *low* ($s = 0$) levels of excitement for the project.

The individual reward of player i is given by $u_i(s, a) = \mathbf{1}_{\sum_i a_i \geqslant N/2} + w_i \mathbf{1}_{\{a_i = s\}} - w_i' a_i$, where the first term represents the common utility derived by everyone if the project is approved, the second term represents the utility derived by a player in approving a high-priority project or disapproving a low-priority project, and the third term corresponds to

the cost of approving the project. Here, we set $w_i = 10\kappa \cdot \frac{N+1-i}{N}$ and $w_i' = \kappa \cdot \frac{i+1}{N}$. Here, parameter κ captures the magnitude of perturbation.

The state transitions from the *high excitement state* to itself with probability λ_1 if more than $N/4$ players approve it; otherwise, it transitions to itself with probability λ_2. In contrast, the state transitions from the *low excitement state* to high with probability λ_3 if there are at least $N/2$ approvers; if there are $N/2$ or fewer approvers, it transitions to high with probability λ_4.

For both games, we perform episodic updates with 20 steps and a discount factor $\gamma = 0.99$. We estimate the Q-functions and the utility functions using the average of mini-batches of size 10. For MCG, Figures 2.2a and 2.2b illustrate the average number of players taking particular action in different states at the converged values of policy. For example, in the state $(0, 0, 0, 1)$ (denoted by the yellow label in Figure 2.2a and 2.2b), facility D is congested, while the other facilities remain in a normal state. In this scenario, only $N/4 = 2$ players select facility D to restore it to a normal state. Simultaneously, $N/2$ players choose facility C, which provides the second-highest reward after D. The number of players at C is within the congestion threshold $(N/2)$, thus ensuring that it remains in a normal state.

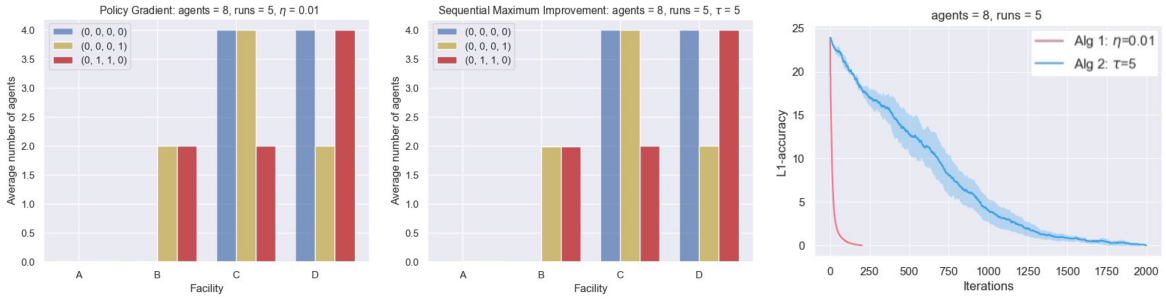
For PMTG, we set $\lambda_1 = \lambda_3 = 1$, $\lambda_2 = \lambda_4 = 0$ and $\kappa = 0.1$. Figures 2.3a and 2.3b illustrate the average number of players taking particular action in different states at the converged values of policy. For example, in the 'high' state of excitement about project (denoted by the red label in Figure 2.3a and 2.3b), almost all players will select to approve as it will always remain in high state thereon. Meanwhile, if the state of excitement is 'low', then at least half of the players select to approve it so that it transitions to 'high' state in future.

Figures 2.2c and 2.3c depict the L_1-accuracy in the policy space at each iteration, defined as the average distance between the current policy and the final policy of all 8 players, i.e., $L_1\text{-accuracy} = \frac{1}{N}\sum_{i\in I} \|\pi_i - \pi_i^{(T)}\|_1$. Figures 2.2c and 2.3c show that Algorithm 1 converges faster for PMTG, while Algorithm 2 converges faster for MCG.

Remark 2.6.1. *We note that the regret bound proposed in our analysis can be loose. In Figure 2.4, we compare growth of regret bound obtained in our theoretical results with that obtained in experiments, where we observe significant gap between the two quantities. This suggests an interesting direction of future research to develop tighter regret bounds.*

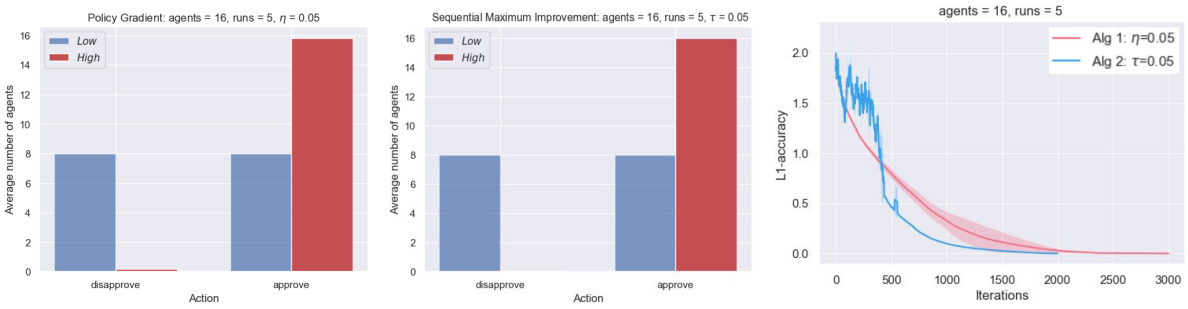
2.7 Concluding Remarks

This chapter introduces a new framework called Markov α-potential games to study Markov games, which generalizes the framework of Markov potential games. We analytically compute upper bounds on α for Markov congestion games and perturbed Markov team games. We also present a semi-infinite linear programming approach to compute an upper bound on α for general discrete-time Markov games. This framework is used to design model-based



(a) (b) (c)

Figure 2.2: **Markov congestion game:** (a) and (b) are distributions of players taking four actions in representative states using $\pi^{(T)}$ given by (a) Algorithm 1 with step-size $\eta = 0.01$; (b) Algorithm 2 with regularizer $\tau_t = 0.999^t \cdot 5$. (c) is mean L1-accuracy with shaded region of one standard deviation over all runs.



(a) (b) (c)

Figure 2.3: **Perturbed Markov team game:** (a) and (b) are distributions of players taking actions in all states: (a) using Algorithm 1 with step-size $\eta = 0.05$; (b) using Algorithm 2 with regularizer $\tau_t = 0.9975^t \cdot 0.05$. (c) is mean L1-accuracy with shaded region of one standard deviation over all runs.

MARL algorithms for Markov games in discrete-time setting, along with associated regret bounds.

The proposed framework opens up new avenues for design and analysis of multi-agent algorithms in dynamic environments. One such example is discussed in Chapter 3, where we use this framework to design high-performance algorithms for autonomous multi-car racing that outperform several existing baselines.

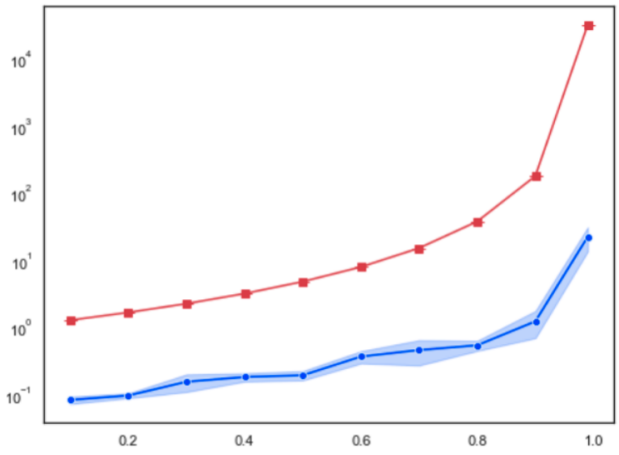


Figure 2.4: Variation of Nash regret with the discount factor for perturbed Markov team game with perturbation parameter $\zeta = 0.1$. The red curve plots the function $1/(1-\gamma)^{9/4}$ (as stated in Theorem 6.1) and the blue shaded region show the Nash regret computed through 10 rounds of experiments with random initialization. Note that the scale on y-axis is in log.

Chapter 3

Competitive Algorithm for Real-time Autonomous Multi-car Racing

In this chapter, we discuss an interesting application of the framework of Markov $\alpha-$potential functions, introduced in Chapter 2. In particular, we use that framework to design real-time algorithms for multi-car autonomous racing.

Autonomous racing is a challenging task in autonomous vehicle development, requiring efficient planning, reasoning, and action in high-speed, dynamic, and constrained environments—key for addressing edge cases in broader autonomous driving. Recent advances, such as deep reinforcement learning (RL), have enabled vehicles to outperform human drivers [433, 209]. However, challenges remain in optimizing strategies against other autonomous agents and in reducing the extensive training times required for achieving competitive performance.

A key challenge in multi-agent autonomous racing is developing real-time competitive strategies that consider the presence of other autonomous vehicles while maintaining high speeds [52]. Algorithms must balance lap-time optimization with aggressive driving, collision avoidance, and dynamic responses to competitors. While existing research on single-agent racing focuses on computing race lines based on vehicle dynamics and track constraints [353, 27, 390, 179, 249, 201], these approaches cannot accommodate complexities of multi-agent settings, where interactions such as blocking and overtaking become crucial. To address these challenges, multi-agent racing requires strategies that account for the interdependent behaviors of agents, with frameworks like Nash equilibrium offering a way to anticipate and adapt to competitors' actions in a competitive environment.

Computation of Nash equilibrium is generally computationally challenging. Several studies have investigated its computation in autonomous racing, but many limitations persist. Many works use kinematic vehicle models [319, 417, 418, 250, 197, 428, 384, 372], which simplify vehicle dynamics but fail to capture the nonlinear tire forces that are critical for high-speed racing. Others rely on open-loop control via receding horizon techniques [76, 200, 417, 418, 250, 197, 372, 454, 455], focusing on finite-horizon planning at the expense of long-term strategies. Additionally, many methods assume two-player zero-sum game [417, 405, 202, 452] or two-team competition [404], which are insufficient for races involving more

than two agents. Alternative approaches, such as Stackelberg games [319, 250], iterative linear quadratic games [372], and local Nash equilibrium [454, 455], provide partial local solutions but do not fully capture the global competitive nature of multi-agent racing.

Towards this end, we propose **α-RACER** (α-potential function based **R**eal-time **A**lgorithm for **C**ompetitiv**E** **R**acing). We study the following key question:

> How to design real-time algorithms for autonomous racing that approximate Nash equilibrium, while accounting for nonlinear tire forces, long-horizon planning and accommodating competitive interaction in any number of competing vehicles?

We highlight three main contributions of this work towards answering the question posed above.

1. Modeling Contribution: We model the multi-agent interaction as an infinite-horizon discounted dynamic game and introduce a novel policy parametrization to enable competitive maneuvers. Specifically, we propose MPC-based policies that track a specially designed, parameterized *reference trajectory*, while avoiding other vehicles. This trajectory is derived by adjusting the optimal single-agent race-line to account for the presence of other agents, enabling competitive racing maneuvers such as overtaking and blocking. Additionally, we structure each agent's immediate utility function to increase its relative progress along the track in comparison to other at each time-step. Moreover, our approach is modular with respect to the vehicle dynamics model. In this work, we use a nonlinear vehicle dynamics model that captures nonlinear tire forces, but we can integrate other sophisticated dynamics. By designing the reference trajectory in this way, we integrate long-term strategic planning for optimizing lap times with tactical planning for effective competition with other cars.

2. Algorithmic Contribution: We present an algorithmic approach to compute Nash equilibrium of the dynamic game, which is inspired by the framework of α-potential functions (developed in Chapter 2). Specifically, we compute an α-potential function that captures the change in each agent's (long-term) value function resulting from unilateral deviations in its policy parameters. The structure of α-potential function allows us to approximate the Nash equilibrium as an optimizer of this function (refer Proposition 3.2.1). We leverage this structure to develop a two-phase algorithmic approach to approximate the Nash equilibrium, consisting of an *offline* and an *online* phase. In the offline phase, we learn an α-potential function using simulated game data. During the online phase, the agent observes the current state of the game and selects the policy parameters that maximize the potential function in that state. This approach enables autonomous agents to engage in competitive maneuvers with reduced computational demands.

3. Numerical Evaluation: We numerically validate the effectiveness of our approach by applying it to three-car autonomous racing scenarios. Our results show that the approxi-

mation gap of our learned potential function is small, enabling us to closely approximate Nash equilibrium strategies. We demonstrate that the maximizer of the potential function effectively captures competitive racing maneuvers. Furthermore, we show that our method outperforms opponent policies—obtained using iterated best response, and finite-horizon self-play reinforcement learning—in most cases.

Additional Related Works.

Several works in the autonomous racing literature use the solution concept of (open-loop) *generalized Nash equilibrium* (e.g., [417, 197]) to incorporate hard collision avoidance constraints, which introduces added computational complexity. In contrast, our modeling framework incorporates collision avoidance in two ways: (i) through the MPC controller used to track the reference trajectory, and (ii) by slowing down vehicles when they enter a pre-defined radius around each other.

Our algorithmic approach also contributes to the growing literature on computational game theory for multi-robot systems [224, 100, 148, 263, 341, 210, 360, 416] by offering a new method to compute approximate Nash equilibrium.

A consolidated summary of comparison with previous literature is provided in Table 3.0.1.

Organization. In Section 3.1, we describe our model of multi-car racing as a dynamic game, including the policy parametrization, utility function, and vehicle dynamics model. In Section 3.2, we present our algorithmic approach. In Section 3.3, we evaluate the performance of our approach using a numerical racing simulator. Finally, we conclude with remarks in Section 3.4.

3.1 Modeling Multi-car Autonomous Racing

In this section, we first present the necessary preliminaries on dynamic game theory, followed by a novel model of multi-car autonomous racing as an instantiation of a dynamic game.

Preliminaries on Dynamic Game Theory

Consider a game involving N strategic players, with the set of players denoted as $I := \{1, 2, \ldots, N\}$. The game proceeds in discrete time steps, indexed by $t \in \mathbb{N}$. At each time step t, the state of player $i \in I$ is represented as $\mathbf{x}_t^i \in \mathcal{X}^i$, where $\mathcal{X}^i$ is the set of all possible states for player i. The joint state of all players at time t is denoted by $\mathbf{x}_t = \left(\mathbf{x}_t^i\right)_{i \in I} \in \mathcal{X}$, where $\mathcal{X} := \times_{i \in I} \mathcal{X}^i$. At every time step t, each player $i \in I$ selects an action $\mathbf{u}_t^i \in \mathcal{U}^i$, where $\mathcal{U}^i$ is the set of feasible actions for player i. The joint action at time t is expressed as $\mathbf{u}_t = \left(\mathbf{u}_t^i\right)_{i \in I} \in \mathcal{U}$, where $\mathcal{U} := \times_{i \in I} \mathcal{U}^i$.

Ref.	Vehicle Model	Game Model	# of cars	Planning
[76]	Dynamic	Known Dynamic Obstacle	2	Local
[200]	Dynamic	Stackelberg Game	$\geqslant 2$	Local
[319]	Kinematic	Zero-sum (IBR)	2	Local
[417]	Kinematic	Zero-sum (IBR) Game	2	Local
[418]	Kinematic	Non-zero sum game	$\geqslant 2$	Local
[250]	Kinematic	Stackelberg/Nash Game	2	Local
[197]	Kinematic	Potential Game	$\geqslant 2$	Local
[428]	Kinematic	IBR	$\geqslant 2$	Local
[384]	Kinematic	Robust RL / Self play	$\geqslant 2$	Local
[372]	Kinematic	iLQG	$\geqslant 2$	Local
[405]	Dynamic	Zero-sum	2	Global
[202]	Dynamic	Zero-sum	2	Global
[454, 455]	Dynamic	General-sum	$\geqslant 2$	Local
Ours	Dynamic	General-sum	$\geqslant 2$	Global

Table 3.0.1: Comparison to previous literature on game theoretic planning for multi-car autonomous racing.

For each player $i \in I$, given the current state $\mathbf{x}_t^i$ and the joint action $\mathbf{u}_t$, the state transitions to a new state at time $t+1$ according to the dynamics $\mathbf{x}_{t+1}^i = f^i(\mathbf{x}_t, \mathbf{u}_t^i)$, where $f^i : \mathcal{X} \times \mathcal{U}^i \to \mathcal{X}^i$ describes the state transition dynamics of player i. Finally, at each time step t, player $i \in I$ receives a reward $r_t^i(\mathbf{x}_t, \mathbf{u}_t, \mathbf{x}_{t+1})$, which depends on the current and next joint state, and joint action of all players. We assume that players select their actions based on a state feedback strategy, denoted by $\pi^i : \mathcal{X} \to \mathcal{U}^i$, such that $\mathbf{u}_t^i = \pi^i(\mathbf{x}_t)$. The set of strategies of player $i \in I$ is denoted by Π^i, and the set of joint strategies by $\Pi = \times_{i \in I} \Pi^i$. In our study, we consider a parameterized set of policies. Specifically, for any player $i \in I$, the player adopts a policy $\pi^i(\cdot; \theta^i) \in \Pi^i$, where $\theta^i \in \Theta^i$ represents the policy parameter. We consider non-myopic players who aim to maximize their long-term discounted utility starting from any state $\mathbf{x} \in \mathcal{X}$. Specifically, each player $i \in I$ selects a parameter $\theta^i \in \Theta^i$ to optimize the discounted infinite-horizon objective,

$$V^i(\mathbf{x}, \theta^i, \theta^{-i}) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t^i(\mathbf{x}_t, \mathbf{u}_t, \mathbf{x}_{t+1})\right],$$

where $\gamma \in [0, 1)$ is the discount factor, $\mathbf{x}_0 = \mathbf{x}$, $\mathbf{u}_t^i = \pi^i(\mathbf{x}_t; \theta^i)$, $\mathbf{u}_t^{-i} = \pi^{-i}(\mathbf{x}_t; \theta^{-i})$[1], and the state transitions according to $\mathbf{x}_{t+1}^i = f^i(\mathbf{x}_t^i, \mathbf{u}_t^i)$.

[1]We use the standard game theory notation of π^{-i} (resp. θ^{-i}) to denote the joint policy (joint policy parameters) of all players except player i.

Definition 3.1.1. *For every* $\mathbf{x} \in \mathcal{X}$*, a joint strategy profile* $\theta^* \in \Theta$ *is called a* ϵ-Nash equilibrium *if, for every* $i \in I, \theta^i \in \Theta^i$*,* $V^i(\mathbf{x}, \theta^{*,i}, \theta^{*,-i}) \geqslant V^i(\mathbf{x}, \theta^i, \theta^{*,-i}) - \epsilon$*. If* $\epsilon = 0$*, the strategy profile* θ^* *is referred to simply as a Nash equilibrium.*

Multi-car Racing as a Dynamic Game

In this subsection, we formulate the multi-car racing problem as a dynamic game by detailing the various components of the game.

Set of Players: Each car is modeled as a strategic player in the dynamic game.

Set of States and Actions: For every car $i \in I$, the state $\mathbf{x}^i = (p_x^i, p_y^i, \phi^i, v_x^i, v_y^i, \omega^i)$, where (p_x^i, p_y^i) denote the longitudinal and lateral position of car i in the *Frenet frame* along the track; ϕ^i denotes the orientation of the car in the Frenet frame along the track; (v_x^i, v_y^i) denote the longitudinal and lateral velocities of car i in the Frenet frame; and ω^i denotes the angular velocity of car i in the body frame. Additionally, $\mathbf{u}^i = (d^i, \delta^i)$, where $\delta^i \in [\delta_{\min}, \delta_{\max}]$ denotes the steering angle of car i and $d^i \in [d_{\min}, d_{\max}]$ is the throttle input of car i.

Dynamics: The most widely used dynamics models in the context of racing are the *kinematic* [197, 319, 420, 418] and *dynamic bicycle models* [203, 70]. In this work, we use the dynamic model as it can accurately model the high-speed maneuvers of the car. A detailed vehicle model description is in Appendix B. On top of the standard dynamic bicycle model, we also incorporate near-collision behavior in our dynamics. Suppose two cars, i and j, are within an unsafe distance from each other and $p_{x,t}^i > p_{x,t}^j$, we reduce their velocities to replicate the time lost due to collision in an actual race with more penalty for the car behind than the one ahead. More concretely, we update the dynamics as $v_{x,t+1}^i = (1/2) \cdot v_{x,t}^i$ and $v_{x,t+1}^j = (1/3) \cdot v_{x,t}^j$. Moreover, if a car goes out of track boundary, i.e., when $|p_{y,t}^i| > w_{\max}/2$ (where $w_{\max}$ is the width of track) we penalize it by reducing its speed and re-align the car along the track. More concretely, we update the dynamics as $v_{x,t+1}^i = v_{x,t}^i/2$ and $\phi_{t+1}^i = 0$.

Policy Parametrization: In this section, we introduce a *novel* policy parametrization $(\Theta^i)_{i \in I}$ designed to capture competitive driving behaviors, such as overtaking, blocking, apex hugging, late braking, and early acceleration. To achieve this, we restrict the set of policies to MPC controllers that track a *reference trajectory* specifically designed to encode such competitive behaviors in multi-car racing. For each car $i \in I$, the parametrization of the MPC controller and its reference trajectory is represented by θ^i, which is characterized by five variables: $(q^i, \zeta^i, s_1^i, s_2^i, s_3^i)$. Before describing these parameters in detail, we present the MPC controller.

At each time step t, for every car $i \in I$, the MPC controller is determined by solving an optimization problem (cf. (3.1)) over a planning horizon of K steps, indexed by k. The

optimization is parameterized by: (a) the longitudinal and lateral positions on the reference trajectory (to be defined shortly), denoted by $(p_{x|t}^{\mathsf{ref},i,k}, p_{y|t}^{\mathsf{ref},i,k})_{k\in[K]}$; (b) the current state of car i at time t, $\mathbf{x}_t^i$; and (c) the longitudinal and lateral positions of the opponent cars located just behind and just ahead of car i. We use the notation $j^*, j_* \in I$ to denote the car in front and the car behind car i at the start of the planning window, based on longitudinal position. For any opponent vehicle $j \in \{j^*, j_*\}$, we assume that the lateral velocity is zero during the planning horizon for the MPC controller, and the longitudinal velocity remains constant at its value at the start of the planning horizon. That is, for every $k \in [K]$, $p_{y|t}^{j,k} = p_{y,t}^j$, and $p_{x|t}^{j,k+1} = p_{x|t}^{j,k} + \Delta t \cdot v_{x,t}^j$. With this setup, we can now describe the MPC optimization problem:

$$\min_{(\mathbf{x}_t^{i,k})_{k=1}^K, (\mathbf{u}_t^{i,k})_{k=0}^{K-1}} \sum_{k=1}^{K} \left\| \begin{pmatrix} p_{x|t}^{i,k} - p_{x|t}^{\mathsf{ref},i,k} \\ p_{y|t}^{i,k} - p_{y|t}^{\mathsf{ref},i,k} \end{pmatrix} \right\|_{\mathbf{Q}}^2 + \sum_{k=1}^{K-1} \left\| \begin{pmatrix} d_t^{i,k} - d_t^{i,k-1} \\ \delta_t^{i,k} - \delta_t^{i,k-1} \end{pmatrix} \right\|_{\mathbf{R}}^2 \tag{3.1a}$$

$$\text{s.t. } \mathbf{x}_t^{i,k+1} = f^i(\mathbf{x}_t^k, \mathbf{u}_t^{i,k}), \quad \forall k = 0, 1, ..., K-1, \tag{3.1b}$$

$$\mathbf{x}_t^{i,0} = \mathbf{x}_t^i, \tag{3.1c}$$

$$\Delta\delta_{\min}^i \leqslant \delta_t^{i,k} - \delta_t^{i,k-1} \leqslant \Delta\delta_{\max}^i, \quad \forall k = 0, 1, ..., K-1, \tag{3.1d}$$

$$d_{\min}^i \leqslant d_t^{i,k} \leqslant d_{\max}^i, \quad \forall k = 0, 1, ..., K-1, \tag{3.1e}$$

$$|p_{y,t}^{i,k}| \leqslant w_{\max}/2, \quad \forall k = 1, ..., K, \tag{3.1f}$$

$$|p_{y|t}^{i,k} - p_{y|t}^{j,k}| \geqslant p_y^{\min}, \quad \forall k = 1, ..., K, \forall j \neq i, \tag{3.1g}$$

$$|p_{x|t}^{i,k} - p_{x|t}^{j,k}| \geqslant p_x^{\min}, \quad \forall k = 1, ..., K, \forall j \neq i, \tag{3.1h}$$

where $\mathbf{x}_t^{i,k} = (p_{x|t}^{i,k}, p_{y|t}^{i,k}, \phi_t^{i,k}, v_{x|t}^{i,k}, v_{y|t}^{i,k}, \omega_t^{i,k})$ is the state of car i in the Frenet frame at the k^{th} step in the planning horizon; $\mathbf{u}_t^{i,k} = (d_t^{i,k}, \delta_t^{i,k})$ is the control input of car i at the k^{th} step in the planning horizon; $w_{\max}$ is the track length, and $p_x^{\min}$ and $p_y^{\min}$ are the minimum required separation between two cars in the longitudinal and lateral directions, respectively; $d_{\min}^i$ and $d_{\max}^i$ are the throttle limits, and $\Delta\delta_{\min}^i$ and $\Delta\delta_{\max}^i$ are the steering rate limits; and $\mathbf{Q}$ and $\mathbf{R}$ are positive definite matrices. Following the MPC approach, the control input at time t is then $\mathbf{u}_t^{i,0}$.

In (3.1), (3.1a) defines the MPC objective, where the first term penalizes the tracking error relative to the reference trajectory, and the second term penalizes variations in the control input. (3.1b) represents the system dynamics constraint, while (3.1c) ensures the planning horizon begins at the car's current state. (3.1d) and (3.1e) enforce constraints on the control inputs, including throttle limits and steering rate bounds. (3.1f) ensures the car stays within the track boundaries, and (3.1g) and (3.1h) enforce the minimum separation between vehicles in the lateral and longitudinal directions, respectively. Next, we describe the policy parameter $\theta^i = (q^i, \zeta^i, s_1^i, s_2^i, s_3^i) \in \mathbb{R}^5$, which parametrizes the policy.

(i) **Parameter q^i:** In (3.1a), we take $\mathbf{R} = \mathbf{I}$ and $\mathbf{Q} = q^i \cdot \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. A higher q^i value results in a more aggressive controller that closely follows the racing line, but it can introduce oscillations that may increase lap time. In contrast, a lower q^i value allows for smoother merging with reduced oscillations, though it may result in larger lateral errors, increased time loss at corners, and a higher risk of track boundary violations.

(ii) **Parameter ζ^i:** We use this parameter to develop a perturbed version of the optimal (single-agent) race-line (see Appendix B for a discussion on race-line), which is generated by sampling points along the optimal racing line at time intervals of Δt. More formally, let the optimal race-line be denoted by $\mathbf{x}^{\mathsf{rl},i}$. Given the current state $\mathbf{x}_t^i$, we find the closest point on the race-line and consider a race-line starting at time t, denoted by $\bar{\mathbf{x}}_t^{\mathsf{rl},i}$. Using this, we compute a trajectory of length K, denoted by $(p_{x|t}^{\mathsf{pert},i,k}, p_{y|t}^{\mathsf{pert},i,k})_{k\in[K]}$. Specifically, for every $k \in [K]$, we construct:

$$p_{x|t}^{\mathsf{pert},i,k} = p_{x|t}^{\mathsf{pert},i,k-1} + v_{x|t}^{\mathsf{pert},i,k-1} \cdot \Delta t, \quad p_{y|t}^{\mathsf{pert},i,k} = p_{y|t}^{\mathsf{pert},i,k-1} + v_{y|t}^{\mathsf{pert},i,k-1} \cdot \Delta t,$$

where

$$v_{x|t}^{\mathsf{pert},i,k-1} = \zeta^i \bar{v}_{x,t+k-1}^{\mathsf{rl},i}, \quad \text{and} \quad v_{y,t}^{\mathsf{pert},i,k-1} = \zeta^i \bar{v}_{y,t+k-1}^{\mathsf{rl},i}$$

are the perturbed race-line velocities. Higher values of ζ^i capture how aggressively the vehicle wants to follow the optimal race-line.

(iii) **Parameters s_1, s_2, s_3:** The reference trajectory, $(p_{x|t}^{\mathsf{ref},i,k}, p_{y|t}^{\mathsf{ref},i,k})$, is generated by modifying the perturbed race-line $(p_{x|t}^{\mathsf{pert},i,k}, p_{y|t}^{\mathsf{pert},i,k})_{k\in[K]}$ by accounting for the positions and velocities of the cars immediately ahead (in terms of longitudinal coordinates) of the ego car and immediately behind (in terms of longitudinal coordinates) ego car. Let's denote the ego car by i, the car immediately ahead of this car by j^*, and the one immediately behind by j_*. We define the reference trajectory as follows

$$p_{y|t}^{\mathsf{ref},i,k} = \mathsf{clip}(p_{y|t}^{\mathsf{pert},i,k} + p_{y|t}^{\mathsf{ot},i,k} + p_{y|t}^{\mathsf{bl},i,k}, [-w_{\max}, w_{\max}]),$$

where

$$\begin{aligned} p_{y|t}^{\mathsf{ot},i,k} = {} & \mathsf{sign}(p_{y,t}^i - p_{y,t}^{j^*}) \max\left\{(s_1 - |(p_{y,t}^i - p_{y,t}^{j^*})|) \exp\left(-s_2\left(\Delta p_{x|t}^{i,j^*,k}\right)^2\right), 0\right\} \\ & + \mathsf{sign}(p_{y,t}^i - p_{y,t}^{j_*}) \max\left\{(s_1 - |(p_{y,t}^i - p_{y,t}^{j_*})|) \exp\left(-s_2\left(\Delta p_{x|t}^{i,j_*,k}\right)^2\right), 0\right\}, \end{aligned}$$

is an adjustment for overtaking that smoothly changes the trajectory of the ego vehicle opposite the leading vehicle to overtake. Here, for any $j \in \{j^*, j_*\}$, $\Delta p_{x|t}^{i,j,k} = p_{x|t}^{\mathsf{pert},i,k} - p_{x|t}^{j,k}$, and

$$\begin{aligned} \bar{p}_{y|t}^{\mathsf{bl},i,k} = {} & \mathbb{1}(v_{x|t}^{\mathsf{pert},i,k} \leqslant v_{x,t}^{j_*}) \mathbb{1}(p_{x|t}^{\mathsf{pert},i,k} \geqslant p_{x|t}^{j_*,k}) h(p_{y|t}^{j_*,k}, p_{x|t}^{\mathsf{pert},i,k}, v_{x|t}^{\mathsf{pert},i,k}, v_{x,t}^{j_*}, \Delta p_{x|t}^{i,j_*,k}) \\ & + \mathbb{1}(v_{x|t}^{\mathsf{pert},i,k} \leqslant v_{x,t}^{j^*}) \mathbb{1}(p_{x|t}^{\mathsf{pert},i,k} \geqslant p_{x|t}^{j^*,k}) h(p_{y|t}^{j^*,k}, p_{x|t}^{\mathsf{pert},i,k}, v_{x|t}^{\mathsf{pert},i,k}, v_{x,t}^{j^*}, \Delta p_{x|t}^{i,j^*,k}) \end{aligned}$$

is the adjustment for blocking that smoothly changes the trajectory of the ego vehicle towards trailing vehicle to block it. Here, for any $j \in \{j^*, j_*\}$,

$$h(p^{j,k}_{y|t}, p^{\mathsf{pert},i,k}_{x|t}, v^{\mathsf{pert},i,k}_{x|t}, v^{j}_{x,t}, \Delta p^{i,j,k}_{x|t}) \\ = (p^{j,k}_{y|t} - p^{\mathsf{pert},i,k}_{y|t})(1 - \exp(-s_3(v^{\mathsf{pert},i,k}_{x|t} - v^{j}_{x,t}))) \exp(-s_2(\Delta p^{i,j,k}_{x|t})^2).$$

One-step Utility Function: We consider that the one-step utility for every car is to maximize progress along track:

$$\mathbb{R} \ni r^i(\mathbf{x}_t, \mathbf{u}_t, \mathbf{x}_{t+1}) = (p^i_{x,t+1} - \max_{j \neq i} p^j_{x,t+1}) - (p^i_{x,t} - \max_{j \neq i} p^j_{x,t}).$$

3.2 Approximating multi-agent interactions

In this section, we provide a tractable approach to compute an approximate Nash equilibrium for the racing game described in Section 3.1. Core to our approach is the framework of α-potential functions, recently introduced in [171, 276, 170].

α-Potential Function

Definition 3.2.1. *A potential function* $\Phi : \mathcal{X} \times \Theta \to \mathbb{R}$ *is called a* dynamic α-potential function[2] *with* approximation parameter α *if for every* $\mathbf{x} \in \mathcal{X}, i \in I, \theta \in \Theta, \theta^{i\prime} \in \Theta^i$,

$$|(\Phi(\mathbf{x}, \theta^i, \theta^{-i}) - \Phi(\mathbf{x}, \theta^{i\prime}, \theta^{-i})) - (V^i(\mathbf{x}, \theta^i, \theta^{-i}) - V^i(\mathbf{x}, \theta^{i\prime}, \theta^{-i}))| \leqslant \alpha. \tag{3.2}$$

This definition intuitively requires that for any agent, the change in its value function resulting from a unilateral adjustment to its policy parameter can be closely approximated by the corresponding change in the dynamic α-potential function. This property allows us to approximate the Nash equilibrium as an optimizer of the near-potential function.

Proposition 3.2.1. *Given an α-potential function Φ, for any $\mathbf{x} \in \mathcal{X}$, $\lambda > 0$, and any policy θ^* satisfying $\Phi(\mathbf{x}, \theta^*) \geqslant \max_{\theta \in \Theta} \Phi(\mathbf{x}, \theta) - \lambda$, the policy θ^* constitutes a $(\lambda + \alpha)$-approximate Nash equilibrium.*

Proof. Consider a policy parameter θ^* such that $\Phi(\mathbf{x}, \theta^*) \geqslant \max_{\theta \in \Theta} \Phi(\mathbf{x}, \theta) - \lambda$. For any $\theta^i \in \Theta^i$,

$$V^i(\mathbf{x}, \theta^{*,i}, \theta^{*,-i}) - V^i(\mathbf{x}, \theta^i, \theta^{*,-i}) \geqslant \Phi(\mathbf{x}, \theta^{*,i}, \theta^{*,-i}) - \Phi(\mathbf{x}, \theta^i, \theta^{*,-i}) - \alpha \\ \geqslant -\lambda - \alpha,$$

where first inequality is due to (3.2) and the second inequality is because θ^* maximizes Φ. The proof follows using the definition of Nash equilibrium (Definition 3.1.1). □

[2]For convenience, we adopt a slightly different definition of the α-potential function than that in [171, 276]; however, the results of this work extend to the definitions used there.

Computational Approach

Our approach for real-time approximation of Nash equilibrium relies on two phases: offline and online. In the offline phase, we learn an α-potential function using simulated game data. In the online phase, the ego vehicle updates its policy parameters by optimizing the potential function.

Offline Phase: We parameterize the potential function through using a feed-forward neural network with ReLU activation and a BatchNorm layer added at the beginning. More concretely, we define the parametrized potential function as $\Phi(\cdot;\phi) : \mathcal{X} \times A \to \mathbb{R}$, where ϕ denotes the weights of neural network. Using Definition 3.2.1, we cast the problem of learning potential function as a semi-infinite program as shown below:

$$\begin{aligned} \min_{y,\phi} \quad & y \\ \text{s.t.} \quad & \Big|(\Phi(\mathbf{x},\theta^i,\theta^{-i};\phi) - \Phi(\mathbf{x},\theta^{i\prime},\theta^{-i};\phi)) - (V^i(\mathbf{x},\theta^i,\theta^{-i};\upsilon) - V^i(\mathbf{x},\theta^{i\prime},\theta^{-i};\upsilon))\Big| \leqslant y, \\ & \quad \forall i \in I,\ \forall \theta^i,\theta^{i\prime} \in \Theta^i,\ \forall \theta^{-i} \in \Theta^{-i}, \mathbf{x} \in \mathcal{X}, \end{aligned} \tag{3.3}$$

where we use a neural network (same architecture as potential function), with parameters υ, to estimate the value function V^i for every $i \in I$. Let (y^*,ϕ^*) be a solution of the above optimization problem. The main challenge with solving (3.3) is that the there are un-countably many constraints, one constraint corresponding to each value of initial state and policy parameter. Therefore, we numerically solve (3.3) by using simulated game data with randomly chosen starting position and policy parameters. Details of simulated game are discussed in next section.

Online Phase: Leveraging Proposition 3.2.1, the ego vehicle optimizes the learned potential function, i.e. $\Phi(\cdot,\phi^*)$, to approximate the Nash equilibrium policy parameter. More formally, given the current state $\mathbf{x}_t$, the ego vehicle computes $\theta^* \in \arg\max_{\theta\in\Theta} \Phi(\mathbf{x}_t,\theta;\phi^*)$, using a non-linear optimization solver. Using θ^*, the ego vehicle takes action $\mathbf{u}_t^i = \pi^i(\mathbf{x}_t;\theta^{*,i})$.

3.3 Numerical Evaluation

Here, we evaluate our approach on a numerical simulator by focusing on three questions:

(Q1) Can our approach closely approximate Nash equilibrium and generate competitive behavior?

(Q2) How do hyper-parameters like the discount factor γ and the amount of data used to learn the α-potential function affect the performance?

(Q3) How does our approach compare against common baselines?

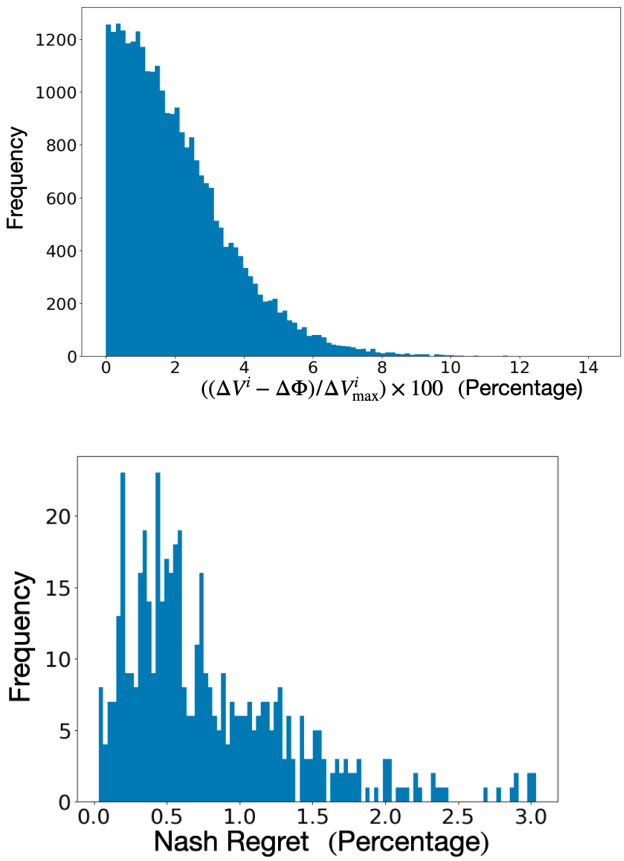


Figure 3.1: Histogram of (a) Relative approximation gap of potential function (b) Nash regret

Experimental Setup: We generate a dataset of 4000 races, each lasting 50 seconds, conducted with randomly chosen policy parameters and involving 3 cars. This dataset is used to first train value function estimators V^1, V^2, and V^3 for each of the cars. These are then used to learn Φ using (3.3). To maximize the learned potential function, we use gradient ascent with a learning rate of 10^{-4} and warm-start by using the solution from the previous time step.

Competitive Behavior by Approximating Nash Equilibrium

We observe that the approximation gap of the learned potential function is small. As shown in Figure 3.1(a), the approximation gap across all states and policy parameters used in the training samples remains within 10% of the value function's range, with a median gap of approximately 2%. Next, we demonstrate that the optimization solver effectively computes the maximizer of the potential function, leading to a lower Nash approximation

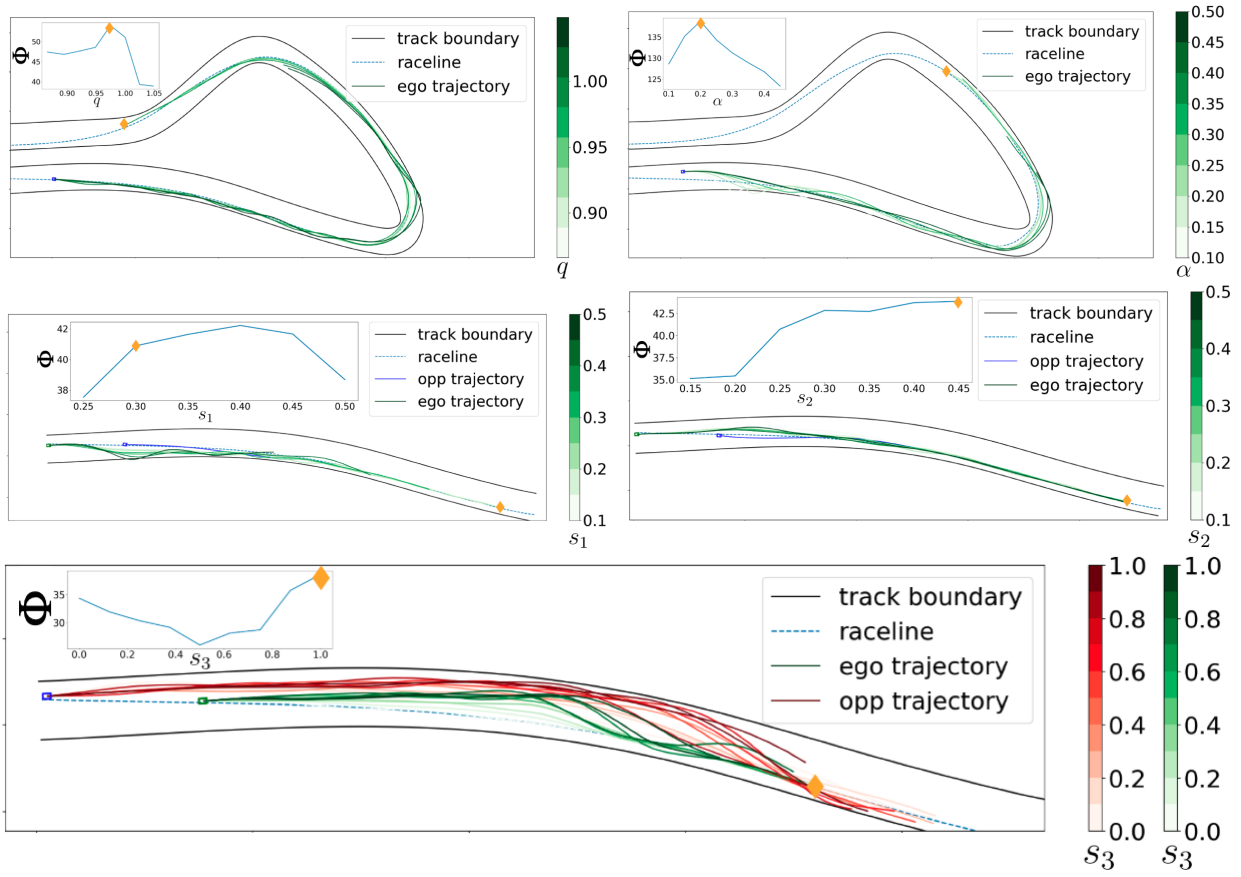


Figure 3.2: Potential values and the trajectories at a given joint state for different (a) q (b) ζ (c) s_1 (d) s_2 (e) s_3 of only the ego agent. We only denote 2 players here (only 1 player for (a) and (b)) and the 3rd player is far away from this position to not affect any players. Additionally, for ease of readability, we only show the impact of variation in trajectory of other player in response to ego in (e) as such deviations are not significant in (c) and (d).

error. In particular, Figure 3.1(b) shows the *Nash regret* for the ego agent, defined as $\max_{\theta^i} V^i(\mathbf{x}, \theta^i, \theta^{*,-i}) - V^i(\mathbf{x}, \theta^{*,i}, \theta^{*,-i})$, where θ^* is the optimizer of the potential function with the starting state $\mathbf{x}$. The regret is plotted for different game states during a race, and we observe that it remains within 3% of the range of the value function. In summary, Figure 3.1 highlights that the dynamic game admits an α-potential function with small α, and that we accurately compute a near-optimal solution to the potential function.

Next, in Figure 3.2, we show that, by fixing all policy parameters but one, the parameter that (approx.) maximizes potential function generates a trajectory that maximizes progress along the track over the next 5 seconds. This is highlighted by yellow diamond in Figure 3.2.

Performance Comparison

To address questions **Q2** and **Q3**, we conduct 99 races involving three agents: ego, O_1, and O_2. Here, ego represents the agent using our proposed algorithm, while O_1 and O_2 are opponents employing other algorithms. In this study, we vary the opponents' algorithms and compute the winning fraction for the ego car. A visualization of the track used in the study, along with the computed optimal race-line, is provided in the Appendix B (cf. Figure B.1). The starting positions for the races are taken from three regions (denoted by R_1, R_2, and R_3), such that R_1 is the furthest ahead on the track, followed by R_2, and then R_3, as shown in Figure B.1. We perform 33 races, each with the ego agent starting in R_1, R_2, or R_3. Also, O_1 is always placed ahead of O_2.

For the ego, we use a discount factor $\gamma = 0.99$ and a training dataset comprising 4000 races. Below, we summarize five different opponent strategies for O_1 and O_2:
(i) **Case I:** Opponents trained with a lower discount factor than ego (i.e., $\gamma = 0.98$);
(ii) **Case II:** Opponents trained with a higher discount factor than ego (i.e., $\gamma = 0.995$);
(iii) **Case III:** Opponents trained using fewer simulated races than ego (i.e., 400 races);
(iv) **Case IV:** Opponents use the *Iterated Best Response* (IBR) algorithm [3], which computes the best response of opponents in round-robin for a fixed number of rounds (i.e., 6), with a planning horizon[4] of length $2s$;
(v) **Case V:** Opponents trained using self-play RL[5]. We use a similar observation and reward as used in [405] and train with $2M$ steps.
The number of races won for all cases are provided in Table 3.3.1, where we see that ego agent trained using our approach has superior performance in comparison to all other opponent strategies.

Performance Variation with Hyperparameters (Cases I–III). The ego car outperforms opponents with lower discount factors (**Case I**) because lower discount factors lead to more myopic behavior, causing opponents to prioritize short-term progress over long-term track performance. Similarly, the ego car also outperforms opponents with higher discount factors (**Case II**) as higher discount factors increase the "effective horizon" of the game, which requires significantly more data to accurately approximate the value and potential functions. Furthermore, the ego car has superior performance against opponents trained with less data of only 400 races (**Case III**), as more data enables the learning of a more accurate α-potential function.

Comparison with opponents using IBR (Case IV). Our approach surpasses IBR for two key reasons. First, IBR may not always converge to a Nash equilibrium. Second, IBR's real-time computation is often restricted to local planning with a short $2s$ horizon, as noted

[3]The IBR algorithm used here is representative of the methods developed in [396, 421], though it may not be an exact implementation, as the original code is not publicly available.

[4]Choice of hyperparameters is such that it roughly takes the same compute time as our approach.

[5]Here, self-play RL represents the approach in [405], excluding the high-level tactical planner.

Racing Scenario	# wins (ours)	# wins ($\mathbf{O}_1$)	# wins ($\mathbf{O}_2$)
Case I (Opponents with low γ)	61	28	10
Case II (Opponents with high γ)	52	40	7
Case III (Opponents trained with less data)	76	16	7
Case IV (Opponents using IBR)	73	22	4
Case V (Opponents trained using self-play RL)	91	7	1

Table 3.3.1: Outcomes of 99 races conducted between 3 cars under three different initial positions

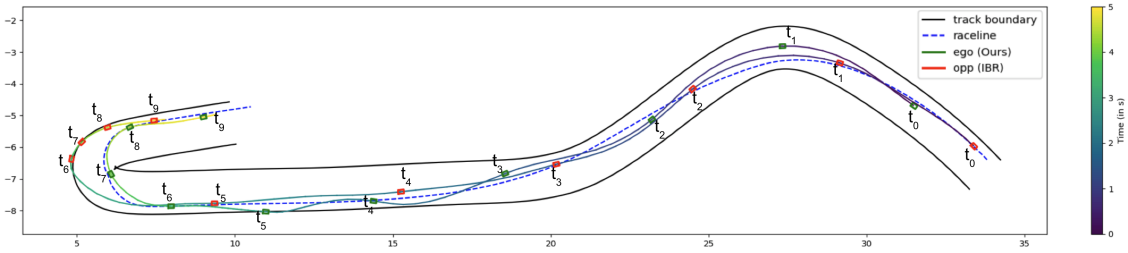


Figure 3.3: Example overtake in a race MPC vs IBR. Opponent (IBR) overtakes from t_4 to t_5 but later suffers at the turn from t_5 to t_9 when the Ego agent (Ours) overtakes back to re-claim it's position

in [396, 417]. In contrast, competitive racing demands prioritizing a global racing line to optimize long-term performance, with strategic deviations for overtaking or blocking rather than short-term gains. This distinction is highlighted in Figure 3.3, where the IBR player achieves higher straight-line speeds to overtake but struggles in turns due to aggressive braking. While our implementation is not a direct comparison with [396], our approach supports longer-horizon planning for real-time control by leveraging offline potential function learning. This is achieved by constraining the policy space to primitive behaviors, closely aligning with practical racing strategies.

Comparison with opponents using self-play RL (Case V). Our approach also outperforms opponents trained via self-play RL. While we do not claim an optimized implementation of self-play RL—acknowledging potential improvements through further training, hyperparameter tuning, or tactical enhancements as in [405]—our method offers clear advantages. It delivers interpretable solutions grounded in realistic racing strategies and approximation guarantees to a Nash equilibrium in races with more than 2 cars. In contrast, self-play RL often requires additional insights to develop effective policies. For example, [405] showed that augmenting self-play RL with high-level tactical plans significantly enhances policy learning.

3.4 Concluding Remarks

In this chapter, we study real-time algorithms for competitive multi-car autonomous racing. Our approach is built on two key contributions: first, a novel policy parametrization and utility function that effectively capture competitive racing behavior; second, the use of dynamic α-potential functions to develop real-time algorithms that approximate Nash equilibrium strategies. This framework enables the learning of equilibrium strategies over long horizons at different game states through the maximization of a potential function.

Chapter 4

Decentralized Learning in Markov Potential Games

As highlighted through Chapters 2-3, Markov games are a useful framework for modeling the strategic interactions among multiple self-interested players in a dynamic environment. This framework has been adopted to study many important applications that include autonomous driving [380], adaptive traffic control [344, 38], e-commerce [221], and AI training in real-time strategy games [414, 68]. In a lot of these applications, players get to interact with one another over long periods of time. Typically, these players interact in an independent and decentralized manner, adapting to the information received through interactions in uncertain and dynamic environments. Coordination and communication may be absent, and players might not even be aware of the existence of others. In such an environment, a natural approach for each agent is to adopt a single-agent reinforcement learning algorithm, which uses only the information from the observed state, each player's individual action, and bandit feedback about their own payoff in each stage. Such learning dynamics are fully decentralized and independent, meaning that each agent updates their own policy as if they are the sole decision-maker in the environment even though they are playing a Markov game.

A canonical example of such decentralized interactions arises in ride-hailing platforms, where drivers must decide where to position themselves, when to go online, and which requests to accept, all while lacking coordination or knowledge of other drivers' strategies. Each driver learns independently based on their own experience—such as location, travel time, and earnings—resulting in emergent behavior that can be modeled as a Markov game with decentralized learning dynamics.

Another example appears in electricity markets with distributed energy resources such as prosumers or energy storage operators, where each agent decides when to consume, store, or sell energy based on local information like price signals, battery levels, and weather forecasts. These decisions impact and are influenced by grid-wide conditions, yet coordination among agents is typically absent. Each agent may apply reinforcement learning to maximize their long-term return, treating the environment as stationary while in fact it is shaped by all agents' actions—again, naturally modeled as a Markov game.

Motivated by this requirement, we study the following important question in this Chapter and the next (i.e. Chapter 5):

> *What is the long run outcome of interactions among players who update their strategies in an independent and decentralized manner using RL algorithms?*

In this chapter, we study the above question in the context of Markov Potential Games (MPGs) ([238, 450, 391, 284, 121, 144, 449, 286, 410]). Recall from Chapter 2, in MPGs the change of utility of any player from unilaterally deviating their policy can be evaluated by the change of the potential function. MPGs can be used to study Markov team games (also known as common interest games) [17], some variant of congestion games [144], and dynamic demand-response in energy marketplaces [313]. Previous studies of MPGs have mainly focused on analyzing convergence properties of gradient-based algorithms in both discounted infinite horizon settings [450, 238, 121, 144] and finite horizon episodic settings [284, 391]. However, the evaluation of gradient (or its estimate) of any player's value function requires players to either have access to a simulator/oracle of the value function or to coordinate and communicate their strategies and payoffs with each other [238, 144, 117, 121]. Such communication and coordination may be restricted in many real-world multi-agent systems due to communication constraints or privacy concerns [221, 320, 239].

We focus on the learning dynamics, where each agent independently and decentralizely adopts an actor-critic algorithm [214] with asynchronous step sizes. In our setting, players do not know the existence of other players participating in the game, and do not have knowledge of state transition probabilities, their own payoff functions or the opponents' payoff functions. Additionally, players do not have access to any information about the potential function or its existence. Each player *only* observes the realized state, and their own realized payoff in each stage. In particular, the multi-agent actor-critic learning dynamics considered in this chapter has the following key features:

(i) The dynamics have *two timescales*: each player updates the q-estimate of their contingent payoff (represented as the Q-function defined in Sec. 4.1) at a faster timescale, and update their policies at a slower timescale.

(ii) Players are *self-interested* in that their updated policy incorporates an *optimal one-stage deviation* that maximizes the expected contingent payoff derived from the current q-estimate.

(iii) Learning is *asynchronous* and *heterogeneous* among players. In every stage, only the q-estimate of the realized state-action pair, and the policy corresponding to the realized state are updated. The remaining elements in q-estimate and policy remain unchanged. Furthermore, the stepsizes of updating the element correspond to each state and action are heterogeneous across players, and are asynchronously adjusted according to the number of times a state and that player's action are realized.

(iv) In each stage, players generate their actions by combining their updated policy with a uniform randomization (exploration) of all of their actions in order to learn the Q-function across all states. The exploration probability can be heterogeneous across players.

Independent and decentralized learning algorithms often do not converge [403, 289, 291]. We develop a new approach to characterize the convergent *set* our learning dynamics. Our approach involves non-trivial extensions of the analysis of single-agent reinforcement learning to MPG, and developing game-theoretic tools to utilize the equilibrium condition of the underlying Markov game. Specifically, we study the asymptotic behavior of discrete-time dynamics using an associated continuous-time dynamical system by exploiting the timescale separation between the updates of q-estimate and policy [60, 408, 337]. In this dynamical system, the fast dynamics – update of the q-estimates – can be analyzed while viewing the policy updates (the slow dynamics) as static, and thus the q-estimate of each player converges to their Q-function (Lemma 4.2.1). Importantly, we show that, for every $\epsilon > 0$, the potential function always increases along the trajectory of the continuous-time dynamical system outside a set of ϵ-stationary Nash equilibrium, given that the total exploration probability of all players is bounded by $L\epsilon$, where $L > 0$ is a game dependent parameter (Lemma 4.2.2). This key lemma allows us to characterize the convergent set of policies. Particularly, we show that the trajectories converge to the smallest super-level set of potential function that contains the ϵ-Nash equilibrium set (Theorem 4.2.1). Furthermore, under additional assumption on the potential function and Nash equilibrium set, we establish convergence to the set of ϵ-stationary Nash equilibrium (Corollary 4.2.1). Finally, we validate the performance of our algorithm on a numerical example.

Additional Related Works

Apart from learning in MPGs, another line of work in multi-agent reinforcement learning focuses on the fully competitive setting of Markov zero-sum games [117, 368, 367, 8, 168, 339]. Most articles in this line of works require players to either observe the opponents' rewards or actions [8, 368, 369], or to coordinate in policy updates [117, 168]. The paper [367] proposed an independent and decentralized learning dynamics, and showed its convergence in Markov zero-sum games. The algorithm in [367] also has timescale separation between policy update and value update, but is in reversed ordering compared to ours. That is, their dynamics update value function at a slower timescale, and update policies at a faster timescale, while our policy update is slower compared to the q-estimate update. Another difference is that our learning dynamics adopts a different stepsize adjustment procedure that allows players to update their step-sizes based on their own counters of states and actions heterogeneously. We emphasize that the convergence analysis in this chapter is different from that in [367] due to the differences in the two learning algorithms and the inherent difference between

Markov zero-sum games and Markov potential games.[1]

Two-timescale based algorithms have also been studied in other non-zero sum games [58, 338, 345, 17, 239]. Specifically, [239] studied the two-timescale based algorithm for static games. The paper [58] proposed an actor-critic algorithm, and showed that certain weighted empirical distribution of realized actions converges to a *generalized Nash equilibrium*. In [17], the authors presented an algorithm in the setting of *acyclic Markov games*, which subsume Markov team games. However, the proposed algorithm require coordination amongst players. The paper [345, 338] proposed actor-critic algorithms with a *fast* value function update – based on temporal difference learning – and a *slow* policy update. In [345], the gradient-based policy update requires the knowledge of opponents' rewards. The paper [338] adopted a best-response based policy update that is similar to our learning dynamics, and proved its convergence in multistage games, which are a class of generalized normal form games with tree structures. Our algorithm is different from the one in [338]. First, we consider a uniform exploration policy which is different from [338] where the authors consider perturbed or smoothed best response (similar to [239]). Second, we consider updates with asynchronous step sizes that are adjusted based on counters of each state and each state-action pair while [338] considers homogeneous step sizes. The proof technique developed in [338] exploits the special tree structure of multistage games, they cannot be applied in our setting. Additionally, [440] proposes a two-loop algorithm where the policies are updated in the *outer-loop* and the value functions are updated in the *inner-loop*. Between any two updates of outer-loop, the algorithm makes multiple rounds of update of the inner-loop. The role of two loop algorithms here is to ensure stationarity in the learning environment by fixing the policy updates while learning the value function. This algorithm requires coordination among agents to decide the length of the inner loop.

Finally, our results also advance the rich literature of learning in stateless potential game that includes continuous and discrete time best response dynamics [306, 401], fictitious play [305, 183, 287], replicator dynamics [328, 184], no-regret learning [178, 218], and payoff-based learning [105, 239]. In particular, our learning dynamics share similar spirit with the payoff-based learning dynamics in stateless potential games [105, 239]. In payoff-based learning, players update their payoff estimates based only on their own payoffs and adjust their mixed strategy using a best response. In MPGs, the payoff estimates of different state-action pairs are updated asynchronously, and the best response becomes an optimal one-stage deviation policy. Therefore, our result is not a direct extension of stateless potential games as it involves using reinforcement learning tools to study long-run behavior.

Outline

Section 4.1 presents Markov potential games. We present our independent and decentralized learning dynamics, and the convergence results in Section 4.2, validate the performance of

[1]Our convergence proof builds on the existence of potential function and the convergence of fast q-learning. On the other hand, the proof of convergence in zero-sum Markov games depends on the Shapley iteration convergence.

algorithm numerically in Section 4.3, and conclude our work in Section 4.4. We include the proofs of technical lemmas in the appendix.

4.1 Model

In this chapter, we consider the same setup of Markov game as in Section 2.1. We review the setup again here for completeness.

We define the Markov game by tuple $\mathcal{G} = \langle I, S, (A_i)_{i\in I}, (u_i)_{i\in I}, P, \gamma\rangle$, where I is a finite set of players; S is a finite set of states; A_i is a finite set of actions with generic member a_i for each player $i \in I$, and $a = (a_i)_{i\in I} \in A = \times_{i\in I} A_i$ is the action profile of all players; $u_i(s, a) : S \times A \to \mathbb{R}$ is the one-stage payoff of player i with state $s \in S$, and action profile $a \in A$; We define $u_{\max} = \max_{i,s,a} |u_i(s, a)|$; $P = (P(s'|s, a))_{s,s'\in S,a\in A}$ is the state transition matrix and $P(s'|s, a)$ is the probability that state changes from s to s' with action profile a; and $\gamma \in (0, 1)$ is the discount factor.

We denote *a stationary Markov policy* $\pi_i = (\pi_i(s, a_i))_{s\in S,a_i\in A_i} \in \Pi_i = \Delta(A_i)^{|S|}$, where $\pi_i(s, a_i)$ is the probability that player i chooses action a_i given state s. For each $i \in I$ and each $s \in S$, we denote $\pi_i(s) = (\pi_i(s, a_i))_{a_i\in A_i}$. The joint policy profile is denoted as $\pi = (\pi_i)_{i\in I} \in \Pi = \times_{i\in I}\Pi_i$. We also use the notation $\pi_{-i} = (\pi_j)_{j\in I\setminus\{i\}} \in \Pi_{-i} = \times_{j\in I\setminus\{i\}}\Pi_j$ to refer to the joint policy of all players except for player i. For concise presentation, we will use the following notation throughout the chapter:

$$u_i(s, a_i, \pi_{-i}) = \sum_{a_{-i}} \pi_{-i}(s, a_{-i}) u_i(s, a_i, a_{-i}),$$

$$P(s'|s, a_i, \pi_{-i}) = \sum_{a_{-i}} \pi_{-i}(s, a_{-i}) P(s'|s, a_i, a_{-i}),$$

and

$$P(s'|s, \pi) = \sum_{a_i} \sum_{a_{-i}} \pi_i(s, a_i)\pi_{-i}(s, a_{-i}) P(s'|s, a_i, a_{-i}).$$

The game proceeds in discrete-time stages indexed by $k = \{0, 1, ...\}$. At $k = 0$, the initial state s^0 is sampled from a distribution μ. At every time step k, given the state s^k, each player's action $a_i^k \in A_i$ is sampled from the policy $\pi_i(s^k)$, and the joint action profile is $a^k = (a_i^k)_{i\in I}$. The state transitions to $s^{k+1} \sim P(\cdot|s^k, a^k)$ based on the current state s^k and action profile a^k. Given an initial state distribution μ, and a stationary policy profile π, the expected total discounted payoff of each player $i \in I$ is given by:

$$V_i(\mu, \pi) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k u_i(s^k, a^k)\right], \tag{4.1}$$

where $s^0 \sim \mu$, $a^k \sim \pi(s^k)$, and $s^k \sim P(\cdot|s^{k-1}, a^{k-1})$. For the rest of the chapter, with slight abuse of notation, we use $V_i(s, \pi)$ to denote the expected total utility of player i when the initial state is a fixed state $s \in S$. Thus, we have $V_i(\mu, \pi) = \sum_{s\in S} \mu(s) V_i(s, \pi)$.

Definition 4.1.1 (Markov potential games (cf. Definition 2.3.1)). *A Markov game $\mathcal{G}$ is a Markov potential game (MPG) if there exists a state-dependent potential function $\Phi : S \times \Pi \to \mathbb{R}$ such that for every $s \in S, i \in I$, $\pi_i, \pi_i' \in \Pi_i$, $\pi_{-i} \in \Pi_{-i}$,*

$$\Phi(s, \pi_i', \pi_{-i}) - \Phi(s, \pi_i, \pi_{-i}) = V_i(s, \pi_i', \pi_{-i}) - V_i(s, \pi_i, \pi_{-i}).$$

Moreover, given an initial state distribution $\mu \in \Delta(S)$, the potential function $\Phi(\mu, \pi) := \sum_{s \in S} \mu(s)\Phi(s, \pi)$ satisfies

$$\Phi(\mu, \pi_i', \pi_{-i}) - \Phi(\mu, \pi_i, \pi_{-i}) = V_i(\mu, \pi_i', \pi_{-i}) - V_i(\mu, \pi_i, \pi_{-i}),$$

for every $i \in I$, $\pi_i, \pi_i' \in \Pi_i$, $\pi_{-i} \in \Pi_{-i}$.

That is, in a MPG, the change of a single deviator's value function can be characterized by the change of the value of the potential function.

We next present the definition of stationary Nash equilibrium, and ϵ-stationary Nash equilibrium.

Definition 4.1.2 (Stationary Nash equilibrium policy). *A policy profile π^* is a stationary Nash equilibrium of $\mathcal{G}$ if for any $i \in I$, any $\pi_i \in \Pi_i$, and any $\mu \in \Delta(S)$, $V_i(\mu, \pi_i^*, \pi_{-i}^*) \geqslant V_i(\mu, \pi_i, \pi_{-i}^*)$.*

Definition 4.1.3 (ϵ-Stationary Nash equilibrium policy). *For any $\epsilon \geqslant 0$, a policy profile π^* is an ϵ-stationary Nash equilibrium of $\mathcal{G}$ if for any $i \in I$, any $\pi_i \in \Pi_i$, and any $\mu \in \Delta(S)$, $V_i(\mu, \pi_i^*, \pi_{-i}^*) \geqslant V_i(\mu, \pi_i, \pi_{-i}^*) - \epsilon$. For any $\epsilon \geqslant 0$, we define the set of all ϵ-stationary Nash equilibrium as $\mathsf{NE}(\epsilon)$. Any ϵ-stationary Nash equilibrium with $\epsilon = 0$ is a Nash equilibrium.*

Both stationary Nash equilibrium and ϵ-stationary Nash equilibrium exist in Markov games with finite states and actions [150]. In a MPG, if there exists a policy π^* such that $\pi^* = \arg\max_{\pi \in \Pi} \Phi(s, \pi)$ for every $s \in S$, then π^* is a stationary Nash equilibrium policy of the MPG. However, computing the Nash equilibrium as the maximizer of $\Phi(s, \cdot)$ is impossible in our setting since the players do not have the knowledge of the potential function $\Phi(s, \cdot)$ or the oracle access to its value. Moreover, even in settings where the potential function is known, e.g. common interest games, computing its maximizer is challenging due to the fact that $\Phi(s, \pi)$ is non-linear and non-concave in π.

4.2 Independent and Decentralized Learning Dynamics

In this section, we present the learning dynamics and characterize its long-run behavior. First, we highlight the information available to every player in the learning process.

Discussion on available information to players. We assume that each player $i \in I$ knows their own action set A_i and the state set S. Players do *not* know the state transition probability matrix P, their own or others' payoff functions $(u_i)_{i\in I}$, and they do not know the initial state distribution $\mu \in \Delta(S)$. Players do not know the existence of other players or the underlying potential function of the game. In each stage $k = 0, 1, 2, \dots$ of the learning algorithm, players observe the realized state s^k that they use to compute the action a^k and in turn obtain the reward $r_i^k = u_i(s^k, a^k)$. We want to emphasize that the players only receive the bandit feedback of reward function, i.e. in any stage they do not receive the reward corresponding to the action they did not choose. Additionally, players do not observe the actions or the rewards of their opponents. Moreover, players do not know the parameters used by other players in the learning dynamics (to be presented shortly).

Given any policy $\pi \in \Pi$, and any initial state $s \in S$, we define the following *Q-function* for each player $i \in I$ and action $a_i \in A_i$:

$$Q_i(s, a_i; \pi) = u_i(s, a_i, \pi_{-i}) + \delta \sum_{s' \in S} P(s'|s, a_i, \pi_{-i}) V_i(s', \pi). \tag{4.2}$$

In (4.2), player i's expected utility in the first stage with state s is derived from playing action a_i and her opponents choosing policy π_{-i}. The expected total utility starting from stage 2 is derived from all players following policy π. Therefore, the Q-function $Q_i(s, a_i; \pi)$ represents player i's expected discounted utility when the game starts in state s, and player i deviates for *one-stage* (namely, the first stage) from her policy to play a_i. With slight abuse of notation, we define $Q_i(s; \pi) = (Q_i(s, a_i; \pi))_{a_i \in A_i} \in \mathbb{R}^{|A_i|}$ for every $i \in I, s \in S, \pi \in \Pi$. Furthermore, we define *optimal one-stage deviation* from policy $\pi \in \Pi$ in state $s \in S$ as

$$\mathrm{br}_i(s; \pi) = \arg\max_{\hat{\pi}_i \in \Delta(A_i)} \hat{\pi}_i^\top Q_i(s; \pi). \tag{4.3}$$

One can obtain equivalent characterization of Nash equilibrium in terms of optimal one-stage deviation. Specifically, a policy is a Nash equilibrium if and only if it is fixed point of optimal one-stage deviation (Lemma C.3.2).

Learning Dynamics

The proposed learning happens in iterates, denoted by t. In every iterate t, each player $i \in I$ updates four components $n^t, \tilde{n}_i^t, \tilde{q}_i^t, \pi_i^t$. In particular, $n^t = (n^t(s))_{s\in S}$ is the vector of state counters, where $n^t(s)$ is the number of times state s is realized before iterate t. For each player $i \in I$, $\tilde{n}_i^t = (\tilde{n}_i^t(s, a_i))_{s\in S, a_i \in A_i}$ is the counter of state-action tuple, where $\tilde{n}_i^t(s, a_i)$ is the number of times that player i has played action a_i in state s before iterate t.

Additionally, $\tilde{q}_i^t = (\tilde{q}_i^t(s, a_i))_{s\in S, a_i \in A_i}$ is player i's estimate of her Q-function, and $\pi_i^t = (\pi_i^t(s, a_i))_{s\in S, a_i \in A_i}$ is player i's policy in iterate t. Given the state-(local) action tuple of any player i, (s^{t-1}, a_i^{t-1}), the estimate $\tilde{q}_i^t(s^{t-1}, a_i^{t-1})$ is updated in (4.4) as a linear combination of the estimate $\tilde{q}_i^{t-1}(s^{t-1}, a_i^{t-1})$ in the previous stage, and a new estimate that is comprised

of the realized one-stage payoff r_i^{t-1} and the estimated total discounted payoff from the next iterate.

The policy $\pi_i^t(s^{t-1})$ updated in (4.5) is a linear combination of the policy $\pi_i^{t-1}(s^{t-1})$ in the previous stage, and player i's optimal one-stage deviation. Particularly, the optimal one-stage deviation is computed using the updated q-estimate $\tilde{q}_i^t$, instead of the actual Q_i, which is unknown to the player.

Each player's action a_i^t is sampled from their policy π_i^t with probability $(1-\theta_i)$ and from a uniform distribution over their action set A_i with probability θ_i, where $\theta_i \in (0,1)$ is the exploration parameter that can be heterogeneous among players (refer to (4.6)). With slight abuse of notation, we define $\theta = (\theta_i)_{i\in I} \in (0,1)^{|I|}$.

Note that the updates of $\tilde{q}_i^t$ (resp. π_i^t), in each iterate, only change the element that corresponds to the realized state and action (s^{t-1}, a^{t-1}) (resp. state s^{t-1}), and the remaining elements stay unchanged. Furthermore, the update speed of $\tilde{q}_i^t(s^{t-1}, a_i^{t-1})$ (resp. $\pi_i^t(s^{t-1})$) is governed by the step size sequence $\alpha_i(n)$ (resp. $\beta_i(n)$) corresponds to the state-action counter $\tilde{n} = \tilde{n}_i^t(s^{t-1}, a_i^{t-1})$ (resp. state counter $n = n^t(s^{t-1})$). Therefore, the update is *asynchronous* in that the stepsizes are different across the elements associated with different states and actions, and stepsizes are different for different players.

Convergence Analysis

Next, we introduce the assumptions that are needed for obtaining the convergent set of the learning dynamics.

Assumption 4.2.1. *The initial state distribution $\mu(s) > 0$ for all $s \in S$. Additionally,*

$$\min_{s,s'\in S, a\in A} P(s'|s,a) > 0.$$

Assumption 4.2.1 ensures that every state is visited infinitely often so that agents can learn the Q-function associated with each state. We note that similar assumptions on the probability transition are commonly made in multi-agent reinforcement learning literature. For example, the paper [240] also assumed that $\min_{s,s'\in S,a\in A} P(s'|s,a) > 0$ and [367] assumed that for any pair of states $(s,s') \in S \times S$ and any infinite sequence of joint actions, the state s' is reachable from s in finite steps.

Next, we make the following assumption on the stepsizes:

Assumption 4.2.2. *The step sizes $\{\alpha_i(n) \in (0,1)\}_{n=0,i\in I}^{\infty}$ and $\{\beta_i(n) \in (0,1)\}_{n=0,i\in I}^{\infty}$ satisfy*

(i) For all $i \in I$, $\sum_{n=0}^{\infty} \alpha_i(n) = \infty, \sum_{n=0}^{\infty} \beta_i(n) = \infty$, $\lim_{n\to\infty} \alpha_i(n) = \lim_{n\to\infty} \beta_i(n) = 0$;

(ii) For every $i \in I$, there exist some $q, q' \geqslant 2$,

$$\sum_{n=0}^{\infty} \alpha_i(n)^{1+q/2} < \infty, \quad \text{and} \quad \sum_{n=0}^{\infty} \beta_i(n)^{1+q'/2} < \infty;$$

Algorithm 3 Independent and decentralized learning dynamics

Initialization: $n^0(s) = 0, \forall s \in S$; $\tilde{n}_i^0(s, a_i) = 0, \tilde{q}_i^0(s, a_i) = 0$, set arbitrary $\pi_i^0(s, a_i),\ \forall (i, a_i, s)$, and $\theta_i \in (0, 1), \forall\ i$.

In iterate 0, each player observes $s^0 \in S$, choose their action $a_i^0 \sim \pi_i^0(s^0)$, and observe $r_i^0 = u_i(s^0, a^0)$.

In every iterate $t = 1, 2, ...,$ each player observes s^t, and independently updates $\{n^t, \tilde{n}_i^t, \tilde{q}_i^t, \pi_i^t\}$.

Update $n^t, \tilde{n}_i^t$:

$$\begin{aligned} n^t(s^{t-1}) &= n^{t-1}(s^{t-1}) + 1, \\ \tilde{n}_i^t(s^{t-1}, a_i^{t-1}) &= \tilde{n}_i^{t-1}(s^{t-1}, a_i^{t-1}) + 1, \end{aligned}$$

Furthermore, for $s \neq s^{t-1}, a_i \neq a_i^{t-1}$,

$$\tilde{n}_i^t(s, a_i) = \tilde{n}_i^{t-1}(s, a_i), \quad n^t(s) = n^{t-1}(s).$$

Update $\tilde{q}_i^t$:

$$\begin{aligned} \tilde{q}_i^t(s^{t-1}, a_i^{t-1}) &= \tilde{q}_i^{t-1}(s^{t-1}, a_i^{t-1}) + \alpha_i(\tilde{n}_i^t(s^{t-1}, a_i^{t-1})) \\ &\cdot \left(r_i^{t-1} + \gamma \pi_i^{t-1}(s^t)^\top \tilde{q}_i^{t-1}(s^t) - \tilde{q}_i^{t-1}(s^{t-1}, a_i^{t-1}) \right), \end{aligned} \tag{4.4a}$$

$$\tilde{q}_i^t(s, a_i) = \tilde{q}_i^{t-1}(s, a_i), \quad \forall\ (s, a_i) \neq (s^{t-1}, a_i^{t-1}), \tag{4.4b}$$

where $r_i^{t-1} = u_i(s^{t-1}, a^{t-1})$.

Update π_i^t*:*

$$\begin{aligned} \pi_i^t(s^{t-1}) &= \pi_i^{t-1}(s^{t-1}) + \beta_i(n^t(s^{t-1})) \cdot \\ &\left(\widehat{\mathsf{br}}_i^{t-1}(s^{t-1}) - \pi_i^{t-1}(s^{t-1}) \right), \end{aligned} \tag{4.5a}$$

$$\pi_i^t(s) = \pi_i^{t-1}(s), \quad \forall\ s \neq s^{t-1}, \tag{4.5b}$$

where $\widehat{\mathsf{br}}_i^t(s) \in \arg\max_{\pi_i \in \Delta(A_i)} \pi_i^\top q_i^t(s)$ for every $s \in S$.

At the end of iterate t, each player chooses their action

$$a_i^t \sim (1 - \theta_i)\pi_i^t(s^t) + \theta_i \cdot (1/|A_i|)\mathbb{1}_{A_i}, \tag{4.6}$$

where $\mathbb{1}_{A_i} \in \mathbb{R}^{|A_i|}$ is a vector with all entries 1.

Each player observes their own reward $r_i^t = u_i(s^t, a^t)$.

(iii) *For every* $i \in I, x \in (0,1)$, $\sup_n \alpha_i([xn])/\alpha_i(n) < A_x < \infty$, $\sup_n \beta_i([xn])/\beta_i(n) < A_x < \infty$, *where* $[xn]$ *denotes the largest integer less than or equal to* xn. *Additionally,* $\{\alpha_i(n)\}, \{\beta_i(n)\}$ *are non-increasing in* n;

(iv) *For every* $i, j \in I$, $\lim_{n\to\infty} \beta_i(n)/\alpha_j(n) = 0$;

(v) *For every* $i, j \in I$, *there exists* $0 < \xi_{ij}^{\alpha} < \zeta_{ij}^{\alpha} < \infty$, *and* $0 < \xi_{ij}^{\beta} < \zeta_{ij}^{\beta} < \infty$ *such that* $\frac{\alpha_i(n)}{\alpha_j(n)} \in [\xi_{ij}^{\alpha}, \zeta_{ij}^{\alpha}]$, *and* $\frac{\beta_i(n)}{\beta_j(n)} \in [\xi_{ij}^{\beta}, \zeta_{ij}^{\beta}]$ *for all* n.

Assumption 4.2.2(i) ensures that the asymptotic properties of our learning dynamics can be studied by using a continuous-time dynamical system [60, 42, 408]. Assumption 4.2.2(ii) ensures that the average cumulative impact of noise terms on the asymptotic behavior of learning dynamics diminishes. Assumption 4.2.2(iii) is a technical condition required for the asynchronous update in the learning dynamics to ensure that the value functions and policies associated with different state-action pairs are updated at the same timescale [337]. Assumption 4.2.2(iv) implies that our learning dynamics have two timescales: the update of $\{\tilde{q}_i^t\}_{t=0}^{\infty}$ is asymptotically faster than the update of $\{\pi_i^t\}_{t=0}^{\infty}$. Assumption 4.2.2(v) suggests that players can employ varying step sizes, provided that the ratio between their step sizes remains bounded between zero and a finite number for all steps. One example of stepsizes that satisfies Assumption 4.2.2 is $\alpha_i(n) = z_i n^{-c_1}$ and $\beta_i(n) = y_i n^{-c_2}$ where $0 < c_1 \leqslant c_2 \leqslant 1$ and y_i, z_i can be any player specific positive scalars.

Next, we state the main result of this chapter that characterizes the convergent set of policy updates of Algorithm 3 as a super-level set of potential function Φ (Theorem 4.2.1). Furthermore, by imposing additional assumption on the potential function and the set of Nash equilibrium, we characterize the convergent set as a set of approximate Nash equilibrium (Corollary 4.2.1).

Theorem 4.2.1. *Under Assumptions 4.2.1 and 4.2.2, for every* $\epsilon > 0$, *the policy sequence* $\{\pi^t\}_{t=0}^{\infty}$ *induced by Algorithm 3 converges to the set*

$$\Pi_{\epsilon}^{*} := \{\pi \in \Pi : \Phi(\mu, \pi) \geqslant \min_{\pi' \in \mathit{NE}(\epsilon)} \Phi(\mu, \pi')\} \tag{4.7}$$

with probability 1 given that

$$\sum_{i \in I} \theta_i < \epsilon \cdot \frac{\eta(1-\gamma)^2}{4u_{\max}|I|(\eta + (D/(1-\gamma)))} =: L\epsilon, \tag{4.8}$$

where

$$D = \frac{1}{1-\gamma} \max_{i, \pi_{-i}, s} \left| \frac{d_{\mu}^{\pi_i^{\dagger}, \pi_{-i}}(s)}{\mu(s)} \right|, \pi_i^{\dagger} \in \arg\max_{\pi_i \in \Pi_i} V_i(\mu, \pi_i, \pi_{-i}),$$
$$\eta = \frac{\zeta}{A_{\zeta}}, \quad \zeta = \min_{s, s' \in S, a \in A} P(s'|s, a),$$

and A_ζ is defined as in Assumption 4.2.2 (iii). Moreover, for any ϵ, ϵ' such that $0 \leqslant \epsilon \leqslant \epsilon'$, $\Pi^*_\epsilon \subseteq \Pi^*_{\epsilon'}$. Additionally, for every $\epsilon \geqslant 0$, $\mathsf{NE}(\epsilon) \subseteq \Pi^*_\epsilon$.

Theorem 4.2.1 guarantees that for any $\epsilon > 0$, for sufficiently small exploration rate that satisfies (4.8), the sequence of policies induced by the learning dynamics asymptotically converges to the smallest super-level set of potential function that contains the set of ϵ-stationary Nash equilibrium (i.e. $\mathsf{NE}(\epsilon)$). Moreover, Theorem 4.2.1 states that for the policy sequence induced by Algorithm 3 to converge to the set of approximate Nash equilibria with smaller approximation gap, the sum of the exploration probabilities of all players should be smaller.

The convergence result of Theorem 4.2.1 can be refined to ensure convergence to an approximate Nash equilibrium set under additional Assumption 4.2.3.

Assumption 4.2.3. *For every $\epsilon > 0$, there exists $h_\epsilon \in \mathbb{R}_+$ such that $\Pi^*_\epsilon \subseteq \mathsf{NE}(\epsilon + h_\epsilon)$, h_ϵ is continuous and non-decreasing in ϵ, and $\lim_{\epsilon \downarrow 0} h_\epsilon = 0$.*

Corollary 4.2.1. *Suppose that Assumptions 4.2.1, 4.2.2 and 4.2.3 hold. For every $\tilde{\epsilon}, \tilde{\epsilon}'$ such that $0 < \tilde{\epsilon} < \tilde{\epsilon}'$, there exist positive scalars $0 < \epsilon < \epsilon'$ such that $\epsilon + h_\epsilon = \tilde{\epsilon}$ and $\epsilon' + h_{\epsilon'} = \tilde{\epsilon}'$, and the sequence of policies $\{\pi^t\}_{t=0}^\infty$ induced by Algorithm 3 converges to the set $\mathsf{NE}(\tilde{\epsilon})$ (resp. $\mathsf{NE}(\tilde{\epsilon}')$) with probability 1, if $\sum_{i \in I} \theta_i < L\epsilon$ (resp. $\sum_{i \in I} \theta_i < L\epsilon'$).*

Corollary 4.2.1 states that for the policy sequence induced by Algorithm 3 to converge to the set of approximate Nash equilibria with a smaller approximation gap, the sum of the exploration probabilities of all players must be smaller. We have included the proof of Corollary 4.2.1 in Section C.2.

Now we prove Theorem 4.2.1. First, we introduce some useful notations. For any $i \in I, s \in S$, we define $\pi_i^\circ(s) = (1/|A_i|) \cdot \mathbb{1}_{A_i}$ to be a uniformly random policy. Additionally, for any $\pi \in \Pi$ and any $\theta = (\theta_i)_{i \in I} \in (0,1)^{|I|}$, we define $\pi^{(\theta)} \in \Pi$ such that for every $s \in S, i \in I$, $\pi_i^{(\theta)}(s) := (1 - \theta_i)\pi_i(s) + \theta_i \pi_i^\circ(s)$ to be a perturbed version of policy π given exploration probability vector θ.

To prove Theorem 4.2.1, we apply the two-timescale asynchronous stochastic approximation theory [337], where we first ensure the convergence of the ("fast") q-estimate updates, $\{\tilde{q}_i^t\}_{t=0}^\infty$, in Lemma 4.2.1. Next, in Lemma 4.2.2-4.2.4, we study the convergent set of the ("slow") policy updates given the convergent values of q-estimates.

Lemma 4.2.1. *Under Assumptions 4.2.1 and 4.2.2, for any $s \in S$ and $i \in I$, $\lim_{t \to \infty} \|\tilde{q}_i^t(s) - Q_i(s; (\pi_i^t, \pi_{-i}^{t,(\theta)}))\|_\infty = 0$ with probability 1.*

The proof of Lemma 4.2.1 is based on two steps. First, we show that under Assumption 4.2.1 and 4.2.2, our learning dynamics satisfies the set of conditions introduced in [337] (restated in Section C.1) so that convergence of q-estimates can be analyzed by the associated continuous-time dynamical system where the policy drifts are treated as asymptotically negligible errors. Second, we argue the global convergence of the continuous-time dynamical

system using the contraction property of the Bellman operator associated with q-estimate update. The complete proof of Lemma 4.2.1 is deferred to Section C.2.

Next, we analyze the policy updates with respect to the convergent values of the q-estimates provided by the fast dynamics as in Lemma 4.2.1. Particularly, the policy $\pi_i^t(s^{t-1})$ in (4.5) becomes a linear combination of $\pi_i^{t-1}(s^{t-1})$, and the optimal one-stage deviation $\mathsf{br}_i(s^{t-1};(\pi_i^{t-1},\pi_{-i}^{t-1,(\theta)}))$ based on the actual Q-function as in (4.3). Under Assumption 4.2.1, the asymptotic behavior of $\{\pi^t\}_{t=0}^{\infty}$ can be analyzed using the following continuous-time differential inclusion, where $\tau\in[0,\infty)$ is a continuous-time index,

$$\frac{d}{d\tau}\varpi_i^\tau(s)\in\gamma_i(s)\left(\mathsf{br}_i(s;(\varpi_i^\tau,\varpi_{-i}^{\tau,(\theta)}))-\varpi_i^\tau(s)\right), \tag{4.9}$$

and $\gamma_i(s)\in[\eta,1]$, for every $s\in S, i\in I$, captures [2] the asynchronous update of policies in different states (cf. (4.5)), and $\eta=\zeta/A_\zeta>0$ [337]. Since $\mathsf{br}_i(\cdot)$ is a non-empty closed, convex and compact set, there exists an absolutely continuous solution of (4.9), ϖ_i^τ for every $i\in I$ [43]. Consequently, for every $i\in I$ and $s\in S$, there exists

$$\widetilde{\mathrm{br}}_i^{\tau,(\theta)}(s)\in\mathrm{br}_i(s;(\varpi_i^\tau,\varpi_{-i}^{\tau,(\theta)}))$$

such that

$$\frac{d}{d\tau}\varpi_i^\tau(s)=\gamma_i(s)\left(\widetilde{\mathrm{br}}_i^{\tau,(\theta)}(s)-\varpi_i^\tau(s)\right). \tag{4.10}$$

To establish the convergence of (4.10), we define a Lyapunov function $\phi:[0,\infty)\to\mathbb{R}$ as follows:

$$\phi(\tau)=\max_{\varpi\in\Pi}\sum_{s\in S}\mu(s)\Phi(s,\varpi)-\sum_{s\in S}\mu(s)\Phi(s,\varpi^\tau), \tag{4.11}$$

which is the difference of the potential function at its maximizer with that of its value at ϖ^τ. We show that $\phi(\tau)$ is strictly decreasing (alternatively, the potential function is strictly increasing) as long as ϖ^τ is outside the set $\mathsf{NE}(\epsilon)$.

Lemma 4.2.2. *Suppose that Assumptions 4.2.1 holds and θ satisfies* (4.8)*. For every $\tau\geqslant 0$ and $\epsilon>0$, if $\varpi^\tau\notin\mathsf{NE}(\epsilon)$ then $d\phi(\tau)/d\tau<0$.*

Unlike static potential games, the potential function is non-concave in each player's policy in Markov potential games. Therefore, we need a new approach to demonstrate that $\phi(\tau)$ decreases outside a neighborhood of approximate Nash equilibrium. To prove Lemma 4.2.2, we need the following technical lemma that extends the single-agent reinforcement learning theory to multi-agent games.

[2] From two-timescale asynchronous stochastic approximation theory [337] (also reviewed in Section C.1), the value of $\gamma_i(s)\geqslant\kappa/A_\kappa$, where κ is a lower bound on the stationary distribution of the Markov chain over the state space induced by any policy. Assumption 4.2.1 ensures that the probability of every state in this stationary distribution, under any policy, is greater than ζ.

Lemma 4.2.3. *(a)* Gradient of value function*: For any* $\mu \in \Delta(S), s \in S, \pi \in \Pi, i \in I, a_i \in A_i$*,*

$$\frac{\partial V_i(\mu,\pi)}{\partial \pi_i(s,a_i)} = \frac{1}{1-\gamma} d_\mu^\pi(s) Q_i(s,a_i;\pi),$$

where

$$d_\mu^\pi(s) \coloneqq (1-\gamma) \sum_{s^0 \in S} \mu(s^0) \sum_{k=0}^{\infty} \gamma^k \mathsf{Pr}(s^k = s | s^0). \tag{4.12}$$

(b) Multi-agent performance difference lemma*: For any policy* $\pi = (\pi_i, \pi_{-i}), \pi' = (\pi_i', \pi_{-i}) \in \Pi$ *and any* $\mu \in \Delta(S)$*,*

$$V_i(\mu,\pi) - V_i(\mu,\pi') = \frac{1}{1-\gamma} \sum_{s'} d_\mu^\pi(s') \Gamma_i(s', \pi_i; \pi'), \tag{4.13}$$

where $\Gamma_i(s,a_i;\pi)$ *is the* advantage function *given by*

$$\Gamma_i(s,a_i;\pi) \coloneqq Q_i(s,a_i;\pi) - V_i(s,\pi), \tag{4.14}$$

for every $i \in I,\ s \in S,\ a_i \in A_i, \pi \in \Pi$*.*

(c) Sensitivity of value function: For any $i \in I, \pi_i \in \Pi_i, \pi_{-i} \in \Pi_{-i}$*,*

$$\max_{s \in S} |V_i(s,\pi_i,\pi_{-i}) - V_i(s,\pi_i,\pi_{-i}^{(\theta)})| \leqslant \frac{2\sum_{k \in I \setminus \{i\}} \theta_k}{(1-\gamma)^2} u_{\max}.$$

(d) Sensitivity of Q-function: For any $i \in I, \pi_i \in \Pi_i, \pi_{-i} \in \Pi_{-i}$*, it holds that*

$$\max_{s,a_i} |Q_i(s,a_i;\pi) - Q_i(s,a_i;\pi_i,\pi_{-i}^{(\theta)})| \leqslant \frac{2\sum_{k \in I \setminus \{i\}} \theta_k}{(1-\gamma)^2} u_{\max}.$$

The proof of Lemma 4.2.3 is included in Section C.2. We are now ready to prove Lemma 4.2.2 based on Lemma 4.2.3.

Proof of Lemma 4.2.2: We compute the derivative of $\phi(\tau)$ with respect to τ, where $\phi(\tau)$ is

given by (4.11).

$$\begin{aligned}
\frac{d}{d\tau}\phi(\tau) &= -\sum_{i,s}\frac{\partial\Phi(\mu,\varpi^\tau)}{\partial\varpi_i(s)}\frac{d\varpi_i^\tau(s)}{d\tau}\\
&\underset{(i)}{=} -\sum_{i,s}\frac{\partial V_i(\mu,\varpi^\tau)}{\partial\varpi_i(s)}\frac{d\varpi_i^\tau(s)}{dt}\\
&\underset{(ii)}{=} \sum_{i,s}\frac{d_\mu^{\varpi^\tau}(s)}{\gamma-1}\gamma_i(s)Q_i(s;\varpi^\tau)^\top\left(\widetilde{\mathrm{br}}_i^{\tau,(\theta)}(s)-\varpi_i^\tau(s)\right),\\
&\underset{(iii)}{=} \sum_{i,s}\frac{d_\mu^{\varpi^\tau}(s)}{\gamma-1}\gamma_i(s)(\Delta Q_i(s))^\top\left(\widetilde{\mathrm{br}}_i^{\tau,(\theta)}(s)-\varpi_i^\tau(s)\right)\\
&+\sum_{i,s}\frac{d_\mu^{\varpi^\tau}(s)}{\gamma-1}\gamma_i(s)(Q_i(s;\varpi_i^\tau,\varpi_{-i}^{\tau,(\theta)}))^\top\left(\widetilde{\mathrm{br}}_i^{\tau,(\theta)}(s)-\varpi_i^\tau(s)\right),
\end{aligned}\tag{4.15}$$

where

$$\Delta Q_i(s) = Q_i(s;\varpi^\tau) - Q_i(s;\varpi_i^\tau,\varpi_{-i}^{\tau,(\theta)}).$$

Here, (i) is due to Lemma C.3.1, (ii) is due to Lemma 4.2.3(a) and (4.10), and (iii) is by adding and subtracting terms. Note that the first term in the RHS of (4.15) can be bounded as

$$\begin{aligned}
&\sum_{i,s}\frac{d_\mu^{\varpi^\tau}(s)}{\gamma-1}\gamma_i(s)(\Delta Q_i(s))^\top\left(\widetilde{\mathrm{br}}_i^{\tau,(\theta)}(s)-\varpi_i^\tau(s)\right)\\
&\leqslant \frac{1}{1-\gamma}\sum_{i\in I}\max_{s\in S}\left|(\Delta Q_i(s))^\top\left(\widetilde{\mathrm{br}}_i^{\tau,(\theta)}(s)-\varpi_i^\tau(s)\right)\right|\\
&\leqslant \frac{2}{1-\gamma}\sum_{i\in I}\max_{s\in S,a_i\in A_i}\left|(\Delta Q_i(s,a_i))\right| \leqslant \frac{4|I|\sum_{i\in I}\theta_i u_{\max}}{(1-\gamma)^3},
\end{aligned}\tag{4.16}$$

where the first inequality is because $\gamma_i(s)\leqslant 1$ as it denotes the fraction of time state s is visited, and last inequality is due to Lemma 4.2.3(d).

Recall that $\widetilde{\mathrm{br}}_i^{\tau,(\theta)}(s)\in \mathrm{br}_i(s;(\varpi_i^\tau,\varpi_{-i}^{\tau,(\theta)}))$ where $\mathrm{br}_i(s;\varpi) = \underset{\hat{\varpi}_i\in\Delta(A_i)}{\arg\max}\ \hat{\varpi}_i^\top Q_i(s;\varpi)$ for every $s\in S,\varpi\in\Pi$. Therefore, $Q_i(s;(\varpi_i^\tau,\varpi_{-i}^{\tau,(\theta)}))^\top\left(\widetilde{\mathrm{br}}_i^{\tau,(\theta)}(s)-\varpi_i^\tau(s)\right)\geqslant 0$. Additionally, given the fact that $\gamma_i(s)\geqslant\eta$ for all $i\in I, s\in S$, (4.15), and (4.16), we obtain

$$\frac{d}{d\tau}\phi(\tau)\leqslant\frac{4|I|\sum_{i\in I}\theta_i u_{\max}}{(1-\gamma)^3}-\frac{\eta}{1-\gamma}\sum_{i,s}d_\mu^{\varpi^\tau}(s)\cdot Q_i(s;(\varpi_i^\tau,\varpi_{-i}^{\tau,(\theta)}))^\top\left(\widetilde{\mathrm{br}}_i^{\tau,(\theta)}(s)-\varpi_i^\tau(s)\right).\tag{4.17}$$

Additionally, let $\pi_i^\dagger\in\arg\max_{\pi_i\in\Pi_i}V_i(\mu,\pi_i,\varpi_{-i}^\tau)$ be a best response of player i if the joint strategy of other players is ϖ_{-i}^τ. Note that $\pi_i^\dagger$ maximizes the total payoff instead of

just maximizing the payoff of one-stage deviation. Therefore, $\pi_i^\dagger$ can be different from the optimal one-stage deviation policy. We drop the dependence of $\pi_i^\dagger$ on ϖ_{-i}^τ for notational simplicity.

Recall that $D = \frac{1}{1-\gamma}\max_{i,\varpi_{-i}} \|d_\mu^{\pi_i^\dagger,\varpi_{-i}^\tau}/\mu\|_\infty$. We note that D is finite under the assumption that μ has full support (Assumption 4.2.1). Next, we provide an inequality that is crucial to bound (4.17). Note that

$$
\begin{aligned}
&\sum_{i,s} d_\mu^{\pi_i^\dagger,\varpi_{-i}^{\tau,(\theta)}}(s) Q_i(s;\varpi_i^\tau,\varpi_{-i}^{\tau,(\theta)})^\top \left(\pi_i^\dagger(s)-\varpi_i^\tau(s)\right)\\
&\leqslant \sum_{i,s} d_\mu^{\pi_i^\dagger,\varpi_{-i}^{\tau,(\theta)}}(s)\cdot \max_{\hat\pi_i\in\Delta(A_i)} Q_i(s;\varpi_i^\tau,\varpi_{-i}^{\tau,(\theta)})^\top\left(\hat\pi_i(s)-\varpi_i^\tau(s)\right)\\
&\leqslant \sum_{i,s} d_\mu^{\varpi^\tau}(s)\left\|\frac{d_\mu^{\pi_i^\dagger,\varpi_{-i}^{\tau,(\theta)}}}{d_\mu^{\varpi^\tau}}\right\|_\infty \cdot \max_{\hat\pi_i\in\Delta(A_i)} Q_i(s;\varpi_i^\tau,\varpi_{-i}^{\tau,(\theta)})^\top\left(\hat\pi_i(s)-\varpi_i^\tau(s)\right)\\
&\overset{(i)}{\leqslant} D\sum_{i,s} d_\mu^{\varpi^\tau}(s)\cdot \max_{\hat\pi_i\in\Delta(A_i)} Q_i(s;\varpi_i^\tau,\varpi_i^{\tau,(\theta)})^\top\left(\hat\pi_i(s)-\varpi_i^\tau(s)\right)\\
&= D\sum_{i,s} d_\mu^{\varpi^\tau}(s)\cdot Q_i(s;\varpi_i^\tau,\varpi_{-i}^{\tau,(\theta)})^\top\left(\widetilde{\mathrm{br}}_i^{\tau,(\theta)}(s)-\varpi_i^\tau(s)\right),
\end{aligned}
$$

where *(i)* is using[3] $d_\mu^{\varpi^{\tau,(\theta)}}(s)\geqslant(1-\gamma)\mu(s)$ along with the definition of D. Therefore,

$$
\begin{aligned}
&\sum_{i,s} d_\mu^{\varpi^\tau}(s)\cdot Q_i(s;\varpi_i^\tau,\varpi_{-i}^{\tau,(\theta)})^\top\left(\widetilde{\mathrm{br}}_i^{\tau,(\theta)}(s)-\varpi_i^\tau(s)\right)\\
\geqslant&\frac{1}{D}\sum_{i,s} d_\mu^{\pi_i^\dagger,\varpi_{-i}^{\tau,(\theta)}}(s) Q_i(s;\varpi_i^\tau,\varpi_{-i}^{\tau,(\theta)})^\top\left(\pi_i^\dagger(s)-\varpi_i^\tau(s)\right).
\end{aligned}
\tag{4.18}
$$

Then, from (4.17) and (4.18), we have

$$
\frac{d}{dt}\phi(\tau)\leqslant\frac{4|I|\sum_{i\in I}\theta_i u_{\max}}{(1-\gamma)^3}-\frac{\eta}{D(1-\gamma)}\sum_{i,s} d_\mu^{\pi_i^\dagger,\varpi_{-i}^{\tau,(\theta)}}(s)\cdot Q_i(s;\varpi_i^\tau,\varpi_{-i}^{\tau,(\theta)})^\top\left(\pi_i^\dagger(s)-\varpi_i^\tau(s)\right).
\tag{4.19}
$$

[3]This is obtained by dropping all terms corresponding to $k\geqslant 1$ in (4.12).

Next, we note that

$$\begin{aligned}
&Q_i(s;\varpi_i^\tau,\varpi_{-i}^{\tau,(\theta)})^\top(\pi_i^\dagger(s)-\varpi_i^\tau(s))\\
&\overset{(i)}{=}\sum_{a_i\in A_i}\Big(Q_i(s,a_i;\varpi_i^\tau,\varpi_{-i}^{\tau,(\theta)})-V_i(s,\varpi_i^\tau,\varpi_{-i}^{\tau,(\theta)})\Big)\cdot(\pi_i^\dagger(s,a_i)-\varpi_i^\tau(s,a_i))\\
&\overset{(ii)}{=}\sum_{a_i}\Gamma_i(s,a_i;\varpi_i^\tau,\varpi_{-i}^{\tau,(\theta)})\pi_i^\dagger(s,a_i)=\Gamma_i(s,\pi_i^\dagger;\varpi_i^\tau,\varpi_{-i}^{\tau,(\theta)}),
\end{aligned}\tag{4.20}$$

where (i) holds because V_i is not dependent on actions, and (ii) follows from the definition of advantage function in (4.14). Combining (4.19) and (4.20),

$$\begin{aligned}
&\frac{d\phi(\tau)}{dt}\leqslant\frac{4|I|\sum_{i\in I}\theta_i u_{\max}}{(1-\gamma)^3}-\frac{\eta}{D(1-\gamma)}\sum_{i,s}d_\mu^{\pi_i^\dagger,\varpi_{-i}^{\tau,(\theta)}}(s)\Gamma_i(s,\pi_i^\dagger;\varpi_i^\tau,\varpi_{-i}^{\tau,(\theta)})\\
=&\frac{4|I|\sum_{i\in I}\theta_i u_{\max}}{(1-\gamma)^3}-(\eta/D)\cdot\sum_i\Big(V_i(\mu,\pi_i^\dagger,\varpi_{-i}^{\tau,\theta})-V_i(\mu,\varpi_i^\tau,\varpi_{-i}^{\tau,\theta})\Big)
\end{aligned}\tag{4.21}$$

where the equality follows from the multi-agent performance difference lemma (Lemma 4.2.3(b)). Next, note that

$$\begin{aligned}
&-(\eta/D)\cdot\sum_i\Big(V_i(\mu,\pi_i^\dagger,\varpi_{-i}^{\tau,(\theta)})-V_i(\mu,\varpi_i^\tau,\varpi_{-i}^{\tau,(\theta)})\Big)\\
&\leqslant-(\eta/D)\cdot\sum_i\Big(V_i(\mu,\pi_i^\dagger,\varpi_{-i}^{\tau})-V_i(\mu,\varpi_i^\tau,\varpi_{-i}^{\tau})\Big)\\
&\quad+(2\eta/D)\cdot\max_{\pi_i\in\Pi_i}\sum_i|V_i(\mu,\pi_i,\varpi_{-i}^{\tau,(\theta)})-V_i(\mu,\pi_i,\varpi_{-i}^{\tau})|\\
&\leqslant-\frac{\eta}{D}\sum_i\Big(V_i(\mu,\pi_i^\dagger,\varpi_{-i}^{\tau})-V_i(\mu,\varpi)\Big)+\frac{4\eta|I|\sum_{i\in I}\theta_i u_{\max}}{D(1-\gamma)^2},
\end{aligned}\tag{4.22}$$

where the first inequality is based on adding and subtracting the term

$$-(\eta/D)\sum_i(V_i(\mu,\pi_i^\dagger,\varpi_{-i}^{\tau})-V_i(\mu,\varpi_i^\tau,\varpi_{-i}^{\tau}))$$

, arranging terms and taking maximum over π_i, and the last inequality is due to Lemma 4.2.3(c).

Combining (4.21) and (4.22), we obtain

$$\begin{aligned}
&d\phi(\tau)/d\tau\\
&\leqslant\frac{4|I|\sum_{i\in I}\theta_i u_{\max}}{(1-\gamma)^3}+\frac{4\eta|I|\sum_{i\in I}\theta_i u_{\max}}{D(1-\gamma)^2}-(\eta/D)\cdot\sum_i\Big(V_i(\mu,\pi_i^\dagger,\varpi_{-i}^{\tau})-V_i(\mu,\varpi^\tau)\Big).
\end{aligned}$$

Since $\pi_i^\dagger$ is a best response to ϖ_{-i}^τ, $V_i(\mu, \pi_i^\dagger, \varpi_{-i}^\tau) - V_i(\mu, \varpi^\tau) \geqslant 0$ for all i. Furthermore, for any $\varpi^\tau \notin \mathsf{NE}(\epsilon)$, there must exist at least one player $i \in I$ and best response policy $\pi_i^\dagger \in \Pi_i$ such that $V_i(\mu, \pi_i^\dagger, \varpi_{-i}^\tau) \geqslant V_i(\mu, \varpi_i^\tau, \varpi_{-i}^\tau) + \epsilon$. Therefore, given that θ satisfies (4.8), we have $d\phi(\tau)/d\tau < 0$.

■

Next, we present the following result that leverages Lemma 4.2.2 to obtain the convergent set of the sequence of policies induced by Algorithm 3.

Lemma 4.2.4. *Under the conditions stated in Theorem 4.2.1, the sequence of policies $(\pi^t)_{t=0}^\infty$ given by Algorithm 3 almost surely converges to Π_ϵ^* with probability 1.*

Proof. By applying the asynchronous stochastic approximation theory [337, Theorem 4.7], the asymptotic behavior of the policies $(\pi^t)_{t=0}^\infty$ is the same as the asymptotic behavior of (4.10). Therefore, it is sufficient to show that any absolutely continuous trajectory of (4.10) will converge to Π_ϵ^*.

From Lemma 4.2.2, we know that the potential function always (strictly) increases along the trajectory of (4.10) outside $\mathsf{NE}(\epsilon)$. Since the potential function is bounded[4], any absolutely continuous trajectory of (4.10) will enter the set $\mathsf{NE}(\epsilon)$ in finite time[5]. Since $\mathsf{NE}(\epsilon) \subseteq \Pi_\epsilon^*$, once the trajectory of (4.10) enters the set $\mathsf{NE}(\epsilon)$, it is already inside the set Π_ϵ^*. The proof concludes by showing that Π_ϵ^* is an invariant set. Indeed, by noting that Π_ϵ^* is a super-level set of the potential function and Π_ϵ^* contains $\mathsf{NE}(\epsilon)$, we conclude that Π_ϵ^* is an invariant set using Lemma 4.2.2. Thus, any absolutely continuous trajectory of (4.10) will enter the set Π_ϵ^* and remain inside it forever. □

Proof of Theorem 4.2.1. From Lemmas 4.2.1-4.2.4, we conclude that the policy sequence $\{\pi^t\}_{t=0}^\infty$ induced by Algorithm 3 converges to the set Π_ϵ^* .

Next, we show that for any $0 \leqslant \epsilon \leqslant \epsilon'$, $\Pi_\epsilon^* \subseteq \Pi_{\epsilon'}^*$. This follows from the fact that $\mathsf{NE}(\epsilon) \subseteq \mathsf{NE}(\epsilon')$. Thus, $\min_{\pi \in \mathsf{NE}(\epsilon')} \Phi(\mu, \pi) \leqslant \min_{\pi \in \mathsf{NE}(\epsilon)} \Phi(\mu, \pi)$. As a result, if $\pi \in \Pi_\epsilon^*$, then $\pi \in \Pi_{\epsilon'}^*$.

Finally, we show that for every $\epsilon \geqslant 0$, $\mathsf{NE}(\epsilon) \subseteq \Pi_\epsilon^*$. Suppose that $\pi \in \mathsf{NE}(\epsilon)$, then $\Phi(\mu, \pi) \geqslant \min_{\pi' \in \mathsf{NE}(\epsilon)} \Phi(\mu, \pi')$. This ensures that $\pi \in \Pi_\epsilon^*$.

4.3 Numerical Experiments

In this section, we demonstrate the performance of the proposed learning dynamics (Algorithm 3) in a Markov routing game (inspired by the example presented in [238]). Consider a parallel link network comprising of L links which is repeatedly used by I travelers (i.e. players). At every stage k, each player i picks a link $a_i^k \in [L]$ to commute. The state of the

[4]The boundedness of potential function is without loss of generality as it is shift invariant.

[5]Even though we state that any absolutely continuous trajectory of (4.10) enters $\mathsf{NE}(\epsilon)$, but it may leave that set and re-enter.

network is $s = (s_\ell)_{\ell \in [L]}$, where $s_\ell = 1$ represents that link ℓ is unsafe and $s_\ell = 0$ represents that link ℓ is safe. The probability that a particular link becomes unsafe in the next time step $k+1$ is λ_1 if the number of players using that link in stage k is larger than or equal to a threshold T, and this probability is λ_2 otherwise.

Here, we consider the common interest rewards. That is, given network state s and joint action a, the utility of any player $i \in I$ is

$$u_i(s,a) = u(s,a) = \sum_{i \in I} \sum_{\ell=1}^{L} \mathbb{1}(a_i = \ell)\Big(b_\ell - (1+s_\ell) m_\ell \sum_{j \in I} \mathbb{I}(a_j = \ell)\Big).$$

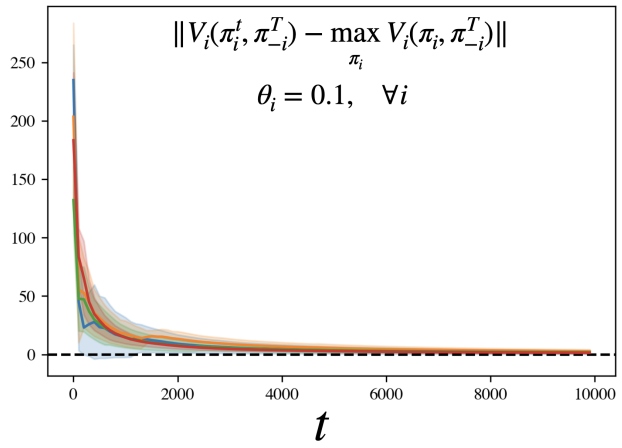
Here, b_ℓ is the fixed utility of using link $\ell \in [L]$, m_ℓ is a link-dependent constant which weights the effect of congestion on the cost and $(1+s_\ell) m_\ell \sum_{j \in I} \mathbb{I}(a_j = \ell)$ represents the cost of using ℓ given the total number of users on ℓ and the network state. The goal of every player $i \in I$ is to choose a policy $\pi_i : S \to \Delta([L])$ to maximize the long run expected discounted payoff $\mathbb{E}[\sum_{k=0}^{\infty} \gamma^k u(s^k, a^k)]$. Due to the common reward structure the game is a Markov potential game. In this example, we set $|I| = 4$, $L = 2$, $T = 2$, $\lambda_1 = 0.8$, $\lambda_2 = 0.2$, $m_1 = 2$, $b_1 = 9$, $m_2 = 4$, and $b_2 = 16$. We simulate for $T = 10^4$ stages. The step size schedule $\alpha_i(n) = 1/n^{0.5}$, $\beta_i(n) = 1/n$ for all $i \in I$. The initial state of every link is sampled uniformly randomly.

To study the Nash approximation error with respect to the converged policy, we study the function: $\|V_i(\cdot, \pi_i^t, \pi_{-i}^T) - \max_{\pi_i' \in \Pi_i} V_i(\cdot, \pi_i', \pi_{-i}^T)\|_1$, where π_i^T is the converged policy update and π_i^t is k^{th} policy update. We observe that decreasing the exploration probabilities asymptotically leads to lower Nash approximation error (Figure 4.1(a)-4.1(b)).

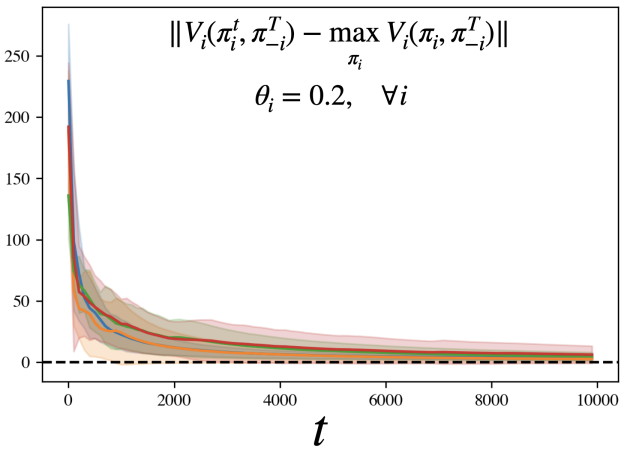
4.4 Concluding Remarks

In this chapter, we analyzed the long-run behavior of a multi-agent decentralized reinforcement learning algorithm in infinite-horizon discounted Markov potential games. We demonstrated that when agents independently employ actor-critic algorithm using only their local information, their learning processes converge to an approximate Nash equilibrium set. The size of the set shrinks if the exploration rates decrease.

In the next chapter, we extend this analysis beyond the framework of Markov potential games, exploring how these methods perform in general-sum Markov games.



(a)



(b)

Figure 4.1: Variation of Nash approximation gap during 10^4 steps of Algorithm 3. The first (resp. second) figure shows the variation with exploration probability $\theta_i = 0.1$ (resp. $\theta_i = 0.2$), for every $i \in I$. In each of the figures the four curves correspond to four players. Each curve represents the mean value of the quantity over 5 trials, and we give error margins of ± 1 standard deviation.

Chapter 5

Decentralized Learning General-Sum Markov Games

This chapter is an extension of the problem studied in Chapter 4. Recall, in Chapter 4, we characterized the asymptotic convergent set of strategies when players use actor-critic learning algorithm in a decentralized manner, using their own local information, in Markov potential games. This chapter expands the scope of that analysis to any general-sum Markov game.

Our analysis builds on a new Markov near-potential function (MNPF) framework. In this framework, given any Markov game with finite state and action space, we can construct an MNPF such that the rate of the change in MNPF with respect to an agent's policy deviation approximates the rate of change in the agent's own value function (see Definition 5.2.1), and the difference between the two rates is captured by the *closeness parameter*. In the special case of the Markov potential game, which is heavily studied in literature, the MNPF becomes the exact Markov potential function, and the closeness parameter becomes zero.

The idea of MNPF builds on the notion of near potential game in static games (introduced in [81, 80]), and is related to the concept of Markov α-potential games discussed in Chapter 2. In static games, a near potential function approximates the *absolute* change of an agent's utility with their own strategy. The paper [80] analyzed the convergence of fictitious play, showing that the convergent set of strategies is related to the closeness parameter of the near potential function. This idea was extended to Markov games as the Markov α-potential function. Our MNPF generalizes this by approximating the rate of change in agents' value functions with their policies, rather than absolute value changes (Remark 5.2.1). This generalization allows us to approximate the gradient of the value function (Lemma 5.2.1), which is essential for characterizing the convergence set of the decentralized actor-critic algorithm (Remark 5.2.2). We show that MNPFs always exist for Markov games with finite state and action sets (Proposition 5.2.2).

Similar to Chapter 4, the convergence result leverages the timescale separation in decentralized actor-critic dynamics, where agents update their local estimate of the Q-function on a faster timescale and their policies on a slower one, using only bandit feedback on state tran-

sitions and their own reward. Using two-timescale stochastic approximation theory [337], we show that the fast updates converge to the Q-function's value while treating the slow policy updates as static. The agents' policy trajectories can then be analyzed as a continuous-time dynamical system, with the MNPF acting as an approximate Lyapunov function. We prove that these trajectories converge to a level set of the MNPF, which can be viewed as a set of approximate Nash equilibria (Theorem 5.3.1). When the MNPF is Lipschitz continuous and the set of Nash equilibria is finite, the dynamics converge to the neighborhood of a single equilibrium (Theorem 5.3.2). In both theorems, the convergent set is characterized by the closeness parameter of the MNPG. We evaluate the effectiveness of our results through a numerical experiment in Section 5.4.

A schematic summarizing our approach is presented in Figure 5.1.

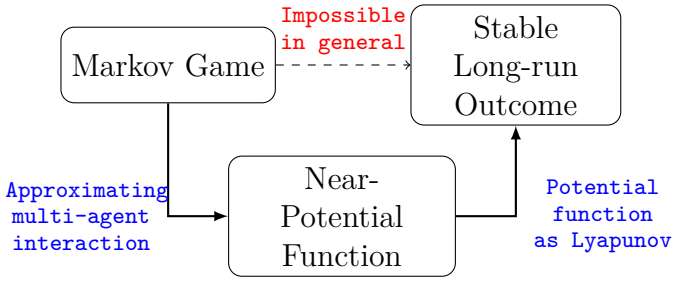


Figure 5.1: Schematic of the our approach.

Additional Related Works

Existing literature on design and analysis of decentralized learning algorithms has focused on the competitive setting represented as the two player zero-sum games (see [367, 117, 424] and references therein), the cooperative setting represented by the Markov team games (see [427, 17, 439] and references therein), and their generalizations to Markov potential games (see Chapter 4, [144] and references therein) or weakly acyclic games (see [438] and references therein). However, these studies fail to capture the complexity of large-scale multi-agent interactions in the real world, which often involves a mixture of both cooperative and competitive dynamics. Recent research has explored decentralized learning in general-sum Markov games (e.g., [283, 198, 265]), but are only concerned with convergence to weaker equilibrium concepts, such as the correlated equilibrium and coarse-correlated equilibrium, rather than reaching a Nash equilibrium.

Notations. We denote the set of all probability distributions over a set X by $\Delta(X)$. For any function $f : X \times Y \to \mathbb{R}$ we define $f(x, p) = \mathbb{E}_{y\sim q}[f(x, y)]$, where $p \in \Delta(X)$ and $q \in \Delta(Y)$. We use the notation $\mathbb{1}_X$ to denote a vector of dimension $|X|$ with all entries to be one. We use $\times_{i\in[N]}X_i$ to denote $X_1 \times X_2 \times ... \times X_N$. Unless otherwise stated, we use $\|\cdot\|$ to mean l_2-norm.

5.1 Setup

A (*general-sum*) *Markov game* $\mathcal{G}$ is given by the tuple $\langle I, S, (A_i)_{i\in I}, (u_i)_{i\in I}, P, \gamma\rangle$, where I is a finite set of players (where $|I| = N$); S is a finite set of states; A_i is a finite set of actions for each player $i \in I$, with joint action profile $a = (a_i)_{i\in I} \in A = \times_{i\in I} A_i$; $u_i : S \times A \to \mathbb{R}$ is the one-stage payoff function of player i and $u_{\max} := \max_{i\in I, s\in S, a\in A} |u_i(s,a)|$; $P = (P(s'|s,a))_{s,s'\in S, a\in A}$ is the state transition matrix and $P(s'|s,a)$ is the probability that state changes from s to s' with action profile a; and $\gamma \in [0,1)$ is the discount factor. For each player i, a *stationary Markov policy* $\pi_i : S \to \Delta(A_i)$ specifies the probability $\pi_i(s,a_i)$ of choosing action a_i in state s. The set of all stationary policies for player i is $\Pi_i := \Delta(A_i)^{|S|}$, and the *joint policy profile* of all players is $\pi := (\pi_i)_{i\in I} \in \Pi := \times_{i\in I}\Pi_i$. Similarly, $\pi_{-i} := (\pi_j)_{j\in I\setminus\{i\}}$ is the joint policy of all players except i.

The game proceeds in discrete-time stages indexed by $k \in \{0, 1, \dots\}$. At $k = 0$, the initial state s^0 is sampled from a distribution $\mu \in \Delta(S)$. At each time step k, given state $s^k \in S$, each player samples $a_i^k \sim \pi_i(s^k)$, forming the joint action profile $a^k := (a_i^k)_{i\in I}$. The next state is $s^{k+1} \sim P(\cdot|s^k, a^k)$. For an initial distribution μ and a stationary policy profile π, the expected total discounted payoff for each player $i \in I$ is $V_i(\mu,\pi) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k u_i(s^k, a^k)\right]$, where $s^0 \sim \mu$, $a^k \sim \pi(s^k)$, and $s^{k+1} \sim P(\cdot|s^k, a^k)$. We define the discounted state occupancy measure as $d_\mu^\pi(s) := (1-\gamma)\sum_{s^0\in S} \mu(s^0) \sum_{k=0}^{\infty} \gamma^k \mathbf{Pr}(s^k = s|s^0)$, with $\sum_{s\in S} d_\mu^\pi(s) = 1$.

Definition 5.1.1 (Stationary Nash equilibrium)**.** *For any $\epsilon \geqslant 0$, a policy profile $\pi^* \in \Pi$ is an* ϵ-stationary Nash equilibrium *of $\mathcal{G}$ if for any $i \in I$, any $\pi_i \in \Pi_i$, $V_i(\mu, \pi_i^*, \pi_{-i}^*) \geqslant V_i(\mu, \pi_i, \pi_{-i}^*) - \epsilon$. Any ϵ-Nash equilibrium with $\epsilon = 0$ is a* Nash equilibrium*. For any $\epsilon \geqslant 0$, we use the notation $\mathsf{NE}(\epsilon)$ to denote the set of all ϵ-stationary Nash equilibria.*

5.2 Markov Near-Potential Function

We introduce the notion of Markov near potential function which is crucial for subsequent disposition.

Definition 5.2.1 (Markov near-potential function)**.** *A bounded function $\Phi : S \times \Pi \to \mathbb{R}$ is called a* Markov near potential function *(MNPF) for a game $\mathcal{G}$ with* closeness parameter *$\kappa \geqslant 0$, if for all $s \in S$, $i \in I$, $\pi_i, \pi_i' \in \Pi_i$, and $\pi_{-i} \in \Pi_{-i}$,*

$$\left|\left(\Phi\left(s, \pi_i', \pi_{-i}\right) - \Phi\left(s, \pi_i, \pi_{-i}\right)\right) - \left(V_i\left(s, \pi_i', \pi_{-i}\right) - V_i\left(s, \pi_i, \pi_{-i}\right)\right)\right| \leqslant \kappa \|\pi_i' - \pi_i\|, \quad (5.1)$$

where the $\|\pi_i' - \pi_i\| := \sqrt{\sum_{s\in S}\sum_{a_i\in A_i}(\pi_i'(s,a_i) - \pi_i(s,a_i))^2}$.

Definition 5.2.1 ensures that the difference between the rate of change in value function of any player with respect to the unilateral change in their policy and that of the potential function is upper bounded by κ.

Remark 5.2.1. *Our MNPF framework generalizes the framework of Markov α-potential function (cf. Chapter 2). Specifically, if Φ is a Markov α-potential function for a game G, then for all $s \in S$, $i \in I$, $\pi_i, \pi_i' \in \Pi_i$, and $\pi_{-i} \in \Pi_{-i}$,*

$$| \Big(\Phi(s, \pi_i', \pi_{-i}) - \Phi(s, \pi_i, \pi_{-i})\Big) - \Big(V_i(s, \pi_i', \pi_{-i}) - V_i(s, \pi_i, \pi_{-i})\Big) | \leqslant \alpha, \tag{5.2}$$

which requires that the difference between an agent's value function change and the Markov α-potential function change due to a unilateral policy shift is uniformly bounded *by α. In contrast, our Definition 5.2.1 bounds this difference based on the magnitude of policy changes. Comparing* (5.1) *and* (5.2)*, if Φ is an MNPF for $\mathcal{G}$ with parameter κ, then $\mathcal{G}$ is a Markov α-potential game with $\alpha \leqslant \kappa\sqrt{2|S|}$.*

Remark 5.2.2. *Our Markov near-potential function framework* (5.1) *offers an advantage over the Markov α-potential function* (5.2) *by quantifying the gap between each agent's value function gradient and that of a potential function (Lemma 5.2.1). This property is essential for characterizing the convergent set of the decentralized actor-critic algorithm, as demonstrated in Theorems 5.3.1 and 5.3.2. To establish convergence, we show a positive correlation between policy update rates and the gradient of agents' value functions. Since this gradient closely approximates that of the potential function, with the closeness parameter defining the gap, we prove that the potential function consistently increases outside a neighborhood of the Nash equilibrium. This neighborhood, which defines the convergence set, varies in size based on the closeness parameter.*

Next, we show that the gradient of MNPF provides an approximation of gradient of value function of players in their local strategies.

Lemma 5.2.1. *For any Markov game G and an associated MNPF Φ with closeness parameter κ, it holds that*

$$|v_i^\top \frac{\partial \Phi(\mu, \pi)}{\partial \pi_i} - v_i^\top \frac{\partial V_i(\mu, \pi)}{\partial \pi_i}| \leqslant \kappa \|v_i\|_2, \quad \forall i \in I, v_i \in \mathbb{R}^{|S||A_i|}.$$

Proof. For any $v_i \in \mathbb{R}^{|S|\cdot|A_i|}$ it holds that

$$\begin{aligned} v_i^\top \frac{\partial \Phi(\mu, \pi)}{\partial \pi_i} &= \lim_{h \to 0} \frac{\Phi(\mu, \pi_i + h v_i, \pi_{-i}) - \Phi(\mu, \pi_i, \pi_{-i})}{h}, \\ v_i^\top \frac{\partial V_i(\mu, \pi)}{\partial \pi_i} &= \lim_{h \to 0} \frac{V_i(\mu, \pi_i + h v_i, \pi_{-i}) - V_i(\mu, \pi_i, \pi_{-i})}{h}. \end{aligned} \tag{5.3}$$

Additionally, using the definition of MNPF, we obtain

$$\begin{aligned} & V_i(\mu, \pi_i + h v_i, \pi_{-i}) - V_i(\mu, \pi_i, \pi_{-i}) - \kappa h \|v_i\|_2 \\ \leqslant\ & \Phi(\mu, \pi_i + h v_i, \pi_{-i}) - \Phi(\mu, \pi_i, \pi_{-i}) \\ \leqslant\ & V_i(\mu, \pi_i + h v_i, \pi_{-i}) - V_i(\mu, \pi_i, \pi_{-i}) + \kappa h \|v_i\|_2. \end{aligned} \tag{5.4}$$

Dividing everything by h in (5.4) and using (5.3), we obtain

$$|v_i^\top \frac{\partial \Phi(\mu,\pi)}{\partial \pi_i} - v_i^\top \frac{\partial V_i(\mu,\pi)}{\partial \pi_i}| \leqslant \kappa \|v_i\|_2, \quad \forall i \in I, v_i \in \mathbb{R}^{|S|\cdot|A_i|}.$$

This concludes the proof. □

The following proposition shows that a near-potential function can be constructed for any Markov game, and the approximate optimal solution of the near-potential function is an approximate Nash of the original game.

Proposition 5.2.2. *For any game $\mathcal{G}$, there exists a tuple (Φ, κ) such that Φ is a MNPF of $\mathcal{G}$ with closeness parameter κ. Furthermore, for any $\epsilon > 0$ and any $\pi^* \in \Pi$ such that $\Phi(s,\pi^*) \geqslant \sup_{\pi\in\Pi} \Phi(s,\pi) - \epsilon$ for all $s \in S$, π^* is a $(\kappa\sqrt{2|S|} + \epsilon)-$stationary Nash equilibrium.*

Proof. For any game G, we set $\Phi(s,\pi) = 0$ for every $s \in S, \pi \in \Pi$. We note that

$$\begin{aligned}
&|V_i(s,\pi_i',\pi_{-i}) - V_i(s,\pi_i,\pi_{-i})| \\
\overset{(a)}{\leqslant} &\frac{1}{1-\gamma} \sum_{s'\in S, a_i \in A_i} \left| d_\mu^\pi(s')(\pi_i(s',a_i) - \pi_i'(s',a_i)) \right| \cdot |Q_i(s',a_i;\pi')| \\
\overset{(b)}{\leqslant} &\frac{u_{\max}}{(1-\gamma)^2} \|\pi_i - \pi_i'\|,
\end{aligned}$$

where *(a)* is due to multi-agent performance difference lemma (see Lemma D.1.5), and *(b)* is due to the fact that for any $\pi \in \Pi$,

$$\max_{s,a_i} |Q_i(s,a_i;\pi)| \leqslant \frac{u_{\max}}{(1-\gamma)}$$

and $d_\mu^\pi(s) \in [0,1]$ for every $s \in S$. Thus, Φ is a MNPF for $\mathcal{G}$ with the closeness-parameter $\kappa = \frac{u_{\max}}{(1-\gamma)^2}$.

Next, we show that for any $\epsilon > 0$ and any $\pi^* \in \Pi$ such that $\Phi(s,\pi^*) \geqslant \sup_{\pi\in\Pi} \Phi(s,\pi) - \epsilon$ for all $s \in S$, and thus π^* is a $(\kappa\sqrt{2|S|} + \epsilon)-$stationary Nash equilibrium. Particularly, using (5.1), we observe that for any $i \in I, \pi_i' \in \Pi_i$,

$$\begin{aligned}
&V_i(\mu,\pi^*) - V_i(\mu,\pi_i',\pi_{-i}^*) \\
\geqslant &\Phi_i(\mu,\pi^*) - \Phi_i(\mu,\pi_i',\pi_{-i}^*) - \kappa\|\pi_i' - \pi_i^*\| \geqslant -\epsilon - \kappa\sqrt{2|S|},
\end{aligned}$$

where we note that $\|\pi_i' - \pi_i^*\| \leqslant \sqrt{2|S|}$ using Cauchy Schwartz inequality. □

A game $\mathcal{G}$ may be associated with multiple MNPF with different κ. Proposition 5.2.2 suggests that an MNPF with a smaller closeness parameter κ is a better approximation of the original game in that the optimum of the potential function is a closer approximation of the Nash equilibrium in the original game. Following [171], we can compute the MNPF with the smallest closeness parameter of a game as a semi-infinite linear program. We omit those details here for concise presentation.

5.3 Decentralized Actor-Critic Algorithm

In this section, we restate the decentralized learning algorithm discussed in Chapter 4. Recall, in such dynamics each player makes decisions based solely on the current state information and their local reward feedback. Players do not need any knowledge about other players. Using MNPF, we provide theoretical guarantees on the long-run outcomes of the algorithm.

For every $i \in I, s \in S, a_i \in A_i, \pi \in \Pi$, we define *Q-function* as

$$Q_i(s, a_i; \pi) := u_i(s, a_i, \pi_{-i}) + \gamma \sum_{s' \in S} P(s'|s, a_i, \pi_{-i}) V_i(s', \pi),$$

which is player i's expected long-horizon discounted utility when the game starts in state s and they play action a_i in the first stage and then employs policy π_i from the second stage onwards, and other players always employ policy π_{-i}. With slight abuse of notation, we define $Q_i(s; \pi) = (Q_i(s, a_i; \pi))_{a_i \in A_i} \in \mathbb{R}^{|A_i|}$. Furthermore, given Q_i and policy $\pi \in \Pi$, we define the *optimal one-stage deviation* of player i in state $s \in S$ as

$$\mathrm{br}_i(s; \pi) = \arg\max_{\hat{\pi}_i \in \Delta(A_i)} \hat{\pi}_i^\top Q_i(s; \pi). \tag{5.5}$$

Decentralized Learning Algorithm and Preliminaries

We study the discrete-time decentralized learning algorithm proposed in Algorithm 3 in Chapter 4, where each player adopts an actor-critic algorithm in a decentralized manner. In each iterate t of the algorithm, every player $i \in I$ updates the following quantities: *(i)* the counters $n^t = (n^t(s))_{s \in S}$ and $\tilde{n}_i^t = (\tilde{n}_i^t(s, a_i))_{s \in S, a_i \in A_i}$, which keep track of the number of visits of all states and all state-action pairs up to the current iteration; *(ii)* their estimate of the local Q-functions q_i^t, which is updated as a linear combination of the previous estimate and a new estimate based on the realized one-stage reward and the long-horizon discounted value from the next state as estimated from the q-function estimate and policy from previous iterate (refer (5.6)); and *(iii)* their local policies π_i^t, which is updated as a linear combination of the policy in the previous iterate, and player i's optimal one-stage deviation (refer (5.7)). Finally, every player samples an action $a_i^t \sim \pi_i^t$ with probability $(1 - \theta)$ and from the uniform distribution over their action set A_i with probability θ, where $\theta \in (0, 1)$ is the exploration parameter[1].

Remark 5.3.1. *The updates in Algorithm 4 are* asynchronous *because each player updates the counters, q-function estimate, and policy* only *for the most recently visited state-(local) action pair, rather than for all state-action pairs.*

Next, we state assumptions that are central to analyze the convergence of Algorithm 4.

[1]For the sake of concise notations, we take the exploration rate to be same for all the players, but the analysis in this chapter extends to the scenario when it is heterogeneous.

Algorithm 4 Decentralized Learning Algorithm

Initialization: $n^0(s) = 0, \forall s \in S$; $\tilde{n}_i^0(s, a_i) = 0, \tilde{q}_i^0(s, a_i) = 0$, $\pi_i^0(s) = 1/|A_i|$, $\forall (i, a_i, s)$, and $\theta \in (0, 1)$. In stage 0, each player observes s^0, choose their action $a_i^0 \sim \pi_i^0(s^0)$, and observe $u_i^0 = u_i(s^0, a^0)$.

In every iterate $t = 1, 2, ...,$ each player observes s^t, and independently updates $\{n_i^t, \tilde{n}_i^t, \tilde{q}_i^t, \pi_i^t\}$.

Update $n^t, \tilde{n}_i^t$: $n^t(s^{t-1}) = n^{t-1}(s^{t-1}) + 1, \tilde{n}_i^t(s^{t-1}, a_i^{t-1}) = \tilde{n}_i^{t-1}(s^{t-1}, a_i^{t-1}) + 1$.

Update $\tilde{q}_i^t$: Using the one-stage reward u_i^{t-1}, update

$$\begin{aligned} \tilde{q}_i^t(s^{t-1}, a_i^{t-1}) &= \tilde{q}_i^{t-1}(s^{t-1}, a_i^{t-1}) + \alpha(\tilde{n}_i^t(s^{t-1}, a_i^{t-1})) \\ &\cdot \left(r_i^{t-1} + \gamma \pi_i^{t-1}(s^t)^\top \tilde{q}_i^{t-1}(s^t) - \tilde{q}_i^{t-1}(s^{t-1}, a_i^{t-1}) \right). \end{aligned} \tag{5.6}$$

Update π_i^t**:** Pick $\widehat{\mathsf{br}}_i \in \arg\max_{\pi_i \in \Delta(A_i)} \pi_i^\top q_i^{t-1}(s^{t-1})$ and set

$$\pi_i^t(s^{t-1}) = \pi_i^{t-1}(s^{t-1}) + \beta(n^t(s^{t-1})) \cdot (\widehat{\mathsf{br}}_i - \pi_i^{t-1}(s^{t-1})). \tag{5.7}$$

Sample action and observe reward:

$$a_i^t \sim (1 - \theta)\pi_i^t(s^t) + \theta \cdot (1/|A_i|)\mathbb{1}_{A_i}. \tag{5.8}$$

Each player observes their own reward $u_i^t = u_i(s^t, a^t)$.

Assumption 5.3.1. *(a) The initial state distribution $\mu(s) > 0$ for all $s \in S$.*
Additionally, $\min_{s,s' \in S, a \in A} P(s'|s, a) > 0$.
(b) The step-sizes satisfy

(i) $\sum_{n=0}^{\infty} \alpha(n) = \infty, \sum_{n=0}^{\infty} \beta(n) = \infty$, $\lim_{n\to\infty} \alpha(n) = \lim_{n\to\infty} \beta(n) = 0$;

(ii) There exist some $q, q' \geqslant 2$, $\sum_{n=0}^{\infty} \alpha(n)^{1+q/2} < \infty$ and $\sum_{n=0}^{\infty} \beta(n)^{1+q'/2} < \infty$;

(iii) $\sup_n \alpha([xn])/\alpha(n) < \infty$, $\sup_n \beta([xn])/\beta(n) < \infty$ *for all $x \in (0, 1)$, where $[xn]$ denotes the largest integer less than or equal to xn. Additionally, $\{\alpha(n)\}, \{\beta(n)\}$ are non-increasing in n;*

(iv) $\lim_{n\to\infty} \beta(n)/\alpha(n) = 0$.

Assumption 5.3.1-(a) is a standard assumption to ensure ergodicity of the Markov state transition for learning Q-functions. Additionally, Assumption 5.3.1-(b) is standard assumption on step sizes in actor-critic algorithms [275].

To study the convergent set of Algorithm 4, we apply two-timescale asynchronous stochastic approximation (TTASA) theory (see Section C.1). The discrete-time updates in Algorithm 4, along with Assumption 5.3.1, satisfy the conditions in Section C.1, as noted in

Chapter 4. TTASA theory ensures two things: First, the (fast) q-function estimates asymptotically track the Q-functions of the current policy. Using Assumption 5.3.1 and the contraction property of temporal difference operator, we observe that

$$\lim_{t\to\infty} \|\tilde{q}_i^t(s) - Q_i(s; \pi_i^t, \pi_{-i}^{t,\theta})\|_\infty = 0$$

holds with probability 1 for all $s \in S$ and $i \in I$, where

$$\pi_{-i}^{t,\theta}(s) := (1-\theta)\pi_{-i}^t + \theta(1/|A_i|) \cdot \mathbb{1}_{A_i}.$$

Second, the convergent set of policy updates is the same as the convergent limit of any absolutely continuous trajectory of the following differential inclusion:

$$\frac{d}{d\tau}\varpi_i^\tau(s) \in \bar{\eta}(s)\left(\mathsf{br}_i(s; \varpi_i^\tau, \varpi_{-i}^{\tau,\theta}) - \varpi_i^\tau(s)\right), \tag{5.9}$$

where $\tau \in [0, \infty)$ is a continuous-time index,

$$\varpi_{-i}^{\tau,\theta}(s) := (1-\theta)\varpi_{-i}^\tau(s) + \theta(1/|A_{-i}|) \cdot \mathbb{1}_{A_{-i}}, \quad \forall\, s \in S, i \in I,$$

and $\bar{\eta}(s) \in [\eta, 1]$ for some positive scalar η that depends on the ergodicity of the probability transition function.

Convergence Guarantees

We now present the first main result of this paper, which characterizes the convergent set of policy updates in Algorithm 4 in terms of the superlevel set of a MNPF over the set of approximate Nash equilibria. This characterization is based on the closeness parameter κ associated with the MNPF.

Theorem 5.3.1. *Consider a Markov game $\mathcal{G}$ and an associated MNPF $\mathbf{\Phi}$ with closeness-parameter κ. Under Assumptions 5.3.1, the sequence of policies $\{\pi^t\}_{t=0}^\infty$ induced by Algorithm 4, with the exploration parameter*

$$\theta \leqslant \lambda \frac{\sqrt{2|S|}(1-\gamma)^2}{4u_{\max}\left(\frac{1}{(1-\gamma)} + \frac{\eta}{D}\right)}, \tag{5.10}$$

converge almost surely to the set

$$\Lambda := \left\{\pi : \Phi(\mu, \pi) \geqslant \min_{y \in \mathsf{NE}(\Theta(\kappa+\lambda))} \Phi(\mu, y)\right\},$$

where λ is a positive scalar,

$$\Theta := DN^2\sqrt{2|S|}/\eta, D = \frac{1}{1-\gamma}\max_{i,\pi_{-i},s}\left|d_\mu^{\pi_i^\dagger,\pi_{-i}}(s)/\mu(s)\right|, \pi_i^\dagger \in \arg\max_{\pi_i \in \Pi_i} V_i(\mu, \pi_i, \pi_{-i}),$$

and η is a positive scalar that depends on the ergodicity of the probability transition function.

Proof. Following Section 5.3, it is sufficient to show that every absolutely continuous trajectory of (5.9) converges to the set Λ.

We construct a Lyapunov function candidate $\phi : [0, \infty) \to \mathbb{R}$ as

$$\phi(\tau) = \max_{\varpi \in \Pi} \Phi(\mu, \varpi) - \Phi(\mu, \varpi^\tau),$$

which is the difference of the MNPF at its maximizer with that of its value at ϖ^τ. The key step in the proof is to show that $\phi(\tau)$ is weakly decreasing in τ as long as $\varpi^\tau \notin \mathsf{NE}(\Theta(\kappa+\lambda))$. We claim that it is sufficient to establish that for any ϖ^τ that is an $\epsilon-$stationary Nash equilibrium, the following equation holds

$$\frac{d\phi(\tau)}{d\tau} \leqslant (\kappa+\lambda)N^2\sqrt{2|S|} - \frac{\eta}{D}\epsilon = \frac{\eta}{D}\Theta(\kappa+\lambda) - \frac{\eta}{D}\epsilon. \tag{5.11}$$

Indeed, if (5.11) holds, then for any $\epsilon > \Theta(\kappa+\lambda)$, $\phi(\tau)$ decreases at a rate $\eta/D \cdot (\epsilon - \Theta(\kappa+\lambda))$. Since ϕ is bounded, any absolutely continuous trajectory of (5.9) will enter the set $\mathsf{NE}(\Theta(\kappa+\lambda))$ in finite time, starting from any initial policy. Subsequently, even if trajectories leave this set, the function $\Phi(\mu, \cdot)$ cannot decrease below $\min_{\pi \in \mathsf{NE}(\Theta(\kappa+\lambda))} \Phi(\mu, \pi)$, and once the trajectory leaves this set the potential function will always increase (as $d\phi(\tau)/d\tau < 0$). Thus, it only remains to show that (5.11) hold. Towards that goal, we note that

$$\begin{aligned}
\frac{d}{d\tau}\phi(\tau) &= -\sum_{i\in I, s\in S}\left(\frac{\partial \Phi(\mu,\varpi^\tau)}{\partial \varpi_i(s)}\right)^\top \frac{d\varpi_i^\tau(s)}{d\tau} \\
&= \sum_{i\in I, s\in S}\left(\frac{\partial V_i(\mu,\varpi^\tau)}{\partial \varpi_i(s)} - \frac{\partial \Phi(\mu,\varpi^\tau)}{\partial \varpi_i(s)}\right)^\top \frac{d\varpi_i^\tau(s)}{d\tau} \\
&\quad - \sum_{i\in I, s\in S}\left(\frac{\partial V_i(\mu,\varpi^\tau)}{\partial \varpi_i(s)}\right)^\top \frac{d\varpi_i^\tau(s)}{d\tau} \\
&\overset{(i)}{\leqslant} \kappa \sum_{i\in I}\sqrt{\sum_{s\in S, a_i\in A_i}(\mathsf{br}_i(s,a_i;\varpi_i^\tau,\varpi_i^{\tau,\theta}) - \varpi_i^\tau(s,a_i))^2} \\
&\quad + \sum_{i\in I, s\in S}\frac{d_\mu^{\varpi^\tau}(s)}{\gamma-1}\bar{\eta}(s)Q_i(s;\varpi^\tau)^\top\left(\widetilde{\mathrm{br}}_i^{\tau,\theta}(s) - \varpi_i^\tau(s)\right), \\
&\leqslant \kappa N\sqrt{|S|\sum_{a_i\in A_i}\left(\mathsf{br}_i(s,a_i;\varpi_i^\tau,\varpi_i^{\tau,\theta}) + \varpi_i^\tau(s,a_i)\right)} \\
&\quad + \sum_{i\in I, s\in S}\frac{d_\mu^{\varpi^\tau}(s)}{\gamma-1}\bar{\eta}(s)Q_i(s;\varpi^\tau)^\top\left(\widetilde{\mathrm{br}}_i^{\tau,\theta}(s) - \varpi_i^\tau(s)\right), \\
&= \kappa N\sqrt{2|S|} \\
&\quad + \sum_{i\in I, s\in S}\frac{d_\mu^{\varpi^\tau}(s)}{\gamma-1}\bar{\eta}(s)Q_i(s;\varpi^\tau)^\top\left(\widetilde{\mathrm{br}}_i^{\tau,\theta}(s) - \varpi_i^\tau(s)\right),
\end{aligned} \tag{5.12}$$

where $\widetilde{\mathrm{br}}_i^{\tau,\theta}(s) \in \mathrm{br}_i(s; \varpi_i^\tau, \varpi_{-i}^{\tau,\theta})$, and (i) is due to Lemma 5.2.1. Next, by adding and subtracting the term $\sum_{i\in I, s\in S} \frac{d_\mu^{\varpi^\tau}(s)}{\gamma-1}\bar{\eta}(s)Q_i(s;\varpi_i^\tau,\varpi_{-i}^{\tau,\theta})^\top \left(\widetilde{\mathrm{br}}_i^{\tau,\theta}(s) - \varpi_i^\tau(s)\right)$ on the RHS in (5.12), we obtain

$$\begin{aligned}\frac{d\phi(\tau)}{d\tau} \leqslant\ & \kappa N\sqrt{2|S|} \\ & + \underbrace{\sum_{i\in I, s\in S} \frac{d_\mu^{\varpi^\tau}(s)}{\gamma-1}\bar{\eta}(s)(Q_i(s;\varpi^\tau) - Q_i(s;\varpi_i^\tau,\varpi_{-i}^{\tau,\theta}))^\top \left(\widetilde{\mathrm{br}}_i^{\tau,\theta}(s) - \varpi_i^\tau(s)\right)}_{\texttt{Term 1}} \\ & + \underbrace{\sum_{i\in I, s\in S} \frac{d_\mu^{\varpi^\tau}(s)}{\gamma-1}\bar{\eta}(s)Q_i(s;\varpi_i^\tau,\varpi_{-i}^{\tau,\theta})^\top \left(\widetilde{\mathrm{br}}_i^{\tau,\theta}(s) - \varpi_i^\tau(s)\right)}_{\texttt{Term 2}}, \end{aligned} \tag{5.13}$$

First, we bound `Term 1` in the above equation. Note that

$$\begin{aligned}\texttt{Term 1} &\leqslant \frac{1}{1-\gamma}\sum_{i\in I}\max_{s\in S}\left|(Q_i(s;\varpi^\tau) - Q_i(s;\varpi_i^\tau,\varpi_{-i}^{\tau,\theta}))^\top \left(\widetilde{\mathrm{br}}_i^{\tau,\theta}(s) - \varpi_i^\tau(s)\right)\right| \\ &\leqslant \frac{2}{1-\gamma}\sum_{i\in I}\max_{s\in S, a_i\in A_i}|(Q_i(s,a_i;\varpi^\tau) - Q_i(s,a_i;\varpi_i^\tau,\varpi_{-i}^{\tau,\theta}))| \\ &\leqslant \frac{4\theta N^2}{(1-\gamma)^3}u_{\max}, \end{aligned} \tag{5.14}$$

where the last inequality is due to Lemma D.1.4. Next, we analyze `Term 2` in (5.13). Using (5.5), it holds that, for every $i \in I$,

$$Q_i(s;\varpi_i^\tau,\varpi_{-i}^{\tau,\theta})^\top \left(\widetilde{\mathrm{br}}_i^{\tau,\theta}(s) - \varpi_i^\tau(s)\right) \geqslant 0.$$

Furthermore, given the fact that $\bar{\eta}(s) > \eta$ and $d_\mu^{\varpi^\tau}(s) \geqslant 0$ for all $i \in I, s \in S$, we bound

$$\texttt{Term 2} \leqslant -\frac{\eta}{1-\gamma}\sum_{i,s} d_\mu^{\varpi^\tau}(s)Q_i(s;\varpi_i^\tau,\varpi_{-i}^{\tau,\theta})^\top \left(\widetilde{\mathrm{br}}_i^{\tau,\theta}(s) - \varpi_i^\tau(s)\right). \tag{5.15}$$

We claim that

$$\begin{aligned}&\sum_{i,s} d_\mu^{\varpi^\tau}(s)Q_i(s;\varpi_i^\tau,\varpi_{-i}^{\tau,\theta})^\top \left(\widetilde{\mathrm{br}}_i^{\tau,\theta}(s) - \varpi_i^\tau(s)\right) \\ &\geqslant 1/D \cdot \sum_{i,s} d_\mu^{\pi_i^\dagger,\varpi_{-i}^{\tau,\theta}}(s)Q_i(s;\varpi_i^\tau,\varpi_{-i}^{\tau,\theta})^\top \left(\pi_i^\dagger(s) - \varpi_i^\tau(s)\right), \end{aligned} \tag{5.16}$$

where $\pi_i^\dagger \in \arg\max_{\pi_i \in \Pi_i} V_i(\mu, \pi_i, \varpi_{-i}^\tau)$ be a best response[2] of player i, given that the strategy of other players is ϖ_{-i}^τ. Indeed, note that

$$\begin{aligned}
&\sum_{i,s} d_\mu^{\pi_i^\dagger, \varpi_{-i}^{\tau,\theta}}(s) Q_i(s; \varpi_i^\tau, \varpi_{-i}^{\tau,\theta})^\top \left(\pi_i^\dagger(s) - \varpi_i^\tau(s)\right) \\
&\leqslant \sum_{i,s} d_\mu^{\pi_i^\dagger, \varpi_{-i}^{\tau,\theta}}(s) \max_{\hat{\pi}_i \in \Delta(A_i)} Q_i(s; \varpi_i^\tau, \varpi_{-i}^{\tau,\theta})^\top \left(\hat{\pi}_i(s) - \varpi_i^\tau(s)\right) \\
&\leqslant \sum_{i,s} d_\mu^{\varpi^\tau}(s) \left\| \frac{d_\mu^{\pi_i^\dagger, \varpi_{-i}^{\tau,\theta}}}{d_\mu^{\varpi^\tau}} \right\|_\infty \cdot \max_{\hat{\pi}_i \in \Delta(A_i)} Q_i(s; \varpi_i^\tau, \varpi_{-i}^\tau)^\top \left(\hat{\pi}_i(s) - \varpi_i^\tau(s)\right) \\
&\overset{(i)}{\leqslant} D \sum_{i,s} d_\mu^{\varpi^\tau}(s) \cdot \max_{\hat{\pi}_i \in \Delta(A_i)} Q_i(s; \varpi_i^\tau, \varpi_i^{\tau,\theta})^\top \left(\hat{\pi}_i(s) - \varpi_i^\tau(s)\right) \\
&= D \sum_{i,s} d_\mu^{\varpi^\tau}(s) \cdot Q_i(s; \varpi_i^\tau, \varpi_{-i}^{\tau,\theta})^\top \left(\widetilde{\mathrm{br}}_i^{\tau,\theta}(s) - \varpi_i^\tau(s)\right),
\end{aligned} \tag{5.17}$$

where *(i)* is due to the fact that $d_\mu^{\varpi^{\tau,\theta}}(s) \geqslant (1-\gamma)\mu(s)$ along with the definition of D.

Using (5.16) in (5.15), we obtain

$$\begin{aligned}
&\texttt{Term 2} \\
&\leqslant -\frac{\eta}{D(1-\gamma)} \sum_{i,s} d_\mu^{\pi_i^\dagger, \varpi_{-i}^{\tau,\theta}}(s) \cdot \max_{\hat{\pi}_i \in \Delta(A_i)} Q_i(s; \varpi_i^\tau, \varpi_{-i}^{\tau,\theta})^\top \left(\hat{\pi}_i(s) - \varpi_i^\tau(s)\right) \\
&\leqslant -\frac{\eta}{D(1-\gamma)} \sum_{i,s} d_\mu^{\pi_i^\dagger, \varpi_{-i}^{\tau,\theta}}(s) Q_i(s; \varpi_i^\tau, \varpi_{-i}^{\tau,\theta})^\top \left(\pi_i^\dagger(s) - \varpi_i^\tau(s)\right),
\end{aligned} \tag{5.18}$$

where the last inequality is because $\pi_i^\dagger \in \Delta(A_i)$. Finally, note that

$$\begin{aligned}
&\sum_{i,s} d_\mu^{\pi_i^\dagger, \varpi_{-i}^{\tau,\theta}}(s) Q_i(s; \varpi_i^\tau, \varpi_{-i}^{\tau,\theta})^\top \left(\pi_i^\dagger(s) - \varpi_i^\tau(s)\right) \\
&= \sum_{i,s} d_\mu^{\pi_i^\dagger, \varpi_{-i}^\tau}(s) \left(Q_i(s; \varpi_i^\tau, \varpi_{-i}^{\tau,\theta}) - V_i(s, \varpi_i^\tau, \varpi_{-i}^{\tau,\theta})\right)^\top \left(\pi_i^\dagger(s) - \varpi^\tau(s)\right) \\
&= (1-\gamma) \sum_i V_i(\mu, \pi_i^\dagger, \varpi_{-i}^{\tau,\theta}) - V_i(\mu, \varpi_i^\tau, \varpi_{-i}^{\tau,\theta}),
\end{aligned} \tag{5.19}$$

where the last inequality is due to multi-agent performance difference lemma (Lemma D.1.5).

[2]Note that $\pi_i^\dagger$ maximizes the total payoff instead of just maximizing the payoff of one-stage deviation. Therefore, $\pi_i^\dagger$ is different from the optimal one-stage deviation policy. We drop the dependence of $\pi_i^\dagger$ on ϖ_{-i}^τ for notational simplicity.

Combining (5.18) and (5.19), we obtain

$$\begin{aligned}
\texttt{Term 2} \leqslant & -\frac{\eta}{D}\sum_i \left(V_i(\mu,\pi_i^\dagger,\varpi_{-i}^{\tau,\theta}) - V_i(\mu,\varpi_i^\tau,\varpi_{-i}^{\tau,\theta})\right) \\
\leqslant & -\frac{\eta}{D}\sum_i \left(V_i(\mu,\pi_i^\dagger,\varpi_{-i}^{\tau}) - V_i(\mu,\varpi_i^\tau,\varpi_{-i}^{\tau})\right) \\
& + \frac{2\eta}{D}\max_{\pi_i\in\Pi_i}\sum_i |V_i(\mu,\pi_i,\varpi_{-i}^{\tau,\theta}) - V_i(\mu,\pi_i^\dagger,\varpi_{-i}^{\tau})| \\
\leqslant & -\frac{\eta}{D}\sum_i \left(V_i(\mu,\pi_i^\dagger,\varpi_{-i}^{\tau}) - V_i(\mu,\varpi_i^\tau,\varpi_{-i}^{\tau})\right) + \frac{4\eta\theta N^2}{D(1-\gamma)^2}u_{\max}, \qquad (5.20)
\end{aligned}$$

where the last inequality is due to Lemma D.1.3. Using (5.14) and (5.20) in (5.13), we obtain

$$\begin{aligned}
\frac{d\phi(\tau)}{d\tau} \leqslant & \kappa N\sqrt{2|S|} + \frac{4\theta N^2}{(1-\gamma)^3}u_{\max} + \frac{4\eta\theta N^2}{D(1-\gamma)^2}u_{\max} \\
& - \frac{\eta}{D}\sum_i \left(V_i(\mu,\pi_i^\dagger,\varpi_{-i}^{\tau}) - V_i(\mu,\varpi_i^\tau,\varpi_{-i}^{\tau})\right),
\end{aligned}$$

where we used the fact that $\eta \leqslant 1$. Since

$$\theta \leqslant \lambda\frac{\sqrt{2|S|}(1-\gamma)^2}{4u_{\max}\left(\frac{1}{(1-\gamma)}+\frac{\eta}{D}\right)},$$

it ensures that

$$\frac{d\phi(\tau)}{d\tau} \leqslant (\kappa+\lambda)N^2\sqrt{2|S|} - \frac{\eta}{D}\sum_i \left(V_i(\mu,\pi_i^\dagger,\varpi_{-i}^{\tau}) - V_i(\mu,\varpi_i^\tau,\varpi_{-i}^{\tau})\right).$$

From the definition of best response, $V_i(\mu,\pi_i^\dagger,\varpi_{-i}^\tau) \geqslant V_i(\mu,\pi_i,\varpi_{-i}^\tau)$ for every $i \in I$. Furthermore, if ϖ^τ is not an $\epsilon-$Nash equilibrium then there exists a player $i \in I$ and a policy π_i such that $V_i(\mu,\pi_i,\varpi_{-i}^\tau) - V_i(\mu,\varpi^\tau) \geqslant -\epsilon$. Therefore,

$$\frac{d\phi(\tau)}{d\tau} \leqslant (\kappa+\lambda)N^2\sqrt{2|S|} - \frac{\eta}{D}\epsilon.$$

This proves (5.11) and concludes the proof. □

Remark 5.3.2. *Theorem 5.3.1 can be viewed as a generalization of Theorem 4.2.1. In particular, when players have homogeneous exploration rates and the setting corresponds to a Markov potential game (i.e., $\kappa = 0$), Theorem 5.3.1 recovers the result of Theorem 4.2.1. Setting $\lambda = \frac{\epsilon\eta}{DN^2\sqrt{2|S|}}$ establishes the equivalence between the two.*

Next, we show that when game $\mathcal{G}$ has a finite number of equilibria and the function $\Phi(\mu, \cdot)$ is Lipschitz (Assumption 5.3.2), Theorem 5.3.1 can be strengthened. Specifically, we show that the learning dynamics converge to neighborhood of equilibrium set instead of just converging to a level set of the MNPF.

Assumption 5.3.2. *The equilibrium set is finite* $\mathsf{NE}(0) = \{\pi^{*1}, \pi^{*2}, ..., \pi^{*K}\}$.

Assumption 5.3.2 has been adopted in previous literature in MARL (e.g. [144]). In fact, [125] has shown that Assumption 5.3.2 holds generically. For any $\delta >$, recall that $NE(\delta)$ is the set of approximate Nash equilibrium with approximation parameter δ. We define $\Gamma(\delta)$ as the maximum distance between an approximate equilibrium in $NE(\delta)$ and the equilibrium set.

$$\Gamma(\delta) := \max_{\pi \in \mathsf{NE}(\delta)} \min_{k \in [K]} \|\pi - \pi^{*k}\|. \tag{5.21}$$

Since $\mathsf{NE}(0)$ denotes the set of Nash equilibria, $\Gamma(0) = 0$.

Theorem 5.3.2. *Consider a Markov game* $\mathcal{G}$ *and an associated MNPF* Φ *with closeness-parameter* κ*. Suppose Assumptions 5.3.1-5.3.2 hold. There exists* $\bar{\kappa}, \bar{\epsilon}$ *such that if* $\kappa + \lambda \leqslant \bar{\kappa}$*, for some* $\lambda > 0$*, the sequence of policies* $\{\pi^t\}_{t=0}^{\infty}$ *induced by Algorithm 4, with the exploration parameter* (5.10)*, converge almost surely to the set* $\tilde{\Lambda} := \left\{\pi | \exists\ k \in [K] :\ \|\pi - \pi^{*k}\| \leqslant \chi\right\}$,

$$\chi := \min_{0 \leqslant \epsilon \leqslant \bar{\epsilon}} \Gamma(\epsilon + \Theta(\kappa + \lambda)) + (2DLN\sqrt{|S|}\Gamma(\Theta(\kappa + \lambda)))/(\eta\epsilon), \tag{5.22}$$

λ *is a positive scalar,* Θ, D, η *are defined in Theorem 5.3.1, and* L *is defined in Lemma D.1.1.*

Remark 5.3.3. *In* (5.22)*,* χ *is weakly increasing with* κ *and* λ*, as* $\Gamma(\cdot)$ *is weakly increasing and upper-semicontinuous (see Lemma D.1.2). If the game is a Markov potential game (i.e.,* $\kappa = 0$*) and players have access to exact Q-functions, then they do not need to explore (i.e.,* $\lambda = 0$*). In this case, the policy updates in* (5.7)*, evaluated at the exact Q-functions, converge to a Nash equilibrium. Indeed, with* $\kappa = 0$ *and* $\lambda = 0$*, one can follow the same steps as in the proof of Theorem 5.3.2 to show that the second term in the optimization problem* (5.22) *does not depend on* ε *since* $\Gamma(0) = 0$*, and consequently,* $\chi = 0$*.*

Proof of Theorem 5.3.2. Similar to the proof of Theorem 5.3.1, this result boils down to show that, under the conditions presented in Theorem statement, any absolutely continuous trajectory of the continuous time differential inclusion (5.9) converges to the set $\tilde{\Lambda}$. The proof comprises of two parts: first, we show that the trajectory of dynamical system (5.9) will eventually converge to a close neighborhood of an approximate equilibrium, where one π^{*k} for some $k \in [K]$. Second, we bound the maximum distance that the policy trajectory can be away from that equilibrium. These two steps together show that equilibrium will be refined in the set $\tilde{\Gamma}$ that is in the neighborhood of one equilibrium π^{*k}.

Before presenting each of these steps in detail, we present a few preliminary results that will be used later. Define

$$d^* := \min_{k,l\in[K],k\neq l} \|\pi^{*k} - \pi^{*l}\|$$

as the minimum distance between a pair of equilibrium policies. Additionally, define

$$\mathcal{B}(\pi; r) := \{\pi' \in \Pi : \|\pi' - \pi\|_2 \leqslant r\}$$

as the neighborhood set of a policy π with radius r. Let ζ be such that $\Gamma(\zeta) \leqslant d^*/4$, which exists due to Lemma D.1.2. This construction ensures that for any $k, l \in [K]$ such that $k \neq l$ it holds that

$$\mathcal{B}(\pi^{*k}; \Gamma(\zeta)) \cap \mathcal{B}(\pi^{*l}; \Gamma(\zeta)) = \varnothing.$$

Moreover, from the definition of $\Gamma(\cdot)$ in (5.21), it must hold that for any $\delta > 0$, $\mathsf{NE}(\delta) \subseteq \cup_{k\in[K]}\mathcal{B}(\pi^{*k}; \Gamma(\delta))$. Additionally, using the fact that for any $k, l \in [K]$ such that $k \neq l$, it must hold that $\mathcal{B}(\pi^{*k}; \Gamma(\zeta)) \cap \mathcal{B}(\pi^{*l}; \Gamma(\zeta)) = \varnothing$,. Therefore, we conclude that $\mathsf{NE}(\zeta)$ is contained in a disjoint union of sets, which is $\sqcup_{k\in[K]}\mathcal{B}(\pi^{*k}; \Gamma(\zeta))$. We select $\bar{\kappa}, \bar{\epsilon}$ such that

$$\bar{\epsilon} + \Theta\bar{\kappa} < \zeta/2, \text{and}, \Gamma(\bar{\epsilon} + \Theta\bar{\kappa}) \leqslant \frac{\zeta\eta d^*}{32NDL\sqrt{S}}.$$

Lemma D.1.2 ensures that such construction exists. Finally, we define $\kappa' := \kappa + \lambda$.

From the proof of Theorem 5.3.1 we know that starting from any initial policy, any solution of (5.9) eventually hits the set $\mathsf{NE}(\bar{\epsilon} + \Theta\kappa')$ in finite time. Suppose that the trajectory leaves the component of the set $\mathsf{NE}(\bar{\epsilon} + \Theta\kappa')$ around the neighborhood of π^{*k} and enters the component in the neighborhood of π^{*l}, for some $k, l \in [K]$ such that $k \neq l$. Since $\bar{\epsilon} + \Theta\kappa' \leqslant \bar{\epsilon} + \Theta\bar{\kappa} < \zeta$, it holds that $\mathsf{NE}(\bar{\epsilon} + \Theta\kappa') \subseteq \mathsf{NE}(\zeta)$, which is contained in the set $\sqcup_{k\in[K]}\mathcal{B}(\pi^{*k}; \Gamma(\zeta))$. Therefore, any trajectory has to leave $\mathcal{B}(\pi^{*k}; \Gamma(\zeta))$ and enter $\mathcal{B}(\pi^{*l}; \Gamma(\zeta))$. Let t_1 denote the time when the trajectory leaves the component of $\mathsf{NE}(\bar{\epsilon} + \Theta\kappa')$ in neighborhood of π^{*k}, t_2 denote the time when it leaves $\mathcal{B}(\pi^{*k}; \Gamma(\zeta))$, t_3 denote the time when the trajectory enters $\mathcal{B}(\pi^{*l}; \Gamma(\zeta))$, and t_4 denotes the time when the trajectory enters the component of $\mathsf{NE}(\bar{\epsilon} + \Theta\kappa)$ around π^{*l}. Since the function $\phi(\tau)$ is decreasing outside $\mathsf{NE}(\bar{\epsilon} + \Theta\kappa')$, it must hold that $\Phi(\mu, \varpi^{t_2}) \geqslant \Phi(\mu, \varpi^{t_1})$ and $\Phi(\mu, \varpi^{t_4}) \geqslant \Phi(\mu, \varpi^{t_3})$. We claim that

$$\|\varpi^{t_2} - \varpi^{t_3}\| \geqslant d^*/2. \tag{5.23}$$

We show this by contradiction. Suppose that $\|\varpi^{t_2} - \varpi^{t_3}\| < d^*/2$. Since ϖ^{t_2} lies on the boundary of $\mathcal{B}(\pi^{*k}; \Gamma(\zeta))$ and ϖ^{t_3} on $\mathcal{B}(\pi^{*l}; \Gamma(\zeta))$, it must hold that

$$\|\varpi^{t_2} - \pi^{*k}\| = \Gamma(\zeta) = \|\varpi^{t_3} - \pi^{*l}\|.$$

Furthermore, using the definition of d^*, we know that $\|\pi^{*k} - \pi^{*l}\| \geqslant d^*$. But, from triangle inequality it must hold that

$$\begin{aligned}\|\pi^{*k} - \pi^{*l}\| &\leqslant \|\pi^{*k} - \varpi^{t_2}\| + \|\varpi^{t_2} - \varpi^{t_3}\| + \|\varpi^{t_3} - \pi^{*l}\| \\ &= 2\Gamma(\zeta) + \|\varpi^{t_2} - \varpi^{t_3}\| < d^*,\end{aligned}$$

which leads to a contradiction. Thus, (5.23) holds.

Using (5.23) and the fact that $\|\dot{\varpi^\tau}\| \leqslant 2N\sqrt{|S|}$, it must hold that $t_3 - t_2 \geqslant d^*/(4N\sqrt{|S|})$. Additionally, using the fact the trajectory in the time interval $[t_2, t_3]$ lies outside $\mathsf{NE}(\zeta)$ and (5.11), we conclude that

$$\phi(t_3) - \phi(t_2) \leqslant -(\zeta - \Theta\bar{\kappa})\frac{\eta d^*}{4ND\sqrt{S}}. \tag{5.24}$$

Define

$$\underline{\phi} := \min_{\pi \in \mathcal{B}(\pi^{*k}; \Gamma(\bar{\epsilon}+\Theta\kappa'))} \Phi(\mu, \pi), \quad \bar{\phi} := \max_{\pi \in \mathcal{B}(\pi^{*l}; \Gamma(\bar{\epsilon}+\Theta\kappa'))} \phi(\pi).$$

We claim that

$$\bar{\phi} \leqslant \underline{\phi}. \tag{5.25}$$

Suppose that this is true, it shows that once the trajectory leaves the component of approximate equilibrium around π^{*k} and enters the component of π^{*l} it will never enter the component of approximate equilibrium around π^{*k} in future. Thus, eventually the trajectories will visit the component of approximate equilibrium only around at most one equilibrium. Now we show (5.25). Let $y_k \in \mathcal{B}(\pi^{*k}; \Gamma(\bar{\epsilon} + \Theta\kappa')), y_l \in \mathcal{B}(\pi^{*l}; \Gamma(\bar{\epsilon} + \Theta\kappa'))$ be such that $\underline{\phi} = \phi(y_k), \bar{\phi} = \phi(y_l)$. This ensures that $\|y_k - \varpi^{t_1}\| \leqslant 2\Gamma(\bar{\epsilon} + \Theta\kappa')$ and $\|y_l - \varpi^{t_4}\| \leqslant 2\Gamma(\bar{\epsilon} + \Theta\kappa')$. Furthermore, from the Lipschitz property of potential function, it holds that

$$\begin{aligned}
\phi(\varpi^{t_2}) - \underline{\phi} \leqslant \phi(\varpi^{t_1}) - \underline{\phi} \leqslant L\|y_k - \varpi^{t_1}\| \leqslant 2L\Gamma(\bar{\epsilon} + \Theta\bar{\kappa}),\\
\bar{\phi} - \phi(\varpi^{t_3}) \leqslant \bar{\phi} - \phi(\varpi^{t_4}) \leqslant L\|y_l - \varpi^{t_4}\| \leqslant 2L\Gamma(\bar{\epsilon} + \Theta\bar{\kappa}).
\end{aligned}$$

Combining the two inequalities and using (5.24), we obtain

$$\bar{\phi} - \underline{\phi} \leqslant 4L\Gamma(\bar{\epsilon} + \Theta\bar{\kappa}) - (\zeta - \Theta\bar{\kappa})\frac{\eta d^*}{4ND\sqrt{S}}.$$

The choice of $\bar{\kappa}$ and $\bar{\epsilon}$ ensures that

$$4L\Gamma(\bar{\epsilon} + \Theta\bar{\kappa}) - (\zeta - \Theta\bar{\kappa})\frac{d^*\eta}{16NLS} < 0,$$

which shows (5.25).

To summarize, so far we have shown that there exists a time T after which the trajectories will only visit the approximate equilibrium set that is close to one equilibrium. Let's say that the equilibrium is π^{*k}. Next, we characterize the maximum distance the trajectory can travel around the equilibrium π^{*k}.

Fix ϵ, ϵ_1 arbitrarily such that $0 \leqslant \epsilon_1 < \epsilon \leqslant \bar{\epsilon}$. Let τ_1 denote the time when the trajectory leaves the component of $\mathsf{NE}(\epsilon_1 + \Theta\kappa')$ in neighborhood of π^{*k}, τ_2 denote the time when

the trajectory leaves $\mathcal{B}(\pi^{*k}; \Gamma(\epsilon + \Theta\kappa'))$, τ_3 denote the time when the trajectory returns to this neighborhood again, and τ_4 denotes the time enters the component of $\mathsf{NE}(\epsilon_1 + \Theta\kappa')$ in neighborhood of π^{*k}.

Let r^* denote the maximum distance the trajectory goes outside of $\mathcal{B}(\pi^{*k}; \Gamma(\epsilon + \Theta\kappa'))$. Since $\|\dot{\varpi}^\tau\| \leqslant 2N\sqrt{|S|}$, it must hold that

$$\tau_3 - \tau_2 \geqslant \frac{r^*}{N\sqrt{|S|}}.$$

Furthermore, since $\dot{\phi} \leqslant -\eta\epsilon/D$ during time interval $[\tau_2, \tau_3]$, it must hold that $\phi(\tau_3) - \phi(\tau_2) \leqslant -\frac{r^*\eta\epsilon}{ND\sqrt{|S|}}$. Moreover, note that $\phi(\tau_2) \leqslant \phi(\tau_1)$, and $\phi(\tau_4) \leqslant \phi(\tau_3)$. This implies

$$\phi(\tau_1) - \phi(\tau_4) \geqslant \phi(\tau_2) - \phi(\tau_3) \geqslant \frac{r^*\eta\epsilon}{ND\sqrt{|S|}}.$$

Furthermore, by Lipschitz continuity

$$\phi(\tau_1) - \phi(\tau_4) \leqslant L\|\varpi^{\tau_1} - \varpi^{\tau_4}\| \leqslant L(\|\varpi^{\tau_1} - \pi^{*k}\| + \|\pi^{*k} - \varpi^{\tau_4}\|) \leqslant 2L\Gamma(\epsilon_1 + \Theta\kappa').$$

Combining previous two inequalities we obtain that

$$\frac{r^*\eta\epsilon}{ND\sqrt{|S|}} \leqslant 2L\Gamma(\epsilon_1 + \Theta\kappa') \implies r^* \leqslant \frac{2DLN\sqrt{|S|}\Gamma(\epsilon_1 + \Theta\kappa')}{\eta\epsilon}.$$

The proof concludes by noting that the maximum distance the trajectories eventually goes away from π^{*k} is

$$\Gamma(\epsilon + \Theta\kappa') + \frac{2DLN\sqrt{|S|}\Gamma(\epsilon_1 + \Theta\kappa')}{\eta\epsilon}.$$

Since ϵ, ϵ_1 are chosen arbitrarily, we conclude the proof of the theorem. □

5.4 Numerical Experiments

We demonstrate the performance of the proposed learning dynamics (Algorithm 4) in a perturbed Markov team game with a parallel link network of L links used by N travelers. At each stage k, player i chooses a link $a_i^k \in [L]$, and the network state is $s = (s_\ell)_{\ell \in [L]}$, where $s_\ell = 1$ means link ℓ is unsafe. A link becomes unsafe with probability ν_1 if the number of users exceeds a threshold T, and ν_2 otherwise.

The utility of player i is a combination of individual and common rewards:

$$u_i(s,a) = \underbrace{\rho_i \sum_{\ell=1}^{L} \mathbb{I}(a_i=\ell)\Big(b_\ell - (1+s_\ell)m_\ell \sum_{j\in I}\mathbb{I}(a_j=\ell)\Big)}_{\text{Individual Reward}} + \underbrace{\sum_{i\in N}\sum_{\ell=1}^{L} \mathbb{I}(a_i=\ell)\Big(b_\ell - (1+s_\ell)m_\ell \sum_{j\in I}\mathbb{I}(a_j=\ell)\Big)}_{\text{Common Reward}}.$$

Here, b_ℓ is the fixed utility of using link $\ell \in [L]$, m_ℓ is a link-dependent constant which weights the effect of congestion on the cost and $(1+s_\ell)m_\ell \sum_{j\in I}\mathbb{I}(a_j=\ell)$ represents the cost of using ℓ given the total number of users on ℓ and the network state. The utility depends on two terms: one is an individual reward function and another is the common reward term. The parameter ρ_i characterizes how much agent i weighs individual reward over the common interest.

Although the game is not a Markov potential game, the expected long-run common reward acts as a near-potential function with closeness parameter depending on $\rho = \max_{i\in I}\rho_i$.

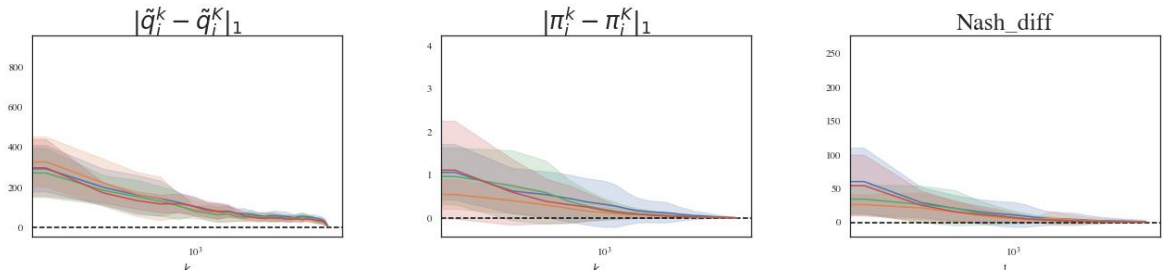
We simulate the system with $N=4$, $L=2$, $T=2$, $\nu_1=0.8$, $\nu_2=0.2$, $m_1=2$, $b_1=9$, $m_2=4$, and $b_2=16$, for 5000 stages. The step sizes are $\alpha_i(n)=1/n^{0.5}$ and $\beta_i(n)=1/n^{0.9}$. We set $\rho_i=(i/N)\cdot K$, with $K\in\{10,100,500,1000\}$. Initial link states are random.

For the exploration parameter $\theta=0.05$ and the perturbation parameter ρ_i with $K=10$, policies converge close to Nash equilibrium (Figure 5.2). Interestingly, we observe that the Nash gap goes close to zero even though $K\neq 0$ indicating that the parameter κ for this game is a very small number. Furthermore, we see that the Nash approximation gaps[3] increase with θ and K (Figure 5.3), which validates our theoretical convergence guarantees.

5.5 Concluding Remarks

In this chapter, we characterize the set of convergent strategies under decentralized actor-critic dynamics in general-sum Markov games. This extends the results from Chapter 5, which focused exclusively on Markov potential games—a special subclass of general-sum Markov games. To facilitate this extension, we introduce a novel framework of *Markov near-potential functions*, inspired by the α-potential functions discussed in Chapter 4. This framework guarantees the existence of a near-potential function for any Markov game, such that the rate of change of a player's value function with respect to their own strategy is well-approximated by the rate of change of the near-potential function. Central to our analysis is the observation that the near-potential function serves as an approximate Lyapunov function for the policy updates in decentralized actor-critic algorithm, enabling us to characterize convergence behavior even beyond potential games.

[3]We define the Nash gap to be $\max_{i\in I}\max_{s\in S}|V_i(s,\pi^{t_{\max}}) - \max_{\pi_i\in\Pi_i} V_i(s,\pi_i,\pi_{-i}^{t_{\max}})|$, where $t_{\max}$ is the number of iterations of Algorithm 4.



(a) (b) (c)

Figure 5.2: Convergence of q-estimate, policies, and Nash error after 10^5 steps of Algorithm 4. In each of the figures the four curves correspond to four players. Each curve represents the mean value of the quantity over 5 trials, and we give error margins of ± 1 standard deviation.

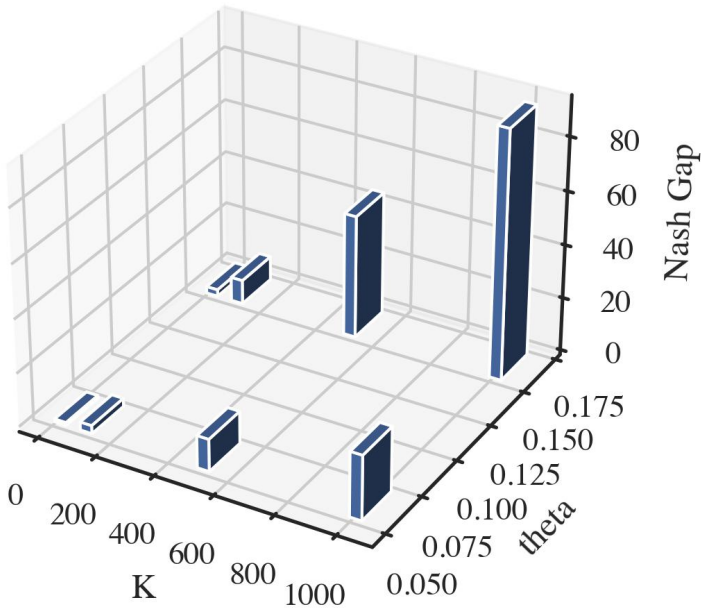


Figure 5.3: Variation of Nash gap with change in the exploration rate θ and the reward perturbation K. Increasing both of these parameters increases the Nash gap.

Part II

Multi-agent Learning in Resource-Constrained / Congested Environments

Chapter 6

Decentralized Learning in Matching Markets

Online decision-making under uncertainty is one of the central problems in modern machine learning, reflecting the need for efficient and high-performing algorithms for real-time learning in real-world settings. Despite being a well-researched area, there remains a broad lack of understanding of how to deploy online learning algorithms in settings where they must compete with each other for resources or information.

While classical problems in online learning focus on balancing the exploration of possible choices with the exploitation of current knowledge (i.e., the exploration–exploitation trade-off [228, 386]), the introduction of competition adds a new dimension to the problem [281, 15]—namely, the challenge of competing (perhaps unsuccessfully) for highly desired outcomes or settling for less desired (but also less competitive) alternatives.

Broadly speaking, the dominant approach to handling competition in machine learning has been to treat opponents as adversarial [87], despite a long-standing literature in economics and game theory [252, 149] showing that agents who understand the competitive structure of their environment can often outperform solutions based on worst-case assumptions.

In this chapter, we address the problem of online learning in competitive settings within the context of two-sided matching markets. Two-sided matching markets match users on one side of the market with those on the other to facilitate the exchange of goods or services.

In such settings, each user on one side of the market has a preference ordering over the users on the other side. Since each user seeks to find their most desired match, this results in a game in which a natural equilibrium concept is that of a stable matching, wherein no two users would prefer switching from their current match to each other, given their preferences. In seminal work, [151] proposed a simple and effective algorithm—the Deferred Acceptance (DA) Algorithm—that users on one side of the market can implement to find such a solution when every user knows their own preferences.

The algorithm has been widely used in applications ranging from kidney exchanges to medical resident matching, where preferences can be assigned or reported to a central au-

thority that performs the matching. However, recent years have seen the emergence of a new form of online matching markets, such as online labor markets (e.g., TaskRabbit, Upwork), online dating platforms (e.g., Tinder, Match.com), and online crowdsourcing platforms (e.g., Amazon Mechanical Turk), where users do not know their preferences a priori and can interact with the market repeatedly to improve their match quality.

Motivated by these applications, we consider a generalization of the problem studied in the seminal paper [151], wherein one side of the market—the agents—do not know their own preferences but are able to interact repeatedly with the market. In particular, we analyze a repeated game in which, at each round, agents can request to match with a user or firm on the other side of the market. If, in a given round, multiple agents request the same firm, the firm—assumed to be a myopic utility maximizer—accepts the request of its most preferred agent (who receives a noisy measurement of their utility from the match, from which they can learn their preferences) and rejects the others (who receive no information about their preferences). This setup has been studied in a line of recent works on online matching markets [256, 257, 363, 35].

Successful algorithms in this framework must simultaneously solve a statistical learning problem (learning about an agent's own preferences) and a competitive problem (ensuring that agents secure their most desired match despite the presence of other self-interested agents). Previous approaches to this problem propose algorithms that are centralized [256] (where agents send their current beliefs about their preferences to a central platform that performs the matching), require coordination between agents (i.e., a choreographed set of strategies to minimize rejections)[363, 35], or assume agents can fully observe the market outcomes of others[257].

In contrast, the DA algorithm—which we take as the full-information benchmark for comparison—is (i) fully decentralized, (ii) coordination-free, and (iii) requires agents to make decisions solely based on their own history of rejections and successful matches. Designing learning algorithms that operate under conditions (i)–(iii) ensures scalability and privacy in large-scale systems, where it is unrealistic to assume agents can track all other agents' matchings. Thus, in this work, we focus on the question:

> **Does there exist a decentralized, coordination-free algorithm—based only on agents' local history of interactions—that provably converges to a stable matching?**

Contributions. In this chapter, we design algorithms for learning while matching in a class of structured matching markets known as α reducible matching markets. This condition ensures the existence of a unique stable matching and encompasses many realistic preference structures, including serial dictatorship and no-crossing conditions [98]. We show that the proposed algorithms incur stable regret with respect to the unique stable matching, which grows at most logarithmically with the time horizon. The particular contributions of this chapter are:

1. We present a general framework for constructing decentralized, communication-free, and coordination-free algorithms for learning while matching. In particular, we combine index-based stochastic bandit algorithms—specifically the Upper Confidence Bound (UCB) algorithm and Thompson Sampling [20, 228, 386]—to address the statistical problem of learning an agent's preferences, with a path-length adversarial bandit algorithm [72, 425] to address the competitive problem. The resulting algorithms are fully decentralized and require no communication or coordination, as each agent selects which firm to request based solely on their own history of collisions, matches, and rewards.

   Furthermore, the algorithms are "any-time" algorithms, meaning they do not require knowledge of the time horizon or any specific parameters of the bandit instance beyond the sub-Gaussian noise parameter.

2. We show that when the agents' and firms' preferences satisfy the α-reducibility condition, and *every* agent uses the proposed algorithm, the regret accumulated by any agent a with respect to the stable matching is

$$O\left(\frac{C_a|A||F|\log(T)}{\Delta^2}\right),$$

   where A is the set of agents, F is the set of firms, Δ is the minimum sub-optimality gap for any agent in the market, and C_a is a constant that depends on the α-reducible structure of the market.

Related works

Sequential decision-making under uncertainty has been extensively studied in machine learning under the framework of multi-armed bandit (MAB) problems. In general, MAB problems can be divided into two distinct categories, which differ in the type of feedback agents receive. Crucially, in both settings, the central challenge is to trade off the exploration of actions with the exploitation of current knowledge.

The first class of MAB problems, known as *stochastic MAB*, provides agents with an unbiased estimate of the utility of an action when it is played. Solutions to this problem generally fall into two dominant algorithmic paradigms. The first, based on the principle of *optimism in the face of uncertainty*, includes the well-known Upper Confidence Bound (UCB) algorithm [228, 223] and its variants. The second approach, based on *Thompson Sampling*, adopts a Bayesian perspective [359, 407]. Each of these approaches is known to achieve optimal performance, measured in terms of *regret*: the expected cumulative utility lost by following the algorithm instead of always selecting the optimal action (i.e., the best action in hindsight with full information) [228, 6]. In particular, both paradigms are known to achieve *logarithmic* regret—regret that grows at most logarithmically over time—which is optimal for this class of problems up to constant factors.

In this chapter, we present an algorithmic framework for learning in matching markets that is compatible with either class of algorithm, and furthermore, achieves logarithmic regret *even* in the presence of competition.

The second class of multi-armed bandit problems, originating from the literature on learning in games, seeks algorithms that perform well against arbitrary feedback sequences [87]. Solutions to this class of problems, known as *adversarial bandit algorithms*, remain an active area of research. While it is well known that simple strategies such as multiplicative weights can guarantee regret against the best fixed action in hindsight on the order of $\sqrt{T}$ in the worst case [87], designing algorithms that improve upon this bound when adversaries are *not* worst-case remains an open problem. In this chapter, we leverage advances in the development of *path-length* adversarial regret algorithms that address this challenge by guaranteeing regret bounds that depend directly on the variation exhibited by the adversary [72, 425].

We briefly remark that several lines of research study multi-agent bandits. One such line focuses on multi-agent bandits with collisions, primarily motivated by applications in wireless spectrum sharing [255, 205, 350, 264, 71]. In these models, the arms do not have preferences, and when multiple agents select the same arm, no agent receives any utility or all incur maximum possible loss. These models differ from our setting, where both sides of the market have preferences over one another, and in the event of a collision, only one agent is matched. Another line of work explores collaborative learning in bandit settings [73, 88, 364], where agents can communicate and jointly learn a single bandit instance. Notably, these settings typically do not model competition—i.e., more than one agent choosing the same arm at the same time.

The particular intersection of multi-armed bandits and two-sided matching markets has recently garnered considerable attention [256, 257, 35, 363]. To the best of our knowledge, [116] presented the first numerical study on using MAB algorithms to learn preferences in matching markets. However, it was only recently that [256] formally introduced the bandit learning problem in matching markets, and generalized the notion of *regret* from the MAB literature to matching markets via the concept of *stable regret*—i.e., the expected cumulative utility benchmarked against the expected cumulative reward that would have been received if every agent had consistently requested their match in a fixed stable matching.[1] Moreover, they proposed a *centralized* UCB-based algorithm that coordinates matching between agents and firms based on each agent's beliefs over preferences and play history, and achieves a regret bound of $\mathcal{O}(|A||F|\log(T))$, where A and F are the sets of agents and firms, respectively, and T is the time horizon.

In follow-up work, [257] proposed a *decentralized* bandit learning algorithm based on UCB that allows agents to make decisions independently while still converging to a stable matching, with a regret bound of $O(\exp(|F|^4)\log^2(T))$. More recently, [215] introduced a Thompson Sampling-based variant of [257]. However, both of these approaches require

[1]Note that the stable matching need not be unique in general. Thus, the notion of stable regret must be specified with respect to a particular stable matching. Typically, the literature focuses on two canonical stable matchings: the *agent-optimal* and *firm-optimal* stable matchings.

knowledge of the outcomes at *other* firms in each round, leaving open the question of whether algorithms based solely on each agent's local history of play can achieve similar guarantees.

Concurrently, [363] proposed a phased algorithm that uses communication between agents to coordinate their actions. Under this information structure, their algorithm achieves regret of $\mathcal{O}(|F|^2|A|^2\log(T))$, but assumes that all firms have homogeneous preferences over agents (also known as the *serial dictatorship* model). Follow-up work by [35] improved the regret for serial dictatorship to $\mathcal{O}(|F||A|\log(T))$ by designing a new algorithm. Additionally, they showed that relaxing the serial dictatorship assumption to a weaker structural condition yields $O(\text{poly}(|A|,|F|)\log(T))$ regret. Although the algorithm in [35] allows for some decentralization, it is still a phase-based approach in which agents follow a coordinated protocol during specific rounds.

In this chapter, we propose a simple, decentralized, communication- and coordination-free algorithm in which agents rely solely on their own local observations to learn while matching. In contrast to prior works [256, 257, 363, 35], which rely exclusively on UCB-based subroutines, we also show that our algorithmic framework naturally extends to a Thompson Sampling variant.

We also remark on another line of research at the intersection of multi-armed bandits and matching markets [188, 199, 85], which considers the problem of learning preferences from the perspective of a platform.

Organization. The chapter is organized as follows: In Section 6.1, we introduce the general problem setup, define matching markets, and describe the structural assumptions on the preferences of agents and firms. In Section 6.2, we present the algorithmic design paradigm along with a specific algorithm based on the Upper Confidence Bound. In Section 6.3, we show that the algorithm incurs $O(\log(T))$ regret and provide a brief sketch of the proof. In Section 6.4, we study the performance of the algorithm through simulations. We conclude the chapter in Section 6.5.

The proofs of our results are relegated to Appendix E. Additionally, we introduce another important algorithmic variant based on Thompson Sampling and present similar results, also in Appendix E.

6.1 Setup

We define a two-sided market M as a collection of agents A and firms F. In the setting under consideration, we assume that each agent $a \in A$ has *unknown* preferences over firms $f \in F$, represented by utilities $u_a(f) \in \mathbb{R}$. Moreover, no two firms provide the same utility to a given agent; that is, $u_a(f) \neq u_a(f')$ for $f \neq f'$.

We assume that each agent seeks to be matched to their most preferred firm, and that firms have preferences over all agents, which are also captured by utilities $u_f(a)$ for each a and f, such that no two agents yield the same utility to a firm; that is, $u_f(a) \neq u_f(a')$ for

$a \neq a'$. Importantly, we assume that firms know their own preference orderings over agents, and that there are more firms than agents, i.e., $|A| \leqslant |F|$.

The interaction between agents and firms unfolds as follows: at each time step $t = 1, \ldots, T$, every agent $a \in A$ independently *requests* a firm $f_a(t) \in F$. Since agents request independently, it is possible that multiple agents request the same firm f. For each $f \in F$, let

$$\mathbb{A}_f(t) \coloneqq \{a \in A : f_a(t) = f\}$$

denote the set of agents that request firm f at time t. At each time step t, we assume that firm f accepts the request of its most preferred agent in $\mathbb{A}_f(t)$, denoted by

$$a_f(t) \coloneqq \arg\max_{a \in \mathbb{A}_f(t)} u_f(a),$$

and rejects the request of all other agents. The agent $a_f(t)$ is said to be *matched* with firm f at time t.

Moreover, every matched agent receives a noisy measurement of their utility, denoted $r_{\mathbf{a},f}$, such that

$$r_{\mathbf{a},f} = u_{\mathbf{a}}(f) + \zeta_{\mathbf{a},f}, \tag{6.1}$$

where $\zeta_{\mathbf{a},f}$ is a zero-mean, one-sub-Gaussian random variable. Meanwhile, all agents who are rejected are said to have *collided* on firm f, and receive no utility; that is, $r_{a,f}(t) = 0$.

We restrict that agents *only* receive the following information at any time step t:

1. $Y_a(t) = \mathbb{1}\left[\text{agent } a \text{ is matched to } f_a(t)\right]$, which indicates whether agent a is matched at time t;

2. if matched, the agent receives a noisy measurement of their utility, $U_{a,f}(t)$.

Remark 6.1.1. *We note that in this setup, an agent does not know* anything *about how other agents are performing in the market. Agents do* not *observe who gets successfully matched to the firms they have requested, nor do they observe with whom they have collided. We emphasize that this is* the same *information structure assumed by the DA algorithm, and it is the key assumption that distinguishes our work from prior approaches to this problem [256, 257, 35, 363].*

In the following subsection, we recall some important results from the matching market literature that are crucial for the subsequent exposition.

Preliminaries on Matching Markets

To analyze the matching market defined in the previous section, we first review key concepts from the literature on matching markets. A matching $\mathbb{M} : A \to F$ is an injective function such that $\mathbb{M}(a) = f$ denotes that agent a is matched with firm f.

A matching is called *unstable* if there exists an agent–firm pair $(a, f) \in A \times F$ such that

$$u_a(\mathbb{M}(a)) < u_a(f) \quad \text{and} \quad u_f(a) > u_f(\mathbb{M}^{-1}(f)).$$

In other words, the agent and the firm prefer each other over their current match; such a pair is referred to as a *blocking pair*. A matching is said to be *stable* if it is not unstable.

It is well known that a market may admit multiple stable matchings. However, for the purposes of this chapter, we focus on markets that are α-reducible, a property first introduced in [9] and further analyzed in [98], which guarantees the existence of a unique stable matching. Before formally describing this property, we introduce the notions of a submarket and a fixed pair.

A *submarket* of M is a market M' such that $M' = A' \cup F'$, where $A' \subseteq A$, $F' \subseteq F$, and $|A'| \leqslant |F'|$.

A pair $(a, f) \in A \times F$ is called a *fixed pair* of the market M if

$$u_a(f) \geqslant u_a(f') \quad \text{for all } f' \in F, \quad \text{and} \quad u_f(a) \geqslant u_f(a') \quad \text{for all } a' \in A.$$

In words, a fixed pair is an agent–firm pair that strictly prefer each other over any other available option in the market.

We now define the notion of α-reducibility.

Definition 6.1.1 (α-reducibility)**.** *A market $M = A \cup F$ is α-reducible if every sub-market of M has a fixed pair.*

The notion of α-reducibility is weaker than the *no-crossing condition* and serial dictatorship [98]. These conditions have been introduced in the effort to characterize the existence and uniqueness of stable matchings. In [98], the authors show that every submarket of M admits a unique stable matching if M is α-reducible.

The preceding property of α-reducible markets will be crucial for establishing regret guarantees for the proposed algorithm in this chapter. Therefore, we assume that M is α-reducible.

Remark 6.1.2. *An important property of the α-reducibility assumption, which is central to the subsequent analysis, is that it allows us to partition the market into a sequence of submarkets by sequentially eliminating fixed pairs. More formally, let us define $\mathcal{A}_0 = \mathcal{F}_0 = \varnothing$ and $M_0 = M$. Now, for $i \geqslant 1$, define inductively:*

$$\mathcal{A}_i \subseteq A \setminus \bigcup_{j=0}^{i-1} \mathcal{A}_j, \quad \mathcal{F}_i \subseteq F \setminus \bigcup_{j=0}^{i-1} \mathcal{F}_j$$

to be the sets of agents and firms that form fixed pairs in the market M_{i-1}. That is, for every agent $a \in \mathcal{A}_i$, there exists a unique firm $f \in \mathcal{F}_i$ such that (a, f) is a fixed pair in M_{i-1}. The market then evolves as:

$$M_i \coloneqq \left(A \setminus \bigcup_{j=0}^{i} \mathcal{A}_j\right) \cup \left(F \setminus \bigcup_{j=0}^{i} \mathcal{F}_j\right).$$

Let K denote the total number of such submarkets $\{M_i\}$. Moreover, this decomposition of the market is unique.

For any agent $a \in A$, we denote by f_a^* its match in the unique stable matching. Furthermore, let

$$\overline{\mathbb{F}}_a \coloneqq \{f \in F : u_a(f) > u_a(f_a^*)\}$$

be the set of firms that agent a prefers over its stable match. We call such firms *super-optimal* firms for a.

Similarly, let

$$\underline{\mathbb{F}}_a \coloneqq \{f \in F : u_a(f) < u_a(f_a^*)\}$$

be the set of firms less preferred than the stable match by agent a. We call these *sub-optimal* firms for a. Note the following lemma, which states a crucial property of super-optimal firms in α-reducible markets.

Lemma 6.1.1. *For any $i \in [K]$ and agent $a \in \mathcal{A}_i$, the set of super-optimal firms is contained in $\bigcup_{j=1}^{i-1} \mathcal{F}_j$.*

An immediate consequence of Lemma 6.1.1 is that it induces a hierarchy in the market. That is, an agent $a \in \mathcal{A}_i$, for some $i \in [K]$, is in a sense "higher ranked" than an agent $a' \in \mathcal{A}_j$ with $j > i$, since the stable match of the former can be super-optimal for the latter. This hierarchy naturally manifests in the learning process, where the learning of agent a creates an *externality* for agent a'.

6.2 Description of the Algorithm

In this section, we present a novel algorithm design principle that enables agents to learn their preferences while ensuring competitive performance compared to the matching they could have achieved had they known their preferences and used the DA algorithm. Throughout this section, we assume that every agent $a \in A$ uses these algorithms to decide which firm to choose at any time t.

The proposed algorithms—by design—use only the feedback information outlined in (1)–(2) in Section 6.1, and involve no implicit or explicit communication or coordination strategies such as phase-based methods with coordinated actions [35] or partial observation of other agents' actions [257]. Thus, the algorithms operate in the same informational regime as the DA algorithm but without assuming that agents know their preferences.

A key aspect of our approach is blending stochastic bandit (SB) algorithms with adversarial bandit (AB) algorithms. In the subsequent exposition, we formally describe our approach and demonstrate its desirable properties in terms of regret and convergence.

Before proceeding, we comment on the difficulties of the problem at hand and what makes the analysis of these algorithms highly non-trivial. The key challenge in designing algorithms for matching while learning is determining when to stop requesting *super-optimal*

firms (i.e., firms that an agent prefers over their stable match) without any prior knowledge of the market structure.

The crux of this problem lies in enabling an agent to learn that certain firms are unattainable due to competition, despite the non-stationarity of the environment arising from the fact that other agents are simultaneously learning, while the agent does not know with whom they collide or who is successfully matched in each round. Furthermore, due to the lack of communication or coordination, agents cannot learn which firms are super-optimal without risking numerous collisions.

A high-level sketch of the algorithm is provided in Algorithm 5, while the precise algorithm—where agents use the UCB algorithm as a subroutine—is detailed in Algorithm 6.

Algorithm 5 High-level algorithmic description

Each agent $a \in A$, at every time $t \in [T]$:

1. Keeps an ordering of firms according to an index computed by a stochastic bandit subroutine.
2. Agent a goes through the firms one by one according to this ordering.
3. Using an adversarial bandit subroutine, decides whether to *request* the firm or to *prune* it:
   a) If a firm is requested, the agent either gets matched or collides.
   b) If pruned, the agent moves to the next firm in the ordering.
4. Updates both the stochastic and adversarial bandit subroutines based on the feedback received.

As per Algorithm 6, each agent is equipped with a stochastic bandit (SB) subroutine. At every time step $t \in [T]$, the SB subroutine of each agent a maintains an ordering of firms in decreasing order of preference according to an index (e.g., UCB). We denote the index of firm f maintained by agent a at time t as $\mathsf{UCB}_{a,f}(t)$.

At time t, each agent considers firms one by one in decreasing order of $\mathsf{UCB}_{a,f}(t)$. For any firm f considered by agent a, the agent decides either to *request* f or to *prune*[2] it (i.e., reject that firm). Specifically, agent a requests firm f with probability $p_{a,f}(t)$. Let $P_{a,f}(t) \sim \text{Bernoulli}(p_{a,f}(t))$.

If a firm is pruned (i.e., $P_{a,f}(t) = 0$), then the agent considers the next best firm from the sorted list, continuing this process until a firm is requested (i.e., $P_{a,f}(t) = 1$). If all firms are pruned, the agent requests the firm with the highest index: $\arg\max_f \mathsf{UCB}_{a,f}(t)$.

Once an agent decides which firm to request, it obtains a noisy utility measurement if it is successfully matched. This feedback is used to update its UCB index. Based on

[2]Note that pruning here is not permanent; it indicates that a particular firm is not considered at that time step.

whether agent a prunes or requests a firm f, it updates $p_{a,f}$ using an adversarial bandit (AB) subroutine. The details of this are given below.[3]

We note that not all firms are considered by agent a at every time t. Once an agent decides to request a firm f, it does not consider firms in the set $\{f' \in F : \mathcal{I}_{a,f'}(t) < \mathcal{I}_{a,f}(t)\}$. Formally, for any agent-firm pair $(a, f) \in A \times F$, define the event that agent a *selects* firm f at time t to decide whether to request or prune it by

$$E^{(\mathsf{c})}_{a,f}(t) = \mathbb{1}\left\{P_{a,f'}(t) = 0, \quad \forall f' : \mathcal{I}_{a,f}(t) \leqslant \mathcal{I}_{a,f'}(t)\right\}.$$

If firm f is considered by agent a, then the event that agent a requests f is denoted by

$$E^{(\mathsf{r})}_{a,f}(t) = \mathbb{1}\left\{P_{a,f}(t) = 1,\ E^{(\mathsf{c})}_{a,f}(t) = 1\right\}.$$

Next, we describe the UCB computation method for the SB subroutine. Following which, we illustrate how the matchings and collisions are used to update the probability $p_{a,f}(t)$ as per an AB subroutine.

Stochastic Bandit Subroutine

The stochastic bandit subroutine is used to efficiently handle the inherent uncertainty in the payoffs obtained upon successful matching. In this section, we develop the theory for the setting in which agents use a UCB-based stochastic bandit (SB) subroutine. Similar results for Thompson Sampling are provided in the Appendix.

To begin, we denote by $M_{a,f}(t)$ the number of times agent a has been successfully matched with firm f up to time t. Similarly, let $C_{a,f}(t)$ denote the number of times agent a has collided with firm f up to time t. Given this notation, the UCB estimate of agent a for firm f at time t [20] is given by

$$\mathsf{UCB}_{a,f}(t) = \hat{\mu}_{a,f}(t-1) + \sqrt{\frac{2\log\left(1 + \bar{M}_a(t)\log^2(\bar{M}_a(t))\right)}{M_{a,f}(t)}},$$

where $\bar{M}_a(t) = \sum_{f \in F} M_{a,f}(t)$ and $\hat{\mu}_{a,f}(t-1)$ is the empirical average of the payoffs received from successfully matching to firm f up to time $t-1$.

The UCB estimate consists of two parts: (i) the empirical mean, which captures the exploitation aspect; and (ii) an exploration bonus that decreases as $M_{a,f}(t)$ increases. We remark that it does not depend on the number of collisions $C_{a,f}(t)$.

Adversarial Bandit Subroutine

A key component of the proposed methodology is the use of an adversarial bandit (AB) subroutine to address the competitive aspect of the problem. In particular, the AB subroutine

[3]The corresponding algorithmic subroutine `pullModule` is presented in the Appendix.

Algorithm 6 UCB based Decentralized Matching Algorithm (UCB-DMA)

Initialize: $\hat{\mu}_{a,f} = 0,\ M_{a,f} = 0,\ p_{a,f} = 0.5,\ x_{a,f} = 0.5,\ L_{a,f} = 0,\ \forall a \in A,\ f \in F$
for $t = 1$ to T **do**
  for each $f \in F$ **do**
    $\bar{M}_a \leftarrow \sum_{f \in F} M_{a,f}$
    $\mathsf{UCB}_{a,f} \leftarrow \hat{\mu}_{a,f} + \sqrt{\frac{2\log(1+(\bar{M}_a+1)\log^2(\bar{M}_a+1))}{M_{a,f}}}$
  end for
  $\mathsf{ArgUCB}_a \leftarrow \mathrm{ArgDescendingSort}(\{\mathsf{UCB}_{a,f}\}_{f \in F})$
  $i \leftarrow 1$
  while $i \leqslant |F|$ **do**
    $f \leftarrow \mathsf{ArgUCB}_a^{[i]}$
    Sample $P_{a,f} \sim \mathrm{Bernoulli}(p_{a,f})$
    if $P_{a,f} = 0$ **then**
      $(x_{a,f}, p_{a,f}, L_{a,f}) \leftarrow \mathsf{AB\_Subroutine}(P_{a,f}, x_{a,f}, p_{a,f}, L_{a,f}, Y_a)$
    else
      if $P_{a,f} = 1$ **then**
        Request firm f and receive (r_a, Y_a)
        $\hat{\mu}_{a,f} \leftarrow Y_a \cdot \frac{\hat{\mu}_{a,f} M_{a,f} + r_a}{M_{a,f}+1} + (1 - Y_a) \cdot \hat{\mu}_{a,f}$
        $M_{a,f} \leftarrow M_{a,f} + Y_a$
        $(x_{a,f}, p_{a,f}, L_{a,f}) \leftarrow \mathsf{AB\_Subroutine}(P_{a,f}, x_{a,f}, p_{a,f}, L_{a,f}, Y_a)$
        break
      end if
    end if
    $i \leftarrow i + 1$
  end while
  if $i = |F| + 1$ **then**
    Request firm $\mathsf{ArgUCB}_a^{[1]}$ and receive (r_a, Y_a)
    $\hat{\mu}_{a,f} \leftarrow Y_a \cdot \frac{\hat{\mu}_{a,f} M_{a,f} + r_a}{M_{a,f}+1} + (1 - Y_a) \cdot \hat{\mu}_{a,f}$
    $M_{a,f} \leftarrow M_{a,f} + Y_a$
  end if
end for

updates the request probabilities $(p_{a,f})_{f\in F}$ so that the agent stops requesting firms on which collisions are frequent, while ensuring it does not miss out on firms that are achievable. Intuitively, by design, the adversarial bandit algorithm learns to prune arms (firms) that lead to frequent collisions and to request firms where successful matches are likely.

We demonstrate this by analyzing the regret of the AB subroutine, showing that high regret is incurred if the algorithm either prunes too often when successful matching is possible or requests a firm that is unachievable due to the frequent presence of higher-ranked agents. By bounding the regret of the AB subroutine, we immediately obtain a bound on the number of collisions.

We now describe the update scheme for $p_{a,f}(t)$ for any (a, f) at any time $t \in [T]$. In this work, we consider an optimistic mirror descent–based adversarial bandit (AB) subroutine specialized from [72]. Interestingly, such AB algorithms have data-dependent regret bounds [425, 72], unlike other AB algorithms like Exp3 [228, 386]. Since the competition in the matching market is not actually adversarial, such data-dependent regret bounds enable us to characterize the competition more effectively in the analysis rather than just treating competition as adversarial.[4] We note that the proof techniques developed here can also be used to analyze an Exp3-based AB subroutine, but the regret bounds of such an approach will not be as sharp.

For a given agent a, our algorithm associates a separate adversarial bandit (AB) subroutine to every firm $f \in F$. Each AB algorithm has *two arms*, which correspond to the actions of requesting the firm f or pruning it, each of which incurs different losses depending on the outcome. In particular, if $P_{a,f}(t) = 0$, then it receives a fixed loss of 0; if $P_{a,f}(t) = 1$, the loss received is $+1$ or -1 if it collides or matches, respectively.

If we denote the loss received by the AB subroutine associated with (a, f) at time t by $L_{a,f}(t)$, then we have

$$L_{a,f}(t) = P_{a,f}(t)\Big(1 - 2Y_a(t)\Big).$$

Note that $Y_a(t)$ is unknown to any agent before requesting any firm, as it depends on the requests made by other agents.

We note that the request probability $p_{a,f}$ is not updated at every time t, but only when $E^{(\mathsf{c})}_{a,f}(t) = 1$ (i.e., if all firms with a higher UCB index have been pruned). As such, the adversarial bandit algorithms can be seen as operating on a randomized timescale

$$\tau_{a,f}(T) = \{t \in [T] : E^{(\mathsf{c})}_{a,f}(t) = 1\},$$

which corresponds to the time steps on which agent a considers firm f. We also note that $p_{a,f}(t + 1) = p_{a,f}(t)$ if $t \notin \tau_{a,f}(T)$.

For the specific AB algorithm we analyze (which is a version of optimistic mirror descent with a log-barrier regularizer first studied in [425]), the simple setup of the losses leads to

[4] We review the required background on optimistic mirror descent–based AB algorithms in the Appendix along with a result that characterizes the corresponding data-dependent regret bounds in the setting of matching markets.

a closed-form update for the probability of requesting or pruning a firm. In particular, for every $(a, f) \in A \times F$ and $t \in \tau_{a,f}(T)$, the optimistic mirror descent AB subroutine creates unbiased estimates of the losses due to pruning and requesting, denoted by $\hat{L}_{a,f}^{(\mathsf{prune})}(t)$ and $\hat{L}_{a,f}^{(\mathsf{pull})}(t)$, respectively.

In particular, if $P_{a,f}(t) = 1$

$$\hat{L}_{a,f}^{(\mathsf{prune})}(t) = \frac{1 + L_{a,f}(t-1)}{2}, \quad \hat{L}_{a,f}^{(\mathsf{pull})}(t) = \frac{1 - 2Y_a(t) - L_{a,f}(t-1)}{2p_{a,f}(t)} + \frac{1 + L_{a,f}(t-1)}{2}.$$

On the other hand, if $P_{a,f}(t) = 0$ then

$$\hat{L}_{a,f}^{(\mathsf{prune})}(t) = \frac{-L_{a,f}(t-1)}{2(1 - p_{a,f}(t))} + \frac{1 + L_{a,f}(t-1)}{2}, \quad \hat{L}_{a,f}^{(\mathsf{pull})}(t) = \frac{1 + L_{a,f}(t-1)}{2}.$$

The term $\frac{1+L_{a,f}(t-1)}{2}$ is an optimistic prediction of the losses based on the last round of interaction [72]. Given these estimators the probability of requesting a firm is updated as:

$$p_{a,f}(t+1) = (1 - \Lambda_{a,f}(t))x_{a,f}(t) + \Lambda_{a,f}(t)P_{a,f}(t),$$

where

$$x_{a,f}(t) = \left(2 + \xi(t) - \sqrt{4 + \xi(t)^2}\right)(2\xi(t))^{-1},$$

and

$$\xi(t) = \eta\left(\hat{L}_{a,f}^{(\mathsf{pull})}(t) - \hat{L}_{a,f}^{(\mathsf{prune})}(t)\right) + \frac{1}{x_{a,f}(t-1)} - \frac{1}{1 - x_{a,f}(t-1)}$$

is the result of a step of mirror descent with the log-barrier regularizer, and $\Lambda_{a,f}(t) = \frac{\lambda(1-L_{a,f}(t))}{2+\lambda(1-L_{a,f}(t))}$, for $\lambda > 0$, promotes exploration. The algorithmic description of this process is stated in Algorithm 7.

6.3 Bounds on the regret of proposed algorithm

To capture the performance of the algorithm we use the natural notion of *stable regret* as introduced in [256]. More formally, the stable regret accrued by any agent $a \in A$ is

$$\mathbb{E}[\mathcal{R}_a(T)] = \mathbb{E}\left[\sum_{t=1}^{T} u_{a,f_a^*} - \sum_{t=1}^{T} u_{a,f_a(t)}\right] \leqslant \sum_{f \in \underline{\mathbb{F}}_a} \Delta_a(f)\mathbb{E}[M_{a,f}(T)] + u_a(f_a^*) \sum_{f \in \mathcal{F}} \mathbb{E}[C_{a,f}(T)], \tag{6.2}$$

where $\Delta_a(f) = u_a(f_a^*) - u_a(f)$ is the gap between the mean that agent a gets upon successfully matching with its stable match as compared firm f. If there are no collisions, then this regret definition is same as that used in stochastic bandits literature (cf. [228]). In the following theorem, we present the regret of any agent using Algorithm 6:

Algorithm 7 Pull Probability Update Module: AB_Subroutine

Input: $P_{a,f}, x_{a,f}, p_{a,f}, L_{a,f}, Y_a$
Parameters: $\eta \leqslant \frac{1}{50},\ \lambda = 8\eta$
if $P_{a,f} = 0$ **then**
  $\hat{L}_{a,f}^{(\text{prune})} \leftarrow \frac{-L_{a,f}}{2(1-p_{a,f})} + \frac{L_{a,f}+1}{2}$
  $\hat{L}_{a,f}^{(\text{pull})} \leftarrow \frac{1+L_{a,f}}{2}$
  $L_{a,f} \leftarrow 0$
else
  if $P_{a,f} = 1$ **then**
    $\hat{L}_{a,f}^{(\text{prune})} \leftarrow \frac{1+L_{a,f}}{2}$
    $\hat{L}_{a,f}^{(\text{pull})} \leftarrow \frac{1-2Y_a-L_{a,f}}{2p_{a,f}} + \frac{1+L_{a,f}}{2}$
    $L_{a,f} \leftarrow 1 - 2Y_a$
  end if
end if
$\xi \leftarrow \eta(\hat{L}_{a,f}^{(\text{pull})} - \hat{L}_{a,f}^{(\text{prune})}) + \frac{1}{x_{a,f}} - \frac{1}{1-x_{a,f}}$
$x_{a,f} \leftarrow \frac{2+\xi-\sqrt{4+\xi^2}}{2\xi}$
$\Lambda_{a,f} \leftarrow \frac{\lambda(1-L_{a,f})}{2+\lambda(1-L_{a,f})}$
$p_{a,f} \leftarrow (1-\Lambda_{a,f})x_{a,f} + \Lambda_{a,f}P_{a,f}$
Output: $L_{a,f}, x_{a,f}, p_{a,f}$

Theorem 6.3.1. *Suppose every agent $a \in A$ uses Algorithm 6. Then, for any $i \in [K]$*

$$\sum_{j=1}^{i} \sum_{a \in \mathcal{A}_j} \mathbb{E}[\mathcal{R}_a(T)] = \mathcal{O}\left(C_i |F||A| \log(T)\left(1 + \frac{1}{\Delta^2}\right)\right),$$

where $\Delta = \min_{a,f} \Delta_{a,f}$ and C_i is a constant dependent on market M_i and $C_1 < C_2 < ... < C_K$.

We see that the regret of any agent $a \in A$ is logarithmic in the horizon T, which matches the lower bound for single-player stochastic bandit algorithms [223]. As such, perhaps surprisingly, we observe that in α-reducible markets, it is possible for agents to learn while competing without incurring drastically worse regret in the long run. It is interesting to note that the learning of an agent depends on its position in the market according to preferences (Remark 6.1.2). An agent low in the hierarchy incurs more regret during the learning process due to the agents higher up in the hierarchy, mainly driven by the larger number of collisions incurred while waiting for agents higher in the hierarchy to stop exploring. We note that, in the worst case, the constant C_i can grow exponentially with the number of agents in the market. This is a consequence of the proof technique and not a fundamental limitation of the algorithmic design paradigm, as we show through numerical studies in the

next section. We leave it as future work to establish tighter regret bounds in terms of the number of agents. In Appendix Chapter E, we also show that if Algorithm 6 uses a SB subroutine based on Thompson Sampling, then a similar regret guarantee can be obtained. We now present a sketch of the proof of Theorem 6.3.1.

Sketch of the proof. Before presenting the sketch, we first define a few notations to make the exposition clear. Let

$$M_{a,\underline{\mathbb{F}}_a}(T) = \sum_{f\in\underline{\mathbb{F}}_a} M_{a,f}(T), \quad M_{a,\overline{\mathbb{F}}_a}(T) = \sum_{f\in\overline{\mathbb{F}}_a} M_{a,f}(T).$$

Moreover, for any $a \in A$, define

$$H_{a,f_a^*}(t) = \left\{\exists a' \in A \text{ s.t. } u_{f_a^*}(a') \geqslant u_{f_a^*}(a),\ f_{a'}(t) = f\right\},$$

which is the event that characterizes whether any other more preferred agent has requested the stable match of agent a at time t.

Against this backdrop, we now present the following crucial lemma:

Lemma 6.3.1. *Suppose every agent uses Algorithm 6 then the following holds:*

(L1) *For any $i \in [K]$, the cumulative regret can be decomposed as*

$$\sum_{j=1}^{i}\sum_{a\in\mathcal{A}_j}\mathbb{E}[\mathcal{R}_a(T)] = \mathcal{O}\Bigg(\sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\left(\mathbb{E}[M_{a,\underline{\mathbb{F}}_a}(T)] + \sum_{\substack{f\in F\\ f\neq\{f_a^*\}}}\mathbb{E}[C_{a,f}(T)] + \mathbb{E}[\sum_{t=1}^{T} H_{a,f_a^*}(t)]\right)\Bigg);$$

(L2) *For any $i \in [K]$, the expected matches with suboptimal firm satisfies*

$$\sum_{j=1}^{i}\sum_{a\in\mathcal{A}_j}\mathbb{E}[M_{a,\underline{\mathbb{F}}_a}(T)] = \mathcal{O}\left(\sum_{j=1}^{i}\sum_{a\in\mathcal{A}_j}\left(|\underline{\mathbb{F}}_a|\log(T)\left(1+\frac{1}{\Delta^2}\right) + \mathbb{E}\left[\sum_{t=1}^{T} H_{a,f_a^*}(t)\right]\right)\right),$$

(L3) *The expected number of collisions between for any agent $a \in A$ satisfies*

$$\sum_{f\in F}\mathbb{E}[C_{a,f}(T)] = \mathcal{O}\left(|F|\log(T) + \mathbb{E}\left[M_{a,\underline{\mathbb{F}}_a}(T) + M_{a,\overline{\mathbb{F}}_a}(T) + \sum_{t=1}^{T}\mathbb{1}\left(H_{a,f_a^*}(t)\right)\right]\right),$$

(L4) *For any $i \in [K]$, we have*

$$\sum_{j=1}^{i}\sum_{a\in\mathcal{A}_j}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f_a^*}(t)\right)\right] = \mathcal{O}\left(C_i\left(\sum_{j=1}^{i}|\mathcal{A}_j|\right)\log(T)\left(1+\frac{1}{\Delta^2}\right)\right),$$

where C_i is a constant dependent on market M_i such that $C_1 < C_2 < ... < C_K$.

(L5) *For any $i \in [K]$, we have*

$$\sum_{j=1}^{i} \sum_{a \in \mathcal{A}_j} \sum_{f \in \overline{\mathbb{F}}_a} \mathbb{E}[M_{a,f}(T)] \leqslant \mathcal{O}\left(C_i \left(\sum_{j=1}^{i} |\mathcal{A}_j| \right) |F| \log(T) \left(1 + \frac{1}{\Delta^2} \right) \right).$$

Theorem 6.3.1 is proved using **(L1)**-**(L5)** from Lemma 6.3.1. Note that **(L1)** follows from (6.2) and the definition of $H_{a,f_a^*}(t)$. From **(L1)** we see that to bound the regret we need to consider three components: (i) expected number of matchings with suboptimal firms, (ii) expected number of collisions with any firm other than stable match, (iii) the *potential collisions* at the stable match[5]. **(L2)** bounds the expected number of matchings with suboptimal firms. Note that the total matchings between agent a and firm f is $M_{a,f}(T) = \sum_{t=1}^{T} \mathbb{1}\left(Y_a(t) = 1, f_a(t) = f\right)$. Thus, we present the following lemma which plays a key role in the proof of **(L2)**:

Lemma 6.3.2. *The event that agent a chooses the firm $f \in \underline{\mathbb{F}}_a$ and successfully matches at time $t \in [T]$ satisfies*

$$\begin{aligned} &\{Y_a(t) = 1, f_a(t) = f\} \\ &\subset \left\{Y_a(t) = 1, \mathsf{UCB}_{a,f_a^*}(t) \leqslant \mathsf{UCB}_{a,f}(t)\right\} \cup \{E_{a,f}^{(r)}(t) = 1, E_{a,f_a^*}^{(r)}(t) = 0\}. \end{aligned}$$

Lemma 6.3.2 separates the challenge associated with uncertainty from that of competition. Note that the first event on the right-hand side is standard in the analysis of the UCB algorithm [228]. Meanwhile, the other event corresponds to the case when the stable firm is pruned by agent a to avoid potential collisions. To bound the latter event, we use the regret bounds for the adversarial bandit subroutine (see Appendix).

To bound **(L3)**, we use the path-length-based regret bounds from [72, 425] for the adversarial bandit subroutine. Meanwhile, to bound **(L4)**, we use the α-reducibility assumption and **(L2)**. In particular, the α-reducibility assumption induces a hierarchy in the market as described in Remark 6.1.2. This decomposition reduces the bound in **(L4)** to an appropriate accounting of the number of matches with suboptimal firms via an induction argument. Finally, **(L5)** follows again due to the hierarchy induced by α-reducibility and using **(L2)**–**(L4)**.

6.4 Experimental Study

In this section, we present numerical experiments that demonstrate and validate the results presented in this chapter. Moreover, we observe that our algorithm performs surprisingly well in general market structures, i.e., in markets that are not α-reducible. We leave it as future work to establish regret bounds for the proposed algorithms in such general markets.

[5]by potential collision at stable match we mean total number of collision that would have been faced by an agent at its stable firm had it always requested the stable firm

In both sets of experiments, we consider a market comprising 5 agents and 5 firms. We study the following two settings:

(S-I). Randomly initialized preferences for agents and randomly initialized (but uniform) preferences for firms. This setting ensures that the market is α-reducible.

(S-II). Randomly initialized preferences for both agents and firms. In this setting, we specifically consider cases where α-reducibility does *not* hold. This provides directions for future research in this area.

In our simulations, for every agent, we randomly sample the preference ordering of firms and assign a mean reward in $[0, 5]$ such that a successful match with the most preferred firm gives mean reward 5, the least preferred firm gives mean reward 0, and the mean rewards from other firms are equally spaced in between. The rewards follow a normal distribution with variance 1. We run both Algorithm 6 and Algorithm 16 25 times for two randomly sampled preference orderings in each of **(S-I)** and **(S-II)**.

In Figure 6.1, we consider **(S-I)** and observe the performance of the algorithms. We see that the mean regret (averaged over 25 runs) accumulated by the algorithms saturates very quickly and agents identify their stable matches.

In Figure 6.2, we consider **(S-II)** and observe the performance of the algorithms. Surprisingly, even without the α-reducibility structure, the mean regret[6] (averaged over 25 runs) accumulated by the algorithms saturates quickly and agents identify their stable matches. This presents an opportunity to further explore the algorithms presented in this chapter for general markets.

Furthermore, in both **(S-I)** and **(S-II)**, we observe that TS-DMA exhibits higher variance but converges faster than UCB-DMA. This is because, compared to UCB-DMA, TS-DMA very rarely encounters scenarios where all firms get pruned by the adversarial bandit module. We also note that in some cases the regret can be negative (which is desirable), as shown in Figure 6.1(c) for the red agent.

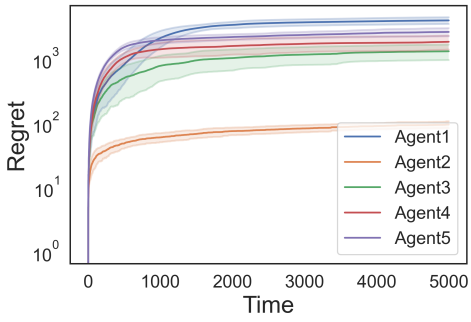
6.5 Concluding Remarks

We consider the problem of bandit learning in two-sided matching markets comprising agents and firms. We focus on the setting where agents have unknown preferences over the firms. In this chapter, we present a simple design principle for a decentralized, communication- and coordination-free algorithm for learning in two-sided matching markets.

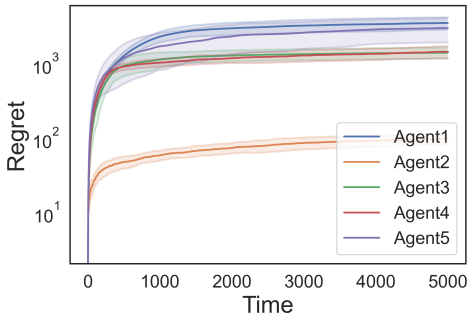
The primary challenge in learning in two-sided matching markets is to balance exploration, exploitation, and collision avoidance. We embed these properties into the algorithm through the novel idea of blending a stochastic bandit subroutine with an adversarial bandit subroutine. The stochastic bandit subroutine manages the exploration-exploitation trade-off, while the adversarial bandit subroutine limits collisions.

As an instance of this design principle, we present an algorithm that uses a stochastic bandit subroutine based on UCB and an adversarial bandit subroutine based on the Opti-

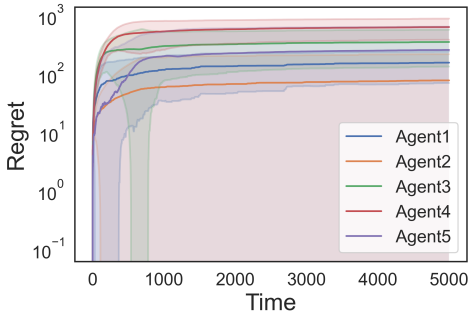
[6]Here, mean regret refers to the agent-optimal stable regret [257]



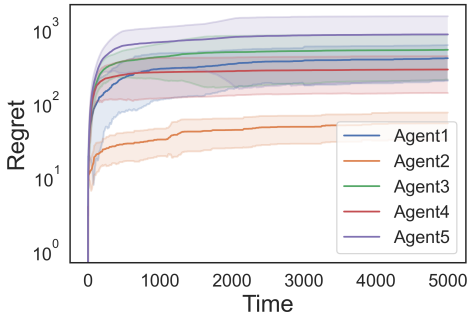
(a) UCB-DMA(Algorithm 6)



(b) UCB-DMA(Algorithm 6)



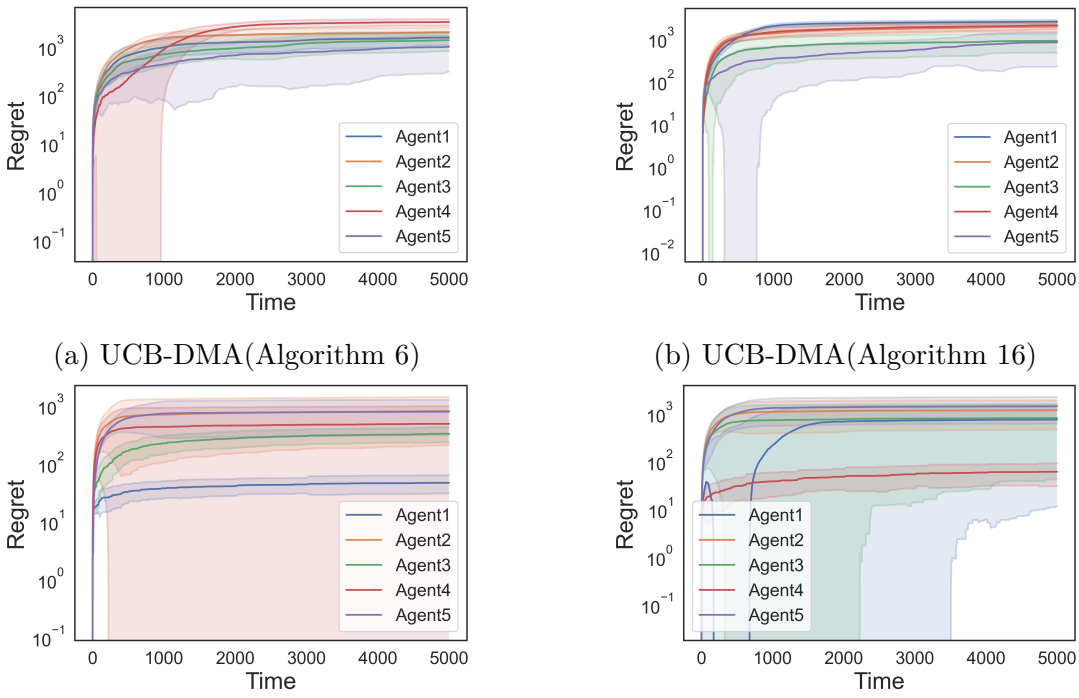
(c) TS-DMA(Algorithm 16)



(d) TS-DMA(Algorithm 16)

Figure 6.1: Performance of UCB-DMA (Algorithm 6) and TS-DMA (Algorithm 16) where the α-reducibility condition is satisfied. We simulated the algorithms for two randomly generated preference orderings that satisfy the α-reducibility condition. The simulation results for one preference ordering are presented in the left column, and for the other in the right column. The bold lines and the corresponding shaded regions denote the mean regret and variance of regret for the agents over 25 runs of the algorithms.

mistic Mirror Descent algorithm. We show that if the agents' preferences satisfy a certain structural property known as α-reducibility, then these algorithms incur regret logarithmic in the time horizon.



(a) UCB-DMA(Algorithm 6)

(b) UCB-DMA(Algorithm 16)

(c) TS-DMA(Algorithm 16)

(d) TS-DMA(Algorithm 16)

Figure 6.2: Performance of UCB-DMA (Algorithm 6) and TS-DMA (Algorithm 16) where the α-reducibility condition is **not** satisfied. We simulated the algorithms for two randomly generated preference orderings that *do not* satisfy the α-reducibility condition. The simulation results for one preference ordering are presented in the left column, and for the other in the right column. The bold lines and the corresponding shaded regions denote the mean regret and variance of regret for the agents over 25 runs of the algorithms.

Chapter 7

Learning in Time-varying Matching Markets

Similar to Chapter 6, this chapter studies matching markets where agents may not initially know their underlying preferences. Unlike the previous chapter's emphasis on decentralized algorithms, we simplify the setting by considering a centralized platform that coordinates the matching process. The primary focus here is on enabling agents to learn effectively when the unknown preferences may itself evolve over time. Agents observe only their local (and potentially noisy) feedback based on the firm assigned by the platform. They then update their preference estimates and report them back to the platform, which uses this information to determine matches in subsequent rounds.

In a recent work, [256] introduced the competing bandits framework to study the interaction of learning and competition in two-sided online matching markets. This framework is inspired by many modern matching markets, which are two-sided and involve multiple rounds of interactions. In each round, agents submit their preferences, and the market is cleared based on these submitted preferences. After the market clears and the matching is made, agents observe the final outcome or reward from their match, which may be subject to statistical uncertainty.

Thus, the feedback agents receive is similar to bandit feedback. The players need to learn their preferences over the other side of the market from this bandit feedback, with the added complexity that the set of options they can explore is influenced by the other agents' actions. The competing bandits framework allows us to study the interaction of learning when multiple agents are simultaneously learning and competing. The ultimate goal is to design algorithms that enable agents to learn their underlying preferences and help the market converge to a stable matching despite the complex interplay of competition and learning.

Contribution We study an extension of the competing bandits problem in which the underlying unknown preferences of agents may vary over time. While [256] introduced the competing bandits framework, they assume that agents' preferences remain fixed. However,

this assumption is often unrealistic, as preferences are likely to evolve in many practical settings.

In this work, we develop an algorithm and provide theoretical guarantees for the setting where agents' preferences are time-varying. While [156] study a similar problem under a serial dictatorship preference structure and assume smoothly varying preferences, *our proposed methodology is significantly more general—it accommodates arbitrary preference structures and arbitrary variation patterns, while remaining computationally efficient.* This flexibility and efficiency constitute our main contribution to this line of work.

We show that with our approach, each agent incurs a regret of $\widetilde{O}(L_T^{1/2}T^{1/2})$, where T is the number of rounds and L_T denotes the total number of changes in the preference orderings across all agents over T rounds. This matches the optimal regret rate for non-stationary single-agent bandit learning [22].

Additional Related Works

Here, we discuss the literature on time-varying bandits and learning in matching markets.

Time Varying Multi-armed Bandits. The non-stationary Multi-Armed Bandits (MAB) problem was first studied by [21], who proposed a variant of the EXP3 algorithm called EXP3.S. They showed that EXP3.S achieves minimax-optimal regret of $\mathcal{O}(\sqrt{KLT})$ up to logarithmic factors, where K is the number of arms, L is the number of abrupt changes in the reward distribution, and T is the number of rounds.

Subsequent works, such as [153] and [10], demonstrated that alternative strategies based on sliding windows or restart mechanisms can also attain minimax-optimal regret. More recently, [22] proposed a near-optimal algorithm that does not require prior knowledge of the number of changes.

In another line of work, [51] and [219] considered the setting of non-stationary MABs with slowly varying reward distributions and showed that a regret of $\widetilde{O}\left(P_T^{1/3}T^{2/3}\right)$ is achievable, where P_T denotes the total variation in the reward sequence over time.

The non-stationary linear bandits problem has also been explored in works such as [93], [358], and [451]. These studies introduced methods like sliding-window least squares, weighted UCB, and restart-based approaches, respectively, and established that a regret of $\widetilde{\mathcal{O}}\left(P_T^{1/4}T^{3/4}\right)$ can be attained.

Learning in Two-sided Matching Markets. Learning in the context of two-sided matching markets is a relatively new and active area of research [116, 256, 257, 188, 274, 35, 84, 115]. The earliest work in this direction was by [116], which employed multi-armed bandit algorithms to learn preferences in matching markets. However, it was only more recently that [256] formalized the problem framework.

The existing literature in this domain can broadly be classified into two categories. The first class assumes that the true underlying preferences of the participants over the firms

remain fixed over time [256, 188, 84, 35, 363, 274, 215]. The second class considers the more general and realistic setting where the underlying preferences may vary with time [301, 156].

Interestingly, the literature on learning in time-varying matching markets remains relatively sparse [301, 156]. [301] consider the setting of a *Markov Matching Market*, where agents' preferences depend on an underlying context that evolves according to a Markovian process, influenced by the planner's policy. In contrast, [156] study learning in *smoothly varying* matching markets, where agents' underlying preferences are allowed to change gradually, bounded by a fixed threshold at each time step. While their framework is decentralized, it focuses on a specific preference structure—namely, a serial dictatorship—limiting its generality.

7.1 Problem Formulation

Consider a two-sided matching market consisting of *players* and *arms*, which we will refer to as the two sides of the market. Each side derives some positive utility when matched with a participant from the other side.

Formally, let the set of N players be denoted by $\mathcal{N} = \{p_1, p_2, \ldots, p_N\}$, and the set of K arms by $\mathcal{K} = \{a_1, a_2, \ldots, a_K\}$. A distinguishing feature of two-sided matching markets is that each side possesses preferences over the participants on the opposite side.

The preferences of an arm $a_j \in \mathcal{K}$ over the players are represented by a utility vector $\pi_j \in \mathbb{R}_+^N$, where $\pi_j(i)$ denotes the utility that arm a_j obtains when matched with player p_i. Similarly, the preferences of a player $p_i \in \mathcal{N}$ over the arms are encoded by a utility vector $\mu_i \in \mathbb{R}_+^K$, where $\mu_i(k)$ denotes the utility that player p_i receives when matched with arm a_k.

For the sake of concise notation, if arm a_j prefers player p_i over player $p_{i'}$, we denote this as $p_i \succ_j p_{i'}$. Similarly, if player p_i prefers arm a_k over arm $a_{k'}$, we write this as $a_k \succ_i a_{k'}$.

In what follows, we first introduce the framework of two-sided matching markets and present relevant background in Section 7.1. Subsequently, we formulate the problem of learning in two-sided time-varying matching markets.

Preliminaries on Matching Markets

To formally present the setup, we begin by recalling some relevant concepts from the literature on two-sided matching markets. A *matching* $\mathbf{m} : \mathcal{N} \to \mathcal{K}$ is defined as an injective map such that $\mathbf{m}(p) = a$ denotes player $p \in \mathcal{N}$ is matched with arm $a \in \mathcal{K}$.

Definition 7.1.1 (Blocking Pair)**.** *We say a tuple $(p_i, a_j) \in \mathcal{N} \times \mathcal{K}$ is a* blocking pair *for a matching $\boldsymbol{m}$ if player p_i is matched to arm $a_{j'} = \boldsymbol{m}(p_i)$, but*

$$a_j \succ_i a_{j'} \quad \textit{and} \quad a_j \textit{ is either unmatched or } p_i \succ_j \boldsymbol{m}^{-1}(a_j).$$

In this case, we say that the triplet $(p_i, a_j, a_{j'})$ blocks *the matching $\boldsymbol{m}$.*

Having defined the notion of blocking pair, we are now ready to define stable matching.

Definition 7.1.2 (Stable matching)**.** *A matching $\boldsymbol{m}$ is called* stable *if there is no blocking pair. Alternatively, a matching is called* unstable *if there exists at least one blocking pair for it.*

Gale and Shapley in 1962 proposed a polynomial time algorithm – referred as Deferred-Acceptance (DA) algorithm – to find a stable matching. Without enforcing any specific assumptions on the underlying preference structure, the stable matching is not unique. Therefore, we define the notion of valid partners of a player which captures the set of all arm to which it can match in some stable matching.

Definition 7.1.3 (Valid partner)**.** *Given the full preference rankings of arms and players, we call arm a_j to be a valid partner of player p_i if there exists a stable matching $\boldsymbol{m}$ such that $\boldsymbol{m}(p_i) = a_j$.*

We now define two important types of matching which are very crucial for subsequent exposition.

Definition 7.1.4 (Optimal matching and pessimal matching)**.** *We say a matching $\overline{\boldsymbol{m}}$ to be* optimal *matching if every player is matched to its most preferred valid partner. Similarly we say a matching $\underline{\boldsymbol{m}}$ to be* pessimal *matching if every player is matched to its least preferred valid partner.*

To identify the optimal and pessimal stable matchings, one can initialize the Deferred Acceptance algorithm from the players' side and the arms' side, respectively. Notably, if the stable matching is unique, then the optimal and pessimal matchings coincide. Interestingly, [208] provide necessary and sufficient conditions on the underlying preference structure that guarantee the uniqueness of the stable matching.

Learning in Time Varying Matching Markets

While a stable matching can be identified directly by employing the Deferred-Acceptance algorithm when players know their preferences, in many modern matching markets, players are often unaware of their true preferences. Furthermore, these underlying preferences can be time-varying.

In this chapter, we study the problem of bandit learning in time-varying matching markets, where the objective is to find a stable match through repeated interactions.

For example, consider an online labor market where the two sides are employers (players) and freelancers (arms). Due to the scale of the system, employers do not know *a priori* the quality of work provided by a freelancer and must repeatedly interact with them to learn. Moreover, the inherent quality of a freelancer's work may change in a nonstationary manner due to factors such as health issues or personal circumstances. Consequently, employers must adapt to these nonstationary changes while updating their preference estimates.

We formulate the repeated market setting as follows. We assume that the preferences of arms are *fixed* and are *common knowledge*. However, the preferences of players are *unknown*

and *time-varying*. The players and arms interact with each other over a total of T rounds, indexed by t. At any time t, the true underlying preference of a player $p_i \in \mathcal{N}$ over an arm $a_j \in \mathcal{K}$ is encapsulated by the mean reward of their interaction, denoted by $\mu_{i,t}(j)$, which is unknown. The players repeatedly interact with arms through a platform to learn these underlying utilities.

At every round t, the players submit their *estimates* of their preferences over the arms to a platform based on past rounds of interaction. The platform then computes a stable matching m_t based on the submitted preferences and assigns the players to the arms accordingly. Upon being assigned an arm $m_t(i)$, player p_i pulls the arm $m_t(i)$ and receives a stochastic reward $X_{i,m_t(i)}$ sampled from a 1-sub-Gaussian distribution with mean $\mu_{i,t}(m_t(i))$. The players use the observed reward to update their estimates of preferences over the arms.

In order to evaluate the performance of any online learning algorithm this setup, we introduce the relevant notion of *regret*. Unlike the single player learning setting, in a two-sided matching market, all of the players cannot be assigned their most preferred arm due to the misaligned preference structure of the different players and arms. With this consideration, [256] proposed a regret metric for the stationary preference setting which we extend here to non-stationary setting. This notion of regret, termed as *arm-stable regret* with respect to a stable matching $\mathbf{m}^\dagger = (\mathbf{m}_t^\dagger)_{t\in[T]}$ corresponding to true underlying preferences, is given by

$$R_T^i(\mathbf{m}^\dagger) = \sum_{t=1}^{T} \mu_{i,t}(\mathbf{m}_t^\dagger(i)) - \sum_{t=1}^{T} \mu_{i,t}(m_t(i)), \tag{7.1}$$

where $\mu_{i,t}(\mathbf{m}_t^\dagger(i))$ is the mean reward of the arm that would be assigned to a player p_i in a stable matching outcome if the players know the true underlying preference.

In any two-sided matching market, there are two important matchings: the *player-optimal* and the *player-pessimal* matchings. Correspondingly, there are two associated regret metrics: the *player-optimal regret*, defined as the regret with respect to the optimal matching $\overline{\mathbf{m}}$, and the *player-pessimal regret*, defined as the regret with respect to the pessimal matching $\underline{\mathbf{m}}$. As noted in [256], achieving sub-linear player-optimal regret is not possible even in a static environment without imposing strong assumptions on the underlying preference structure. Therefore, we adopt the *player-pessimal regret* as our performance metric.

To quantify the non-stationarity in the environment, we introduce the notion of *time-variation*, which captures the variability in the underlying true preferences. In single-player bandit learning, time-variation is typically characterized by quantities such as the total variation [51] or the total number of changes [22]. However, in the matching setting, the interdependence among players is such that changes in other players' preferences can lead to lower rewards than the pessimal match, even without any change in a given player's own preferences. These occurrences are not directly bounded by the magnitude of variation but rather depend on the number of changes in the order of players' preferences.

Therefore, we adopt the number of changes as the measure of variation for the competing bandits setting. Specifically, we define the variation measure as the total number of changes

across all the players:

$$L_T = \sum_{t=2}^{T} \sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{K}} \mathbb{1}[\mu_{i,t}(j) \neq \mu_{i,t-1}(j)], \tag{7.2}$$

where $\mathbb{1}[\cdot]$ is the standard indicator function. Such a definition accounts for all the variations in the market and its impact uniformly across all the players. Naturally, if the players' preferences are fixed then $L_T = 0$.

In any realistic scenario, the preferences do not change arbitrarily over time. Furthermore, to obtain meaningful quantitative guarantees, it is necessary to assume certain structure on the underlying preference dynamics. Against this backdrop, we make the following assumption on the preferences of the players.

Assumption 7.1.1. *We make the following assumptions on the mean reward of arms*

(i) The mean reward of arms are bounded. That is $\mu_{i,t}(k) \leqslant \bar{\mu}$ for all $t \in [T], i \in [N], k \in [K]$;

(ii) The gap between the arms for any player is always greater than Δ. That is, $0 < \Delta = \min_t \min_{i,k,k'} |\mu_{i,t}(k) - \mu_{i,t}(k')|$.

Some remarks about Assumption 7.1.1 are in order. Assumptions 7.1.1-(i) and 7.1.1-(ii) are necessary to obtain any meaningful regret guarantees for competing bandits in non-stationary matching markets without imposing structure on preferences. Furthermore, we argue that these are practical assumptions on the rewards obtained in real-world applications. Our focus here is on changing preferences, on which we impose no restrictions whatsoever. We note that we do not require any assumptions on the magnitude of instantaneous changes in the preferences of any player, which is typically assumed in many prior works on non-stationary bandits [219, 153]. Moreover, the changes in the underlying true preferences of players can occur asynchronously.

7.2 Algorithm and Results

There are three key challenges associated with designing an effective algorithm that works in this setting: uncertainty, competition, and non-stationarity. Overcoming each of these challenges individually has been studied extensively in the existing literature. However, the key challenge is to develop a provably effective algorithm that can overcome all of these challenges in a holistic manner. In particular, any single player cannot explore independently to learn its preferences because the arm it gets matched with is determined by the submitted preferences of all players, who are simultaneously exploring different arms. Furthermore, this challenge is exacerbated by the inherent non-stationarity of the environment, as a change in the preference of any single player can affect the stable matching. Interestingly, we show that

the proposed algorithmic design, to be presented, handles these challenges while achieving low player-pessimal regret.

Algorithmic Description

Our algorithm is an extension of the bandit learning algorithm proposed by [256] for the stationary setting, where players repeatedly interact with the market platform by submitting a preference order over the arms. After receiving the preferences, the platform assigns players to arms according to a stable matching computed using the Deferred-Acceptance (DA) algorithm on the set of preferences submitted by the players. Upon being matched to an arm, the players receive a reward, which they then use to update their preferences before the next round. The key idea in [256] lies in how the players compute their preferences, enabling them to efficiently explore the arms and allowing the market to converge to a stable outcome despite competition among the players. Specifically, each player submits a preference order based on the Upper Confidence Bound (UCB) estimates of the mean rewards of the arms, computed using the collected rewards from previous time steps.

We extend this algorithm to the time-varying setting by adopting a restart strategy, which is a widely used technique in time-varying bandit learning [451]. The core idea of the restart strategy is to reset the base algorithm (such as the UCB algorithm) after a fixed period H. If the number of changes L_T is known, then the restart period H can be chosen optimally a priori. We adopt this straightforward approach to handle the time variations in the competing bandits setting. Specifically, in the proposed algorithm, *the market platform ensures that players restart their local UCB algorithms after every H rounds, and between two restarts, the players run their local UCB algorithms exactly as in the stationary setting of [256].* The complete algorithm is shown in Algorithm 8.

We now introduce the algorithm more formally. Let $T_{i,t}(j)$ denote the number of times player p_i is matched with arm a_j by the platform after the latest restart. After every restart, $T_{i,t}(j)$ is assigned to zero, and $T_{i,t}(j)$ is incremented by one whenever player p_i is matched with arm a_j. Let $\hat{\mu}_{i,t}(j)$ denote the mean of the rewards observed whenever player p_i pulls arm a_j after a restart and till $t-1$. Let S_t denote the latest restart before time step t. Then, $\hat{\mu}_{i,t}(j)$ is given by

$$\hat{\mu}_{i,t}(j) = \frac{\sum_{k=S_t}^{t-1} X_{i,m_k} \mathbb{1}[m_k(i) = j]}{T_{i,t-1}(j)}.$$

Then, for a player p_i, the upper confidence bound for arm a_j in the current period is given by

$$u_{i,t}(j) = \begin{cases} \infty, & \text{if } T_{i,t}(j) = 0, \\ \hat{\mu}_{i,t}(j) + \sqrt{\frac{3\log(t-S_t)}{2T_{i,t-1}(j)}}, & \text{otherwise.} \end{cases} \tag{7.3}$$

At the beginning of every round t, each player updates its UCB for the arms as in (7.3). The players then compute their respective rank ordering $\hat{r}_{i,t}$ according to the updated UCBs.

Algorithm 8 Restart Competing Bandits (RCB) Algorithm

Input: Common restart period H
Set $k \leftarrow 1$
for $t = 1$ to T **do**
  if $k = 1$ **then**
    Set $T_{i,t}(j) \leftarrow 0$ for all $i \in [N], j \in [K]$
  end if
  Each player i computes UCB values $u_{i,t}(j)$ for all $j \in [K]$ using Eq. (7.3)
  Each player i computes rank ordering $\hat{r}_{i,t}$ using $u_{i,t}(j)$ as per Eq. (7.4)
  Match players to arms using optimal stable matching m_t computed from $\hat{r}_{i,t}$ and π_j
  for each player $i \in [N]$ **do**
    Update $T_{i,t+1}(m_t(i)) \leftarrow T_{i,t}(m_t(i)) + 1$
  end for
  if $k \leqslant H$ **then**
    $k \leftarrow k + 1$
  else
    $k \leftarrow 1$
  end if
end for

Specifically, the rank ordering $\hat{r}_{i,t}$ is computed as follows: for any two arms $a_j, a_{j'}$,

$$\hat{r}_{i,t}(j) > \hat{r}_{i,t}(j'), \text{ if } u_{i,t}(j) > u_{i,t}(j'). \tag{7.4}$$

The platform receives these rank orderings from every player, computes a stable matching, and assigns the matched arms to the respective players. This completes a round.

Algorithm 8 has several important features: first, it is a simple algorithm which us computationally efficient and intuitive. Second, the algorithm does not require storing exact rewards obtained in past rounds at the end of both players and platforms. Third, the preferences of players are updated only using local information which is a crucial aspect to ensure privacy and reliability of the system.

Regret Guarantee

Below, we characterize the regret accrued by a player while using the RCB algorithm (Algorithm 8) under time-varying preferences.

Theorem 7.2.1. *Suppose Assumption 7.1.1 holds. Under RCB algorithm (Algorithm 8), with the common restart period $H = L_T^{-1/2} T^{1/2}$, the pessimal regret for each player i is given as*

$$R_T^i = \widetilde{\mathcal{O}}\left(L_T^{1/2} T^{1/2}\left(1 + \frac{1}{\Delta^2}\right)\right).$$

Some comments about Theorem 7.2.1 are in order. First, we note that the constant in the above regret bound can be exponentially large in the size of the market, which is also the case with [256] which considered stationary preferences of players. Second, the achievable regret bounds in a single agent MAB setting is $\widetilde{O}(\sqrt{KL_TT})$ [21]. Our regret bounds are similar to the single agent multi-arm bandit setting up to some constant, although under the assumption that the gap between the arm of any two players is greater than Δ at all the times. As pointed out in [256], the dependence on Δ^2 cannot be improved in general because of the dependence of a player's regret on the gaps of other players in the competing bandit setting. For the same reason, it is not be possible to extend the results to the case where the arm gaps for a player can be arbitrarily close to each other. Third, we note that the regret R_T can be negative, which is desirable as player's can receive better utility than the pessimal matching.

A detailed proof of Theorem 1 can be found in Appendix F. However, we provide a sketch of the proof below.

Proof Sketch: In the following analysis, we outline only the key steps. We begin by analyzing the agent regret for a player p_i within an interval of length H, from one restart to the next. The total regret is then obtained by summing the regrets across all such intervals of length H.

Let the start and end times of the ℓth interval be denoted by t_s^ℓ and t_e^ℓ, respectively. Let R_ℓ^i denote the agent regret incurred by player i within the ℓth interval. Then, the regret of player i in interval ℓ is given by

$$R_\ell^i = \sum_{t=t_s^\ell}^{t_e^\ell} \left(\mu_{i,t}(\underline{\mathbf{m}}_t(i)) - \mu_{i,t}(m_t(i))\right).$$

The critical component in bounding the regret lies in accounting for two key challenges: (i) unlike in the single-agent setting, each player can be assigned an unstable match with lower reward than their valid partners due to the exploration and preference submissions of other players; and (ii) the underlying preferences—and hence the set of stable matchings—are time-varying.

While the first challenge was addressed by [256] for the case of stationary preferences, our analysis extends to handle the non-stationarity in players' preferences, and the added complexity it introduces in how others' preference changes can indirectly lead to unstable matches for a player.

The cornerstone of our analysis is the concept of the *minimal cover* of an unstable match (see Definition F.1.2) computed prior to the first change in preferences across all players within an interval. This concept allows us to isolate the effect of preference changes and bound the regret incurred.

We now present a crucial lemma that bounds the pessimal regret incurred by any player between two restarts. The proof of this lemma is deferred to the appendix.

Lemma 7.2.1. *Suppose Assumption 7.1.1 holds. Then, under the RCB algorithm (Algorithm 8), the pessimal regret for a player i between $(\ell-1)$th restart and ℓth restart is given by*

$$R_\ell^i \leqslant \mathcal{O}(KL_\ell H) + \mathcal{O}\left(K\left(1+\frac{\log(H)}{\Delta^2}\right)\right).$$

The agent-stable regret over all rounds can be bounded by summing the pessimal regrets incurred across the individual intervals. Therefore, we have

$$R_T^i \leqslant \sum_{\ell=1}^{\lfloor T/H \rfloor} R_\ell^i \leqslant \mathcal{O}(L_T H) + \widetilde{\mathcal{O}}\left(\frac{T}{H\Delta^2}\right),$$

where L_T is the total number of preference changes, H is the restart period, and Δ is the minimum preference gap.

Substituting the optimal restart period $H = L_T^{-1/2}T^{1/2}$ yields the final regret bound.

Remark 7.2.1 (Relation to [256])**.** *The bound presented in Lemma 7.2.1 clearly highlights the connection to the stationary setting analyzed in [256]. The lemma quantifies the regret incurred by a player within an interval following a restart. The constant accompanying the main term, $\widetilde{\mathcal{O}}(1+1/\Delta^2)$, is analogous to the constant in the time-invariant case, which corresponds to the sum over the minimal cover induced by the blocking pairs that block an unstable match. The key difference in our setting is that this minimal cover is defined with respect to the set of blocking pairs* prior *to the first change in preferences within the interval. In the absence of time variations, this recovers the constant used in [256]. The first term in Lemma 7.2.1, $\mathcal{O}(L_\ell H)$, captures the additional regret incurred due to time variations. Naturally, this term vanishes when there are no changes in the underlying preferences.*

7.3 Extension to Unknown Time Variation

Algorithm 8 requires knowledge of the time variation L_T in order to compute the optimal restart period H. However, in many real-world applications, L_T is not known *a priori*.

One approach to extend the RCB algorithm to the unknown time variation setting is the *bandits-over-bandits* strategy [451]. The key idea behind this approach is to adaptively tune the restart period by exploring restart periods drawn from an ensemble that ranges from very short to very long intervals. Specifically, a *meta-bandit algorithm* is employed to periodically reset the restart period of the base algorithm, treating each candidate period in the ensemble as an arm. The meta-bandit algorithm selects the restart period by balancing exploration and exploitation, thereby ensuring that the process converges efficiently to the most effective restart period.

The same idea can be extended to learning in two-sided matching markets. The platform can employ a bandit meta-algorithm to explore and adaptively choose the best restart period

for the base RCB algorithm. To ensure that the ensemble of restart periods is sufficiently rich, we consider the ensemble

$$\mathcal{H} = \{H = 2^{j-1} \mid j \in [1, N]\},$$

where $N = \lceil \frac{1}{2} \log T \rceil + 1$.

Let us denote by H^* the optimal restart period defined in Theorem 7.2.1. Then, $H^* = \mathcal{O}(T^{1/2})$ when $L_T = \mathcal{O}(1)$ and $H^* = \mathcal{O}(1)$ when $L_T = \mathcal{O}(T)$. Therefore, the ensemble $\mathcal{H}$ contains the H^* values corresponding to the full range of L_T and is thus sufficiently rich.

To explore the restart periods, the meta-algorithm can partition the total time horizon T into intervals of length $\mathsf{Y} = \mathcal{O}(T^{1/2})$, and within each interval, it explores a specific restart period. A standard bandit algorithm such as EXP3 [21] can be employed to efficiently explore the restart periods and adaptively select the best-performing one.

The key challenge in the competing bandits setting is how to define the reward for the EXP3 meta-algorithm. We discuss this in more detail below. Let the matching under the RCB algorithm with a fixed restart period H^* be denoted by m_t^*. Then, the regret for each player can be split as

$$\begin{aligned} R_T^i &= \sum_{t=1}^{T} \mu_{i,t}(\mathbf{m}_t(i)) - \sum_{t=1}^{T} \mu_{i,t}(m_t(i)) \\ &= \underbrace{\left[\sum_{t=1}^{T} \mu_{i,t}(\mathbf{m}_t(i)) - \sum_{t=1}^{T} \mu_{i,t}(m_t^*(i))\right]}_{\text{Base Regret for } p_i} + \underbrace{\sum_{t=1}^{T} \left[\mu_{i,t}(m_t^*(i))\right] - \sum_{t=1}^{T} \left[\mu_{i,t}(m_t(i))\right]}_{\text{Meta-regret for } p_i}. \end{aligned}$$

The first term corresponds to the regret incurred when following a fixed restart period. The second term captures the difference between the rewards accumulated by adhering to a fixed restart period and those obtained by the meta-algorithm that adaptively tunes the restart period by exploring an ensemble of periods. We refer to this second term as the *meta-regret*.

The current proof technique to bound the base regret requires that all agents restart simultaneously. Therefore, in this setting, we consider that the platform coordinates the restarting period uniformly across all players. Consequently, we can only obtain bounds on the joint regret of all players.

Towards this goal, we define the reward for the meta-algorithm as the sum total of the accumulated rewards over the period Y across all players. The term $\sum_{t=1}^{T} \mu_{i,t}(m_t^*(i))$ corresponds to the total returns with the restart period H^*. Since $H^* \in \mathcal{H}$, the total return must be less than or equal to the returns corresponding to the best restart period for the players. Therefore, the *meta-regret* term can be upper bounded by the actual regret of the meta-algorithm, which can be bounded as in [451, Theorem 3] by $\widetilde{\mathcal{O}}(T^{3/4})$.

Hence, the joint regret when the number of changes is unknown can be bounded as

$$\sum_{i=1}^{\mathcal{N}} R_T^i = \widetilde{\mathcal{O}}\left(N L_T^{1/2} T^{1/2}\right) + \widetilde{\mathcal{O}}(T^{3/4}).$$

The exploration required for tuning the restart period incurs an additional regret on top of the regret bound established in Theorem 7.2.1. Consequently, in the unknown L_T setting, the achievable total regret is dominated by the $\widetilde{\mathcal{O}}(T^{3/4})$ term. It is possible that this regret bound can be improved by designing a more efficient exploration strategy. Moreover, it remains an open problem to derive meaningful bounds on the *individual regret* of the players when the number of changes L_T is unknown.

7.4 Concluding Remarks

In this chapter, we studied the framework of competing bandits in time-varying matching markets. Specifically, we addressed the problem of how players can learn their preferences amidst competition, enabling the platform to converge to a stable matching outcome. While this challenge has been previously explored, we focused on the more realistic setting where players' preferences themselves may vary over time.

To tackle this problem, we proposed the *Restart Competing Bandits* algorithm, which provably achieves sub-linear regret for each player, provided that the total variation in preferences across all players is also sub-linear. A key insight of our work is that non-stationarity does not fundamentally hinder learning in matching markets. To the best of our knowledge, our result is the *first of its kind* for general preference structures.

We also discussed an extension of our algorithm to the setting where the number of preference changes is not known a priori and highlighted several open directions for future research.

Chapter 8

Multi-agent Learning in Congested Networks

Chapters 6–7 addressed multi-agent learning problems under explicit, hard resource constraints. In many practical settings, however, resource limitations manifest implicitly through congestion effects. A canonical example of this phenomenon arises in transportation networks, where users' interactions are naturally modeled by congestion-induced costs that depend on the aggregate usage of shared infrastructure. In this chapter, we focus on such congestion-based models and develop multi-agent learning algorithms that enable users to adaptively learn optimal routing decisions in presence of congestion effects introduced due to decisions by other users.

Traffic assignment models (TAMs) [104, 422, 26, 143, 114, 7, 120] play a central role in congestion modeling for transportation networks, by informing crucial decisions about infrastructure investment, capacity management, and tolling for congestion regulation. The central dogma behind this modeling approach is that self-interested travelers select routes with minimal *perceived* latency (i.e., the Wardrop or user equilibrium), which can be modeled as deterministic [104, 422] or stochastic [114, 120, 7, 26, 143]. Empirical studies confirm that stochastic TAMs achieve greater success at interpreting congestion levels, compared to their deterministic counterparts [383].

There exist two dominant modeling paradigms in TAM: the route-based model [104, 120, 114, 434]—where each traveler makes a single choice between set of available routes from origin to destination—and the arc (or edge) based model [26, 456, 325, 280, 279]—where the traveler sequentially makes routing decision at each node on the network, based on their perception of arc latencies. There are two major drawbacks of route-based models on real-world networks: route correlation and route enumeration. Specifically, the utility generated from different routes is correlated due to overlapping arcs on different routes. Moreover, exhaustive route enumeration is prohibitive in terms of computational cost, memory storage, and information acquisition, since the number of routes in a traffic network can be exponential in the number of arcs.

To avoid explicit route enumeration, Akamatsu [7] proposed the first arc-based stochastic

TAM, which was further generalized by Baillon and Cominetti [26]. More recently, Fosgerau et al. and Mai et al. [143, 280] presented similar arc-based models based on dynamic discrete choice analysis, which are mathematically similar to the models proposed by Akamatsu [7] and Baillon and Cominetti [26]. However, these models suggest that travelers take cyclic routes with positive probability. To overcome this fundamental modeling challenge, Oyama et al. [326, 324] recently proposed various methods to explicitly avoid routing on cyclic routes. Unfortunately, these methods either do not apply beyond acyclic graphs [324] or restrict the set of feasible routes, at the expense of modeling accuracy [326], or restrictive assumptions on cost structure [26]. Sequential arc selection models in network routing have also been studied by Calderone et al. [77, 78] where each arc selection is accompanied by stochastic transitions to the next arc, and a deterministic transition cost. This stands in contrast to the stochastic TAM literature, where transitions from arc to arc are assumed deterministic and the travel cost (latency) is assumed stochastic.

In this work, we propose an arc-based stochastic TAM that explicitly avoids cycles by considering routing on a directed acyclic graph derived from the original network, henceforth referred to as the *Condensed Directed Acyclic Graph* (CoDAG). The CoDAG representation duplicates an appropriate subset of nodes and arcs in the original network, to explicitly avoids cycles while preserving all feasible routes. Travelers sequentially select arcs on the CoDAG network at every intermediate node, based on perceived arc latencies. This route choice behavior is akin to the models prescribed by Akamatsu [7] and Baillon and Cominetti [26], but with routing occurring over the CoDAG associated with original network. We show that the corresponding equilibrium congestion pattern—which we term the *Condensed DAG equilibrium* (CoDAG equilibrium)—can be characterized as the unique minimizer of a strictly convex optimization problem.

Moreover, we propose a discrete-time dynamical system that captures a natural adaptation rule used by self-interested travelers who progressively learn towards equilibrium arc selections. In the game theory literature, an equilibrium notion is only considered useful if there exists an adaptive learning scheme that allows self interested players to converge to it [149]. Despite research progress on both theoretical and algorithmic aspects of stochastic arc-based TAMs, to the best of our knowledge, there has been no research on adaptive learning schemes that ensure convergence to such equilibria. Recently, adaptive learning schemes that converge to equilibria in route-based TAMs have been extensively studied [217, 216, 213, 270, 362, 295, 158], by considering self-interested travelers who repeatedly select routes by observing route latencies in past rounds of interaction. In this work, we extend this line of research to arc-based TAMs by proposing a discrete-time dynamics, in which in every round, travelers update arc selections at every node on the CoDAG network based on previous interactions. We prove that the emergent aggregate arc selection probabilities at every node (and the resulting congestion levels on each arc) globally asymptotically converge to a neighborhood of the CoDAG equilibrium.

To establish convergence, we appeal to the theory of stochastic approximation [57], which requires two conditions: (i) The vector field of the discrete-time dynamical system is Lipschitz, and (ii) The trajectories of an associated continuous-time dynamical system asymp-

totically converge to the CoDAG equilibrium. To prove (i), we establish recursive Lipschitz bounds for vector fields associated with every node. For (ii), we first construct a Lyapunov function using a strictly convex optimization objective associated with the CoDAG representation. We then show that the value of this Lyapunov function decreases along the trajectories of the continuous-time dynamical system. Our contributions are:

1. We introduce a new arc-based traffic equilibrium concept—the *Condensed DAG equilibrium*—which overcomes some limitations of existing traffic equilibrium notions. Furthermore, we show that the Condensed DAG equilibrium is characterized by a solution to a strictly convex optimization problem.

2. We present, to the best of our knowledge, the first adaptive learning scheme in the context of stochastic arc-based TAM. Furthermore, we establish formal convergence guarantees for this learning scheme.

3. We validate our theorems on a simulated traffic network.

The chapter unfolds as follows. Section 8.1 introduces the setup considered in this chapter, and defines the Condensed DAG representation. Section 8.2 defines the Condensed DAG equilibrium, and characterize it as a solution to a strictly convex optimization problem. Section 8.3 presents discrete-time dynamics that converges to the Condensed DAG equilibrium and also provides a proof sketch. In Section 8.4, we numerically study the convergence of the discrete-time dynamics on a simulated traffic network. Finally, Section 8.5 presents concluding remarks and future work directions.

Notation For each positive integer $n \in \mathbb{N}$, we denote $[n] := \{1, \cdots, n\}$. For each $i \in [n]$ in an Euclidean space $\mathbb{R}^n$, we denote by e_i the i-th standard unit vector.

8.1 Condensed DAG Representation

Setup

Consider a traffic network represented by a directed graph $G_O = (I_O, A_O)$, possibly with bidirectional arcs, where I_O and A_O denote nodes and arcs, respectively. An example is depicted in Figure 8.1 (top left). Let the *origin nodes* and *destination nodes* be two disjoint subsets of nodes in G_O. Each traveler enters the network through an origin node to travel to a destination node, by sequentially selecting arcs at every intermediate node. This gives rise to congestion on each arc, which in turn decides the travel times. Specifically, each arc $\tilde{a} \in A_O$ is associated with a strictly increasing *latency function* $\ell_{\tilde{a}} : [0, \infty) \to [0, \infty)$, which gives for each arc the travel time as a function of traffic flow. To simplify our exposition, we assume that there is only one origin-destination tuple (o, d), although the results presented

in this chapter naturally extend to settings where the traffic network has multiple origin-destination pairs. We denote by g_o the demand of (infinitesimal) travelers who travel from the origin o to the destination d.

Remark 8.1.1. *Arc selections made by travelers at different nodes are independent of one another. Therefore, if the underlying network has bidirectional edges, then sequential arc selection by a traveler can result in a cyclic route. For example, sequential arc selection in the original network shown on the top left in Figure 8.1 may lead a traveler to loop between i_2^O and i_3^O before reaching destination. To overcome this, we introduce a directed acyclic graph (DAG) representation of the original graph G_O in the following subsections, called the condensed DAG. Sequential arc selections made on this network encodes the travel history by design and therefore avoids cyclic routes.*

Preliminaries on DAG: Depth and Height

Before introducing condensed DAG representation, we first present the notions of *height* and *depth* of a DAG. These concepts are crucial for the construction and analysis of condensed DAGs in the following sections. For the exposition in this subsection, let G be a DAG with a single origin-destination pair (o, d). Furthermore, let $\mathbf{R}$ be the set of all acyclic routes in G which start at the origin node o and end at the destination node d.

Definition 8.1.1 (**Depth**). *For each $r \in \boldsymbol{R}$ and $a \in r$, let $\ell_{a,r}$ denote the location of arc a in route r, i.e., a is the $\ell_{a,r}$-th arc in the route r, and with a slight abuse of notation, define: $\ell_a := \max_{r \in \boldsymbol{R}: a \in r} \ell_{a,r}$, We say that a is an ℓ_a-*th depth arc *in the Condensed DAG G. Moreover, we define the* depth of a node $i \in I \backslash \{o\}$ *by:*

$$\bar{\ell}_i := \max_{a \in A_i^-} \ell_a$$

with $\bar{\ell}_o = 0$.

Definition 8.1.2 (**Height**). *For each $r \in \boldsymbol{R}$ and $a \in r$, let $m_{a,r}$ denote the location of arc a in route r when enumerating arcs in r backwards from the destination node, i.e., a is the $(|r| - m_{a,r})$-th arc in route r, and with a slight abuse of notation, define: $m_a := \max_{r \in \boldsymbol{R}: a \in r} m_{a,r}$. We say that a is an m_a-*th height arc *in the Condensed DAG G . Moreover, we define the* height of a node $i \in I \backslash \{d\}$ *by:*

$$\bar{m}_i := \max_{a \in A_i^+} m_a$$

with $\bar{m}_d = 0$.

Table 8.1.1: Arc correspondences between the graphs in Figure 8.1: The original network (top left), fully expanded tree (bottom), and the CoDAG (top right).

Original	Tree DAG	CoDAG
a_1^O	$a_1^T, a_2^T, a_3^T, a_4^T, a_5^T$	a_1^C
a_2^O	$a_6^T, a_7^T, a_8^T, a_9^T, a_{10}^T$	a_2^C
a_3^O	a_{12}^T, a_{13}^T	a_4^C
a_4^O	$a_{18}^T, a_{19}^T, a_{20}^T$	a_7^C
a_5^O	$a_{14}^T, a_{15}^T, a_{23}^T, a_{24}^T$	a_5^C, a_9^C
a_6^O	$a_{16}^T, a_{17}^T, a_{21}^T, a_{22}^T$	a_6^C, a_8^C
a_7^O	a_{11}^T, a_{25}^T	a_3^C, a_{10}^C
a_8^O	$a_{26}^T, a_{28}^T, a_{30}^T, a_{32}^T$	a_{11}^C
a_9^O	$a_{27}^T, a_{29}^T, a_{31}^T, a_{33}^T$	a_{12}^C

Construction of Condensed DAG

For ease of description, we illustrate the construction through an example in Figure 8.1. We also present a pseudo-code to generate the condensed DAG representation.

A straightforward way to associate G_O with a DAG would be to brute-force enumerate all acyclic (simple) routes and construct a tree network by replicating arcs and nodes by the number of routes passing through them (see Figure 8.1, bottom). However, the resulting tree network may contain a significantly larger number of arcs and nodes compared with the original network. To ameliorate this, we present the *condensed DAG* representation (Figure 8.1, top right). The condensed DAG is formed by merging superfluous nodes and arcs in the tree network, while ensuring that the graph remains acyclic, and preserving the set of acyclic routes from the original network.

One can design a condensed DAG representation as follows:

(S1) Convert the original network G_O to a tree structure $G_T = (I_T, A_T)$, in which every branch emanating from the origin represents a route. Each node and arc is replicated by the number of acyclic routes that contains it. For every node i in G_T, compute the depth $\bar{\ell}_i$ and height $\bar{m}_i$ (see Definition 8.1.1-8.1.2).

(S2) Generate a partition P_T of I_T such that:

(i) For each $X \in P_T$, all nodes in X replicate the node in I_O that shares the same height or depth in G_T.

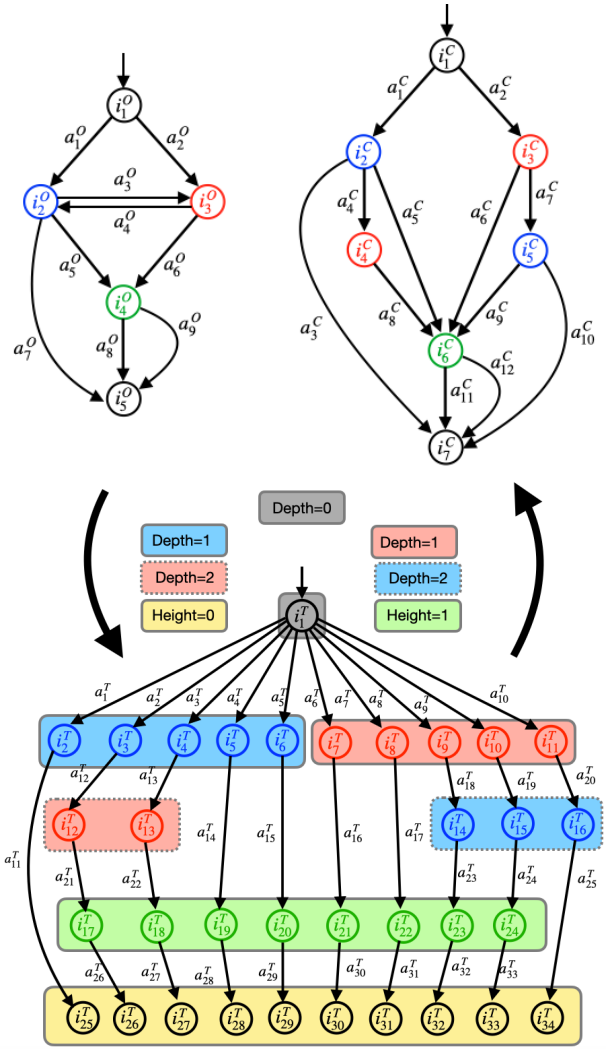


Figure 8.1: Example of a single-origin single-destination original network G_O (top left, with superscript O), and its corresponding tree network (bottom, with superscript T) and condensed DAG G (top right, with superscript C). The blocks in G_T represent a partition P_T (see (S2)). The depth and height of nodes in every partition are denoted above G_T. Arc correspondences between the three networks are given by Table 8.1.1, while node correspondences are indicated by color.

(ii) For any $X, Y \in P_T$, there exists no $i, i' \in X, j, j' \in Y$, such that $\bar{m}_j > \bar{m}_i$ and $\bar{m}_{j'} < \bar{m}_{i'}$

(S3) For each set element X of P_T, merge all nodes in X into a single node. Then, merge arcs which have the same start and end nodes, and are replicas of the same edge in the original network G_O.

We refer to any graph generated via (S1)-(S3) as a *condensed DAG* (CoDAG) representation $G = (I, A)$ of the original network, where I and A are the set of nodes and arcs, respectively. By construction, the CoDAG representation explicitly avoids cyclic routes, and preserves all the acyclic routes from the original network. This is because the tree network constructed in (S1) preserves all acyclic routes from original network. Furthermore, the merging conditions stated in (S3) prohibit both the removal and the addition of routes.

Remark 8.1.2. *A given traffic network with bidirectional arcs may yield several distinct CoDAG representations, any of which would be amenable to our analysis in subsequent sections. The development of an algorithmic procedure to compute a CoDAG with the least number of arcs or nodes is beyond the scope of this work.*

Remark 8.1.3. *The Condensed DAG representation G can be significantly smaller in size, compared to the tree network. There exist original networks whose corresponding tree representation G_T is exponentially larger than its corresponding CoDAG G. For example, consider a network with nodes $i_1, \cdots, i_n$, with two directed arcs connecting i_k to i_{k+1}, for each $k \in [n-1]$. Here, the corresponding tree network would have $2^n - 2$ arcs, while the CoDAG representation only has $2(n-1)$ arcs.*

Remark 8.1.4. *The arc-based TAM literature also considers modified representations of traffic networks with bidirectional arcs. For example, Oyama, Hara et al. [324, 330] consider the Network Generalized Extreme Value (NGEV) representations, which are similar to our CoDAG representation, but applies only to acyclic networks [324]. Thus, NGEV models cannot capture realistic traffic networks where almost all arcs are bidirectional. Meanwhile, Oyama, Hato et al. [326] consider the Choice Based Prism (CBP) representation, which prunes the available set of feasible routes to ameliorate computational inefficiency. While CBP explicitly avoids cyclic routes, it also removes some acyclic routes during the pruning process. In contrast, the CoDAG representation avoids this issue.*

To conclude this section, we introduce some notation used throughout the rest of the chapter. Recall that CoDAGs are formed by replicating the arcs in G_O. To describe this correspondence between arcs, we define $[\cdot] : A \to A_O$ to be a map from each CoDAG arc $a \in A$ to the corresponding arc $[a] \in A_O$. For each arc $a \in A$, let i_a and j_a denote the start and terminal nodes, and for each node $i \in I$, let $A_i^-, A_i^+ \subset A$ denote the set of incoming and outgoing arcs.

8.2 Equilibrium Characterization

In this section, we introduce the *condensed DAG (CoDAG) equilibrium* (Definition 8.2.1), which is based on the CoDAG representation of the original traffic network. Specifically, we show that the CoDAG equilibrium exists, is unique, and solves a strictly convex optimization problem (Theorem 8.2.1).

Condensed DAG Equilibrium

Below, we assume that every traveler knows G_O and has access to the same CoDAG representation of G_O. To avoid cyclic routes, we model travelers as performing sequential arc selection over the CoDAG representation $G = (I, A)$. The aggregate effect of the travelers' arc selections gives rise to the congestion on the network. Concretely, for each $a \in A$, let the *flow* or *congestion level* on arc a be denoted by w_a, and let the total flow on the corresponding arc in the original network be denoted, with a slight abuse of notation, by $w_{[a]} := \sum_{a' \in [a]} w_{a'}$. (Note that unlike existing TAMs, the latency of arcs in G can be coupled through the map $w_{[\cdot]}$, since multiple copies of the same arc in G_O may exist in G.) Then, the perceived latency of travelers on each arc $a \in A$ is described by:

$$\tilde{s}_{[a]}(w_{[a]}) := s_{[a]}(w_{[a]}) + \nu_a,$$

where ν_a is a zero-mean random variable. At each non-destination node $i \in I \backslash \{d\}$, travelers select among outgoing nodes $a \in A_i^+$ by comparing their perceived latencies-to-go $\tilde{z}_a : \mathbb{R}^{|A|} \to \mathbb{R}$, given recursively by:

$$\begin{aligned} \tilde{z}_a(w) &:= \tilde{s}_{[a]}(w_{[a]}) + \min_{a' \in A_{j_a}^+} \tilde{z}_{a'}(w), \quad && j_a \neq d, \\ \tilde{z}_a(w) &:= \tilde{s}_{[a]}(w_{[a]}), && j_a = d. \end{aligned} \tag{8.1}$$

Consequently, the fraction of travelers who arrive at $i \in I \backslash \{d\}$ and choose arc $a \in A_i^+$ is given by:

$$P_{ij_a} := \mathbb{P}(\tilde{z}_a \leqslant \tilde{z}_{a'}, \forall\, a' \in A_i^+). \tag{8.2}$$

An explicit formula for the probabilities $\{P_{ij_a} : a \in A_i^+\}$ in terms of the statistics of $\tilde{z}_a$, is provided by the discrete-choice theory [40]. In particular, define $z_a(w) := \mathrm{E}\left([]\, \tilde{z}_a(w)\right]$ and $\epsilon_a := \tilde{z}_a(w) - z_a(w)$, and define the latency-to-go at each node by:

$$\varphi_i(\{z_{a'}(w) : a' \in A_i^+\}) = \mathbb{E}\left[\min_{a' \in A_i^+} \tilde{z}_{a'}(w)\right]. \tag{8.3}$$

Then, from discrete-choice theory [40]:

$$P_{ij_a} = \frac{\partial \varphi_i(z)}{\partial z_a}, \quad i \in I \backslash \{d\}, a \in A_i^+, \tag{8.4}$$

where, with a slight abuse of notation, we write $\varphi_i(z)$ for $\varphi_i(\{z_{a'} : a' \in A_i^+\})$. To obtain a closed form expression of φ, this work considers the *logit Markovian model* [7, 26], which assumes that the zero-mean noise ϵ is Gumbel-distributed with scale $\beta > 0$. Intuitively, $\beta > 0$ is an entropy parameter that models the degree to which the average traveler's perception of network latency is suboptimal. In this case, the corresponding latency-to-go at each node i in G is:

$$\varphi_i(z) = -\frac{1}{\beta} \ln \left(\sum_{a' \in A_i^+} e^{-\beta z_{a'}} \right). \tag{8.5}$$

Using (8.1) and (8.5), the expected minimum latency-to-go $z_a : \mathbb{R}^{|A|} \to \mathbb{R}$, associated with traveling on each arc $a \in A$, is given by:

$$z_a(w) = \ell_{[a]} \left(\sum_{\bar{a} \in [a]} w_{\bar{a}} \right) - \frac{1}{\beta} \ln \left(\sum_{a' \in A_{ja}^+} e^{-\beta z_{a'}(w)} \right). \tag{8.6}$$

Note that (8.6) is well-posed, as z_a can be recursively computed along arcs of increasing height (Definition 8.1.2) from the destination back to the origin. For more details, please see Appendix G.2 [278].

Against the preceding backdrop, we formally define the central equilibrium solution concept studied in this chapter: the Condensed DAG Equilibrium (CoDAG Equilibrium).

Definition 8.2.1 (**Condensed DAG Equilibrium**). *Given $\beta > 0$, a vector of arc-flow $\bar{w}^\beta \in \mathbb{R}^{|A|}$ is called a* Condensed DAG equilibrium *if, for each $i \in I \backslash \{d\}$, $a \in A_i^+$:*

$$\bar{w}_a^\beta = \left(g_i + \sum_{a' \in A_i^+} \bar{w}_{a'}^\beta \right) \cdot \frac{\exp(-\beta z_a(\bar{w}^\beta))}{\sum_{a' \in A_{i_a}^+} \exp(-\beta z_{a'}(\bar{w}^\beta))}, \tag{8.7}$$

where $g_i = g_o$ if $i = o$, $g_i = 0$ otherwise, and $w \in \mathcal{W}$, with:

$$\mathcal{W} := \Bigg\{ \bar{w}^\beta \in \mathbb{R}^{|A|} : \sum_{a \in A_i^+} \bar{w}_a^\beta = \sum_{a \in A_i^-} \bar{w}_a^\beta, \forall\, i \neq o, d, \tag{8.8}$$

$$\sum_{a \in A_o^+} \bar{w}_a^\beta = g_o,\ \bar{w}_a^\beta \geqslant 0, \forall\, a \in A \Bigg\}.$$

For any CoDAG equilibrium $\bar{w}^\beta$, the fraction of travelers at any node $i \in I \backslash \{d\}$ who selects an arc $a \in A_i^+$ is:

$$\bar{\xi}_a^\beta := \frac{\bar{w}_a^\beta}{\sum_{a' \in A_i^+} \bar{w}_{a'}^\beta}.$$

Remark 8.2.1. *Essentially, at the CoDAG equilibrium, the traveler population at each intermediate node $i \in I\backslash\{d\}$ (with total flow $g_i + \sum_{a' \in A} w_{a'}$) select from outgoing arcs by comparing their costs-to-go using the softmax function. While the CoDAG equilibrium and Markovian Traffic Equilibrium (MTE) share some similarities (see [26]), there also exist two main fundamental differences. First, by design, the CoDAG equilibrium does not yield cyclic routes with strictly positive probability (as is the case with the MTE). Second, unlike the MTE, congestion levels on arcs (which may be replicas of the same arc in G_O) in the CoDAG representation are coupled to each other. Therefore, MTE analysis does not extend straightforwardly to the CoDAG equilibrium.*

Existence and Uniqueness of the CoDAG equilibrium

In this subsection, we show the existence and uniqueness of the CoDAG equilibrium, by characterizing it as the unique minimizer of a strictly convex optimization problem over a compact set. First, for each $[a] \in A_O$, define:

$$f_{[a]}(w) := \int_0^{w_{[a]}} \ell_{[a]}(u)\, du, \tag{8.9}$$

and for each $i \in I\backslash\{d\}$, set:

$$\chi_i(w_{A_i^+}) := \sum_{a \in A_i^+} w_a \ln w_a - \left(\sum_{a \in A_i^+} w_a \right) \ln \left(\sum_{a \in A_i^+} w_a \right). \tag{8.10}$$

Finally, define $F : \mathcal{W} \to \mathbb{R}$ by:

$$F(w) = \sum_{[a] \in A_O} f_{[a]}(w) + \frac{1}{\beta} \sum_{i \neq d} \chi_i(w_{A_i^+}), \tag{8.11}$$

where $w_{A_i^+} \in \mathbb{R}^{|A_i^+|}$ denotes the components of w corresponding to arcs in A_i^+.

Theorem 8.2.1. *The CoDAG equilibrium $\bar{w}^\beta \in \mathcal{W}$ exists, is unique, and is the unique minimizer of F over $\mathcal{W}$.*

To prove Theorem 8.2.1, we first show that $F(\cdot)$ is strictly convex over $\mathcal{W}$ (Lemma 8.2.1), so F has a unique minimizer in $\mathcal{W}$. It then suffices to show that the CoDAG equilibrium definition (Definition 8.2.1) matches the Karush-Kuhn-Tucker (KKT) conditions for the optimization problem (8.11).

Lemma 8.2.1. *The map $F : \mathcal{W} \to \mathbb{R}$ is strictly convex.*

Proof. (**Proof Sketch**) It suffices to show that $f_{[a]}$ and χ_i are convex for each $[a] \in A_O$, $i \in I\backslash\{d\}$. Each $f_{[a]}$ is convex, since it is the composition of a convex function ($w \mapsto$

$\sum_{a\in A_O}\int_0^{w_a} s_a(u)du)$ with a linear function ($w_{[a]} := \sum_{a'\in[a]} w_{a'}$). Furthermore, we establish that for any $i \in I\backslash\{d\}$, $y_i \in \mathbb{R}^{|A_i^+|}$:

$$y_i^\top \nabla_w^2 \chi_i(w) y_i \geqslant 0,$$

where the equality holds if and only if y_i and $w_{A_i^+}$ are scalar multiples of one another. Strict convexity then follows by a contradiction argument showing that there exists at least one node $i \in I\backslash\{d\}$ such that $y_i^\top \nabla_w^2 \chi_i(w) y_i > 0$. □

8.3 Learning Dynamics

In this section, we propose a discrete-time dynamical system (PBR) which captures travelers' preferences for minimizing total travel time, as well as their perception uncertainties, while simultaneously learning about the emergent congestion on the network.

We leverage the constant step-size stochastic approximation theory to prove that these discrete-time dynamics converge to a neighborhood of the CoDAG equilibrium (Theorem 8.3.1). To this end, we first prove that the continuous-time counterpart to (PBR) globally asymptotically converges to the CoDAG equilibrium (Lemma 8.3.1). We then conclude the proof by verifying technical assumptions required to invoke results in stochastic approximation theory [57] (Lemma 8.3.2).

Discrete-time Dynamics

In this subsection, we present a discrete-time dynamical equation that captures the evolution of flows on the network as a result of learning and adaptation by self-interested travelers. More formally, at each discrete time step $n \geqslant 0$, g_o units of travelers arrive at the origin node o. At time step n, every traveler who reaches node $i \in I\backslash\{d\}$ selects some arc $a \in A_i^+$. For any $i \in I\backslash\{d\}, a \in A_i^+$, let $\xi_a[n]$ be the *aggregate arc selection probability*: the fraction of travelers at node i choosing arc a at time n. As a result of the arc selections made by every traveler, a flow of $W[n]$ is induced on the arcs as given below. For every $a \in A$:

$$W_a[n] = \left(g_{i_a} + \sum_{a'\in A_{i_a}^-} W_{a'}[n]\right) \cdot \xi_a[n], \tag{8.12}$$

where, as given in Definition 8.2.1, $g_{i_a} = g_o$ if $i_a = o$, and $g_{i_a} = 0$ otherwise.

At the end of each time step, every traveler reaches the destination and observes a noisy estimate of the latency-to-go, independent across travelers, on every arc in the network (including ones they did not visit in that time step). Note that the latency-to-go for any arc is dependent on the congestion $W[n]$, which in turn depends on aggregate decisions taken by travelers (please refer to (8.12)). Based on the observed latencies, at time $n+1$, at every non-destination node $i \in I\backslash\{d\}$, a $\eta_i[n+1] \cdot K_i$ fraction of travelers at node i switches to

an arc with the minimum observed latency-to-go. Meanwhile, a $1-\eta_i[n+1]\cdot K_i$ fraction of travelers selects the same arc they selected at time step n. We assume that $\{\eta_i[n+1] \in \mathbb{R} : i \in I, n \geqslant 0\}$ are independent bounded random variables in $[\underline{\mu}, \overline{\mu}]$, independent of travelers' perception stochasticities, with $0 < \underline{\mu} < \mu < \overline{\mu} < 1/\max\{K_i : i \in I\backslash\{d\}\}$ and $\mathrm{E}\left(\left[\right)\eta_{i_a}[n+1]\right] = \mu$ for each node index $i \in I$ and discrete time index $n \geqslant 0$. Meanwhile, the constants K_i represent node-dependent update rates. To summarize, the dynamic evolution of arc selections by infinitesimal travelers is captured by the following evolution of $\xi[n]$. For every $i \in I\backslash\{d\}, a \in A_i^+$:

$$\xi_a[n+1] = \xi_a[n] + \eta_{i_a}[n+1]\cdot K_{i_a}\left(-\xi_a[n] + P_{ij_a}\right),$$

where P_{ij_a} is defined in (8.2). Using (8.4) and (8.5), the previous equation can be rewritten as:

$$\begin{aligned}&\xi_a[n+1] \\ =&\xi_a[n] + \eta_{i_a}[n+1]\cdot K_{i_a}\cdot\left(-\xi_a[n] + \frac{\exp(-\beta\left[z_a(W[n])\right])}{\sum_{a'\in A_{i_a}^+}\exp(-\beta\left[z_{a'}(W[n])\right])}\right),\end{aligned} \tag{PBR}$$

The dynamics (PBR) bears close resemblance to perturbed best response dynamics in routing games [362], so we shall refer to (PBR) as *perturbed best response* dynamics.

We assume $\xi_a[0] > 0$ for each $a \in A$, i.e., each arc has some strictly positive initial traffic flow. This is reasonable, since the stochasticity in travelers' perception of network congestion ensures that each arc has a nonzero probability of being selected.

Convergence Results

Our main theorem establishes that the discrete-time dynamics (PBR) asymptotically converges to a neighborhood of the CoDAG equilibrium $\bar{w}^\beta$.

Theorem 8.3.1. *Under the discrete-time flow dynamics* (PBR)*, for each* $\delta > 0$*:*

$$\limsup_{n\to\infty}\mathbb{E}\left[\|\xi[n]-\bar{\xi}^\beta\|_2^2\right] \leqslant O(\mu),$$
$$\limsup_{n\to\infty}\mathbb{P}\left(\|\xi[n]-\bar{\xi}^\beta\|_2^2 \geqslant \delta\right) \leqslant O\left(\frac{\mu}{\delta}\right).$$

To prove Theorem 8.3.1, we leverage the theory of constant step-size stochastic approximation [57]. This requires proving that the continuous-time dynamics corresponding to the discrete-time update (PBR), presented below, converges to the CoDAG equilibrium. For each arc $a \in A$:

$$\dot{\xi}_a(t) = -K_i\left(\xi_a(t) + \frac{\exp(-\beta\cdot z_a(w(t)))}{\sum_{a'\in A_{i_a}^+}\exp(-\beta\cdot z_{a'}(w(t)))}\right), \tag{8.13}$$

where $w(t)$ is the resulting arc flow associated with the arc selection probability $\xi(t)$, similar to (8.12):

$$w_a(t) = \xi_a(t) \cdot \left(g_{i_a} + \sum_{a' \in A_{i_a}^-} w_{a'}(t) \right). \tag{8.14}$$

Lemma 8.3.1 (**Informal**). *Suppose $w(0) \in \mathcal{W}$, i.e., the initial flow satisfies flow continuity. Under the continuous-time flow dynamics* (8.14) *and* (8.13)*, if $K_i \ll K_{i'}$ whenever $\ell_i < \ell_{i'}$, the traffic flow $w(t)$ globally asymptotically converges to the CoDAG equilibrium $\bar{w}^\beta$.*

Proof. (**Proof Sketch**) Recall that $\bar{w}^\beta$ is the unique minimizer of the map $F : \mathcal{W} \to \mathbb{R}$, defined by (8.11). We show that F is a Lyapunov function for the continuous-time flow dynamics (G.6) induced by the arc selection dynamics (8.13). To this end, we first unwind the dynamics (8.13) and (8.14) to obtain the recursive relation:

$$\begin{aligned} \dot{w}_a(t) = -K_{i_a} \left(1 - \frac{1}{K_{i_a}} \cdot \frac{\sum_{a' \in A_{i_a}^-} \dot{w}_{a'}(t)}{\sum_{\hat{a} \in A_{i_a}^+} w_{\hat{a}}(t)} \right) w_a(t) \\ + K_{i_a} \cdot \sum_{a' \in A_{i_a}^-} w_{a'}(t) \cdot \frac{\exp(-\beta z_a(w(t)))}{\sum_{a' \in A_{i_a}^+} \exp(-\beta z_{a'}(w(t)))}. \end{aligned}$$

Then, we establish that along any trajectory starting on $\mathcal{W}$ and following the dynamics given by (8.13), we have for each $t \geqslant 0$:

$$\dot{F}(t) = \dot{w}(t)^\top \nabla_w F(w(t)) \leqslant 0.$$

The proof then follows from LaSalle's Theorem (see [365, Proposition 5.22]). For a precise characterization and detailed proof of Lemma 8.3.1, please see Appendix G.3 [278]. □

Remark 8.3.1. *On a technical level, the statement and proof technique of Theorem 8.3.1 share similarities with methods used to establish the convergence of best-response dynamics in potential games [362]. However, there exist crucial distinctions between the two approaches which render our problem more difficult. First, since the map F defined by* (8.11) *is not a potential function, the mathematical machinery of potential games cannot be directly applied. Moreover, the continuous-time flow dynamics* (8.13) *and* (8.14) *allow couplings between arbitrary arcs in the CoDAG. For more details, please see Appendix G.3 [278].*

Remark 8.3.2. *The assumption that $K_i \ll K_{i'}$ whenever the depth of node $i \in I \backslash \{d\}$ is less than the depth of node $i' \in I \backslash \{d\}$ conforms to the intuition that travelers farther away from the destination node face more complex route selection decisions based on more information regarding traffic flow throughout the rest of the network, and thus perform slower updates.*

Having established the global asymptotic convergence of the continuous-time dynamics (8.13) and (8.14) to the CoDAG equilibrium $\bar{w}^\beta$, it remains to verify the remaining technical conditions necessary to prove Theorem 8.3.1 via stochastic approximation theory. To

this end, we rewrite the discrete ξ-dynamics (PBR) as a Markov process with a martingale difference term:

$$\xi_a[n+1] = \xi_a[n] + \mu\Big(\rho_a(\xi[n]) + M_a[n+1]\Big),$$

where $\rho_a : \mathbb{R}^{|A|} \times \mathbb{R}^{|A_O|} \to \mathbb{R}^{|A|}$ is given by:

$$\rho_a(\xi) := K_{i_a}\left(-\xi_a + \frac{\exp(-\beta \cdot z_a(w))}{\sum_{a' \in A^+_{i_a}} \exp(-\beta \cdot z_{a'}(w))}\right), \tag{8.15}$$

with $w \in \mathbb{R}^{|A|}$ defined arc-wise by $w_a = (g_{i_a} + \sum_{\hat{a} \in A^-_{i_a}} w_{a'}) \cdot \xi_a$, and:

$$M_a[n+1] := \left(\frac{1}{\mu}\eta_{i_a}[n+1] - 1\right) \cdot \rho_a(\xi[n]). \tag{8.16}$$

Here, $W_a[n] = \Big(g_{i_a} + \sum_{a' \in A^-_{i_a}} W_{a'}[n]\Big)$, as given by (8.12).

The following lemma bounds the magnitude of the discrete-time flow $W[n] \in \mathbb{R}^{|A|}$ and the martingale difference terms $M[n] \in \mathbb{R}^{|A|}$.

Lemma 8.3.2. *Given initial flows $W[0]$ and arc selection probabilities $\xi[0]$:*

1. *For each $a \in A$: $\{M_a[n+1] : n \geqslant 0\}$ is a martingale difference sequence with respect to the filtration $\mathcal{F}_n := \sigma\Big(\cup_{a \in A}\,(W_a[1], \xi[1], \cdots, W_a[n], \xi[n])\Big)$.*

2. *There exist $C_w, C_m > 0$ such that, for each $a \in A$, $n \geqslant 0$, we have $W_a[n] \in [C_w, g_o]$ and $|M_a[n]| \leqslant C_m$.*

3. *For each $a \in A$, the map ρ_a, given by (8.15), is Lipschitz continuous over the range of realizable flow and arc selection probability trajectories $\{W[n] : n \geqslant 0\}$ and $\{\xi[n] : n \geqslant 0\}$.*

Proof. (**Proof Sketch**) The first part of Lemma 8.3.2 follows because, with respect to $\mathcal{F}_n$, the only stochasticity in $M_a[n+1]$ originates from the i.i.d. input flows $\eta_{i_a}[n+1]$. The second part follows by invoking the flow continuity equations in (8.12) to recursively upper bound each $W_a[n]$ and $z_a(W[n])$, in increasing order of depth and height, respectively (flows are propagated from origin to destination, and latency-to-go values are computed in the opposite direction). These bounds are then used to recursively establish upper and lower bounds for each $\xi_a[n]$, and consequently each $W[n]$, in order of increasing depth. Finally, the Lipschitz continuity of each ρ_a can be proved by establishing that ρ_a is continuously differentiable, with bounded derivatives over the compact domain defined by the bounds on $W[n]$ established in the second part of the lemma. For details, please see the proofs of Lemmas G.3.2 and G.3.3 in Appendix G.3 [278]. □

Table 8.4.1: Parameters for simulation.

Notation	Default value
k_0	0, 1, 0, 1, 1, 0, 1, 1, 1 (ordered by edge index)
k_1	2, 1, 1, 1, 1, 1, 2, 2, 2 (ordered by edge index)
g_1	1
β	10
$\eta_{i_a}[n]$	Uniform$(0, 0.1)$, $\forall a \in A, i \in I \backslash \{d\}$

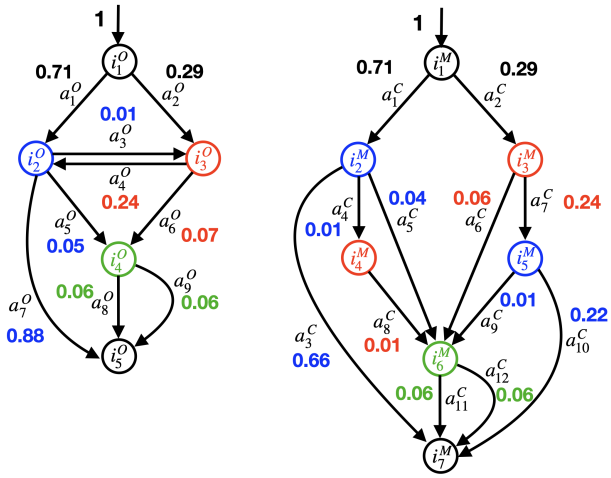


Figure 8.2: Steady state traffic flow on each arc for an original network and condensed DAG. Flows on arcs emerging from same node are represented in same color.

8.4 Numerical Results

In this section, we conduct numerical experiments to validate the theoretical analysis presented in Section 8.3. We show in simulation that, under (PBR), the traffic flows converge to a neighborhood of the condensed DAG equilibrium, as claimed by Theorem 8.3.1.

Consider the network presented in Figure 8.1, with affine edge-latency functions

$$\ell_{[a]}(w_{[a]}) = k_0 + k_1 w_{[a]}$$

for each arc $a \in A$, where $k_0, k_1 > 0$ are simulation parameters provided in Table 8.4.1. To validate Theorem 8.3.1, we evaluate and plot the traffic flow values $W_a[n]$ on each arc $a \in A$ and discrete time $n \geqslant 0$. Figure 8.2 presents traffic flow values at the condensed DAG equilibrium (i.e., w^β) for the original network and condensed DAG. While travelers generally prefer routes of lower latency, each route has a nonzero level of traffic flow at equilibrium. The reason is that under the perturbed best response dynamics, users do not allocate all

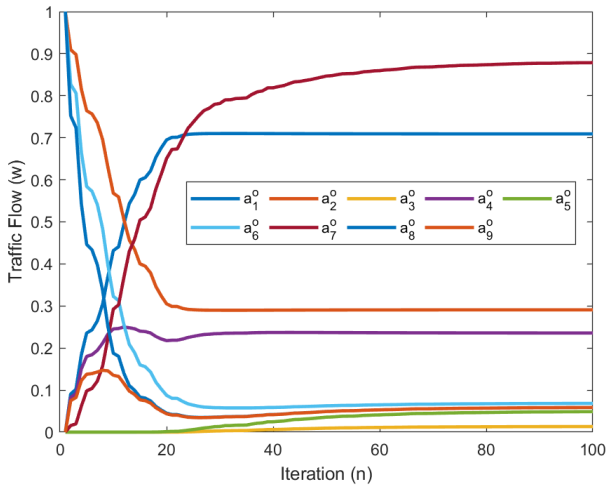


Figure 8.3: Traffic flow $W[n]$ for the network in Fig. 8.2.

the traffic flow to the minimum-cost route, but instead distribute their traffic allocation more evenly. Meanwhile, Figure 8.3 illustrates that w converges to the condensed DAG equilibrium in approximately 100 iterations with some initial fluctuations. The fluctuations are due to the magnitude of the average step-size μ. If μ is small, the discrete-time update is close to the continuous-time dynamics, and the resulting evolution of the traffic flow follows a smoother trend. Note that in practice, flow convergence to the CoDAG equilibrium occurs even when the effects of the constants $\{K_i : i \in I\}$ are ignored, i.e., when each K_i is set to unity.

8.5 Concluding Remarks

We introduce a novel equilibrium concept for stochastic arc-based traffic assignment models (TAMs) that ensures all travelers are routed along acyclic paths. This is achieved by constructing a condensed directed acyclic graph (DAG) representation of the original network through systematic replication of arcs and nodes, which eliminates cyclic routes while preserving the original feasible route set. We rigorously characterize this equilibrium as the unique optimal solution to a strictly convex optimization problem. Building on this formulation, we propose adaptive learning dynamics that model the evolution of traffic flow resulting from the simultaneous learning and strategic adaptation of self-interested travelers. Finally, we establish convergence guarantees, proving that the proposed learning dynamics asymptotically converge to the equilibrium flow allocation.

Part III

Data-driven Mechanisms for Societal Good

Chapter 9

Efficiency and Equity Considerations in Transportation through Data-driven Congestion Pricing

Congestion pricing is an incentivize mechanism for effective utilization of road infrastructure among selfish travelers. Widely adopted in many major cities, both theoretical [435] and empirical [107, 342, 335, 130] studies have shown that congestion pricing can reduce traffic congestion and greenhouse gas emissions, and improve air quality [254, 234, 446, 172]. The revenue generated from congestion pricing is often reinvested to improve the road infrastructure, public transit, and other sustainable mobility initiatives [160, 387]. Despite these benefits, implementation of congestion pricing often faces challenges, and one of the primary concerns is its disproportional impact on low-income travelers [118, 154]. These travelers often have limited access to alternative transportation options, and the additional financial burden of congestion fees may exacerbate existing inequalities.

In this work we present a principled approach to compute congestion pricing schemes that incorporate both *(i)* the *efficiency* objective of minimizing the total travel time on the network, and *(ii)* the *equity-welfare* objective, where the equity is assessed in terms of maximum disparity in relative change in travel costs experienced by different traveler populations following the implementation of tolls, compared to a scenario with no tolls, and welfare is assessed as the average relative change in travel costs experienced by travelers across all types following the implementation tolls, compared to scenario with no tolls.

We consider a non-atomic routing game, where travelers make routing decisions based on the travel time of each route plus the monetary cost that includes tolls and gas prices. The monetary cost is adjusted by the travelers' value-of-time— the amount of money a traveler is willing to pay to save a unit of time. Our game has a finite number of traveler populations, each with a heterogeneous value-of-time. Following the result from [169], the equilibrium flow in our game is unique, and can be computed by solving a convex optimization problem. Moreover, the congestion minimizing edge flow vector (i.e. the edge flow vector that minimizes the total travel time) is unique.

We propose four kinds of congestion pricing schemes that differ in terms of whether (a) tolls are differentiated based on the type of population, and (b) tolls can be set on all edges or a subset of edges. In particular, the four congestion pricing schemes are: (i) *homogeneous pricing scheme with no support constraints*, denoted by hom, where all populations are charged with the same tolls and all edges are allowed to be tolled; (ii) *heterogeneous pricing scheme with no support constraints*, denoted by het, where populations are charged with differentiated toll prices based on their types and all edges can be tolled; (iii) *homogeneous pricing scheme with support constraints*, denoted by hom_sc, where tolls are not differentiated but only a subset of edges can be tolled; (iv) *heterogeneous pricing scheme with support constraints*, denoted by het_sc, where tolls are differentiated and only a subset of edges are tolled.

We compute the tolls in each pricing scheme using a two-step approach. First, we characterize the set of tolls that minimize the total travel time (i.e. efficiency objective). Second, we select a particular toll price in the set of tolls computed in the first step to optimize for an objective that achieves the trade-off between average welfare of all populations and the equity across different populations. Under the hom and het pricing schemes, the set of tolls that minimize the total travel time (as computed in the first step) can be characterized as the set of solutions of a linear program, and the second step of selecting a particular toll price is also an optimal solution of a linear program (Proposition 9.2.3) and het (Proposition 9.2.4). The two step approach and the linear program formulations build on the study of enforceable equilibrium flows in routing games with heterogeneous populations [142, 206, 435, 175]. On the other hand, under hom_sc and het_sc, direct extensions of the two linear programs to include toll support set constraints are not guaranteed to achieve the efficiency goal. In fact, the problem of designing congestion minimizing pricing schemes with support constraints is known to be NP hard without the consideration of heterogeneous value-of-time [182, 56, 174]. Building on the linear programming based approaches developed for the pricing schemes without support constraints, we propose a linear programming based heuristic to compute tolls with support constraints and evaluate their efficiency outcomes in the case study.

We next apply our results to evaluate the performances of the four congestion pricing schemes in the San Francisco Bay Area freeway network. Populations in the San Francisco Bay area exhibit significant socioeconomic disparities. This is evident from the distribution of median annual individual income of each neighborhood as shown in Figure 9.1. Moreover, the area has low public transport coverage and thus majority of the populations commute via car. We can see in Figure 9.2 that the driving population percentage of most zip codes outside of San Francisco and Oakland cities are higher than 60%. Moreover, zip codes that are on the east side of the Bay Area have both a high percentage of driving population and a low median individual income. This observation underscores the importance to design efficient and equitable congestion pricing schemes that account for the socioeconomic and geographic disparities, and the disproportionate impact of tolling on different populations.

We model the freeway network in the San Francisco Bay Area as a network with 17 nodes (Figure 9.3). Each node represents a major work or home location for travelers, and

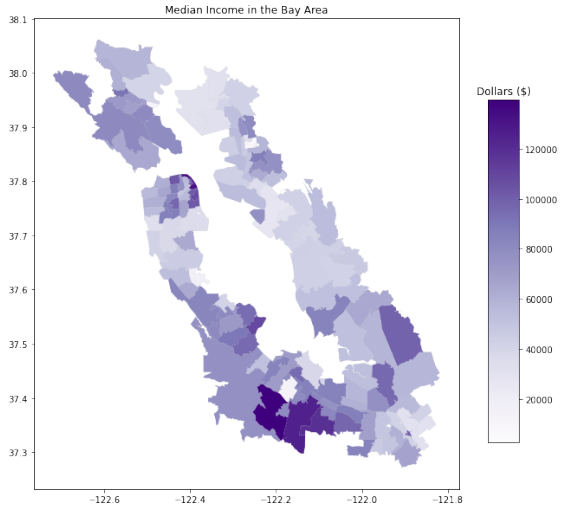


Figure 9.1: Median income.

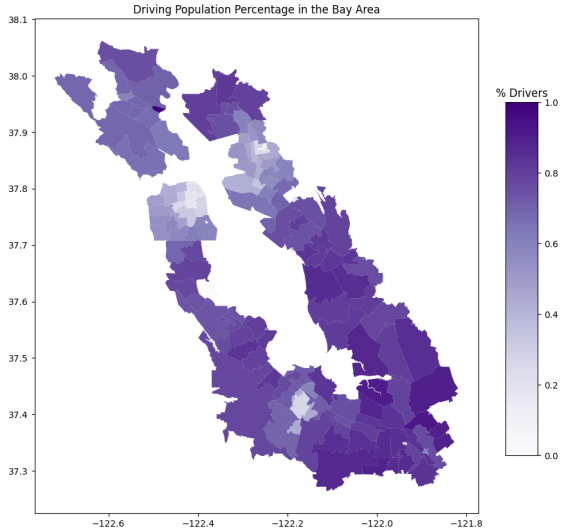


Figure 9.2: Driving population percentage.

the edges represent the primary freeways connecting these locations. Since we differentiate populations based on their value-of-time, which is a latent parameter that cannot be directly estimated from the data, we use the median individual income as a proxy ([19, 167, 423, 406]) to categorize travelers with home at each node into three types of populations with low, middle and high value-of-time, respectively.

Using high-fidelity datasets from Safegraph, the Caltrans Performance Measurement System (PeMS), and the American Community Survey (ACS), we calibrate the latency function of each edge and the demand of each traveler population between each pair of nodes.

The current congestion pricing scheme, denoted as curr sets \$7 price on each of the bridges in the Bay Area (Figure 9.3). We compute the four congestion pricing schemes (hom, het, hom_sc, het_sc), and compare the resulting equilibrium routing behavior in comparison to curr and the zero pricing scheme that set no tolls. We summarize our finding below:

(i) Efficiency and Equity: All four proposed pricing schemes leads to a lower value of total travel time compared to curr. Surprisingly, curr is also marginally outperformed by zero. This is primarily attributed to the fact that the homogeneous toll price of \$7 on all bridges under curr does not account for the heterogeneous distribution of populations between different home-work locations. We show that hom and het achieve the minimum congestion, as indicated by our theoretical result (Proposition 9.2.2). Additionally, hom_sc and het_sc achieve lower value of total travel time than curr and zero but higher than hom and het. Furthermore, we find that the price of anarchy (POA) – the ratio between the total travel time in equilibrium with no tolls and that of the minimum value total travel time [354] – in our setup is 1.04, which is close to 1. This is likely due to high total demand of travelers in the Bay area network since POA always converges to 1 in routing games as the population demand increases [102, 103].

We find that all pricing schemes, except hom, result in lower travel costs for all traveler populations compared to curr. Additionally, our results show that curr is outperformed by all other schemes, except hom, even on the equity metric.

(ii) Revenue Generation: We observe that the revenue generated by hom is the highest as it charges high tolls to all travelers in order to achieve the minimum congestion. Moreover, the revenues generated by het, hom_sc and het_sc are comparable to curr with het being marginally higher and, hom_sc and het_sc, marginally lower.

The rest of the chapter is organized as follows: Section 9.1 presents the model of routing games with heterogeneous populations. Section 9.2 presents the computation methods for the four congestion pricing schemes (hom, het, hom_sc, and het_sc). Section 9.3 presents calibration of the routing game model in the San Francisco Bay Area. Section 9.4 presents the efficiency and equity evaluation of the proposed pricing schemes and the comparison of the emerging congestion patterns.

Related Works

The literature on designing congestion pricing schemes can be categorized into two main threads: *first-best* and *second-best*. First-best pricing schemes allow tolls to be placed on

every edge of the network. The most popular first-best tolling scheme is marginal cost pricing, which sets the toll price to be the marginal cost created by an additional unit of congestion on each edge. [16, 39, 389, 354]. Additionally, an extensive line of research in this thread also focuses on characterizing the set of all congestion-minimizing toll prices (see [435], and references therein). On the other hand, second-best pricing schemes restrict the set of edges that can be tolled. The literature on second-best pricing schemes primarily focuses on formulating the problem as a mathematical program with equilibrium constraints (MPEC) and developing algorithms to approximate the optimal solution (e.g. [436, 67, 138, 222, 227, 247, 332, 412, 230, 128, 204]). The papers [182, 56, 174] studied the problem of characterizing the hardness of the problem of designing second-best tolls. The paper [182] showed that it is NP hard to compute optimal tolls on a subset of edges in general networks and gave a polynomial time algorithm to solve the problem for the parallel link case with affine latency functions. This was extended to allow for non-affine latency functions by [174], and upper bound on the toll values in [56]. In our setup, hom and het are first-best pricing schemes and hom_sc and het_sc are second-best pricing schemes. We contribute to this line of literature by proposing a multi-step linear programming based approach to compute hom and het that account for the equity objective and the heterogeneous traveler populations. Our approach is also an efficient heuristic to solve hom_sc and het_sc with atmost three linear programs instead of iteratively computing the Wardrop equilibrium.

The literature on congestion pricing has mostly focused on homogeneous pricing schemes with a few exceptions. The paper [137] considered tolling schemes that differentiate conventional vehicles from clean energy vehicles. Moreover, differentiated tolls are also used in [233, 232, 293] to study mixed autonomy. The paper [69] studied the impact of differentiated tolling in parallel-link networks with affine cost functions and travelers that have heterogeneous value-of-time.

One effort to ameliorate the inequities resulting from congestion pricing is to redistribute the toll revenue (see [387, 161, 113, 3, 169, 194], or references therein), or provide tradable or untradable travel credits (see [453, 248, 123, 317, 431], or references therein). The papers [160, 387] were amongst the first to propose different ways to redistribute the revenue in form of infrastructure development and tax rebates. The effectiveness of redistribution schemes are theoretically analyzed in single-lane bottleneck models [16, 44], parallel networks [3], and single origin-destination network [129].

Pareto-improving congestion pricing schemes were introduced as another approach to reduce inequality. First proposed by [229], Pareto-improving congestion pricing minimizes the total congestion while ensuring that no travelers are worse off in comparison to no tolls. The paper [392] studied the design of Pareto-improving schemes for travelers with heterogeneous value-of-time, and [231] further proved that such Pareto-improving schemes only exist in special classes of networks. The paper [169] studied the problem of designing Pareto-improving pricing schemes combined with revenue refund. [194] extended this line of research by developing optimal revenue refunding schemes to minimize the congestion and inequity together. In both [169] and [194], the tolls minimize the weighted sum of travel times with weights being each population's value-of-time. This objective is different from

our goal of minimizing the unweighted total travel time, which is a more suitable metric to assess the environmental impact of congestion.

The third approach to addressing inequality is the study of *fairness constrained traffic assignment problem* proposed by [189], where the fairness metric is the maximum difference of travel time experienced by travelers between the same origin-destination pair. [13, 14] extended this line of research by developing algorithmic methods to solve the fairness constrained traffic assignment problem. The problem of devising congestion pricing schemes which could enforce the resulting traffic assignment patters was studied in [195]. Particularly, [195] studies homogeneous pricing scheme that implements the traffic assignment minimizing an interpolation of the potential function (which is used to characterize the equilibrium) and the social cost function.

Our work contributes to all of the above studies on the equity of congestion pricing from three aspects: *(i)* Our equity consideration accounts for both the travel time cost and the monetary cost that includes both the toll and the gas prices. This generalizes the fairness notion that focuses only on the travel time difference; *(ii)* Our tolling scheme minimizes the total congestion in the network (i.e. guarantees the optimal efficiency) while providing the central planner a flexible way to trade-off between the total welfare, equity across heterogeneous populations and total revenue. In particular, by tuning the parameter that governs the trade-off between the average welfare and equity, we can increase or reduce the revenue collected by the our pricing scheme; *(iii)* We provide a comprehensive evaluation of different congestion pricing schemes in terms of efficiency, equity and revenue using real-world data collected in the San Francisco Bay Area.

Another line of research related to this chapter is on developing inverse optimization based tools to estimate model parameters in non-atomic routing games such as demand, latency functions etc [429, 47, 445]. There are several differences between our approach and these works. First, we use high fidelity datasets to directly estimate the latency on every edge and the demand of travelers. Second, we consider heterogeneous population of travelers as opposed to the homogeneous population of travelers considered in these works.

Finally, on the empirical side, [147, 311, 31, 448] focused on understanding the impact of congestion pricing of the San Francisco-Oakland Bay Bridge, which is the most heavily congested segment in the San Francisco Bay Area highway network. Our work generalizes this line of work to the entire Bay Area highway network using high-fidelity mobility and socioeconomic datasets.

9.1 Model

In this section, we introduce the non-atomic networked routing game model that forms the basis for our theoretical and computational results. We introduce equilibrium routing and the four types of congestion pricing schemes we consider in this chapter.

Network

Consider a transportation network $G = (N, E)$, where N is the set of nodes, and E is the set of edges. A set of non-atomic travelers (agents) make routing decisions in the network between their origin and destination. We denote the set of origin-destination (o-d) pairs as K and the set of routes (i.e. sequences of edges) connecting each o-d pair $k \in K$ as R^k.

Travelers for each o-d pair k are grouped into I populations, where each population is associated with a different level of *value-of-time* $\theta^i \in \mathbb{R}_{\geqslant 0}$ that captures the trade-off travelers in population i are willing to make between travel time and monetary cost while selecting between different routes. We refer to agents with value-of-time θ^i as type i agents. The demand vector is given by $D = (D^{ik})_{i\in I,k\in K}$, where D^{ik} is the demand of agents with type i that want to travel between o-d pair k. Throughout this chapter, we operate under the *inelastic demand* assumption: traveler demands on each origin-destination pair are constant. This assumption is reasonable given that *(i)* our analysis focuses on the commuting behavior during the morning rush hour, when the majority of trips are work-related with little elasticity; *(ii)* the availability of public transit is sparse and the cost of car ownership is high [119].

The strategy distribution of agents is denoted $q = (q_r^{ik})_{r\in R^k, i\in I, k\in K}$, where q_r^{ik} is the flow of agents with type i and o-d pair k who take route r. Therefore, given a demand D, the set of feasible strategy distributions is given by:

$$\mathcal{Q}(D) := \left\{ q : \sum_{r\in R^k} q_r^{ik} = D^{ik}, q_r^{ik} \geqslant 0, \forall r \in R^k, i \in I, k \in K \right\}. \tag{9.1}$$

Given a strategy distribution $q \in \mathcal{Q}(D)$, the flow of agents of type $i \in I$ on edge $e \in E$ is given by

$$f_e^i(q) := \sum_{k\in K} \sum_{r\in R^k} q_r^{ik} \mathbb{1}(e \in r), \tag{9.2}$$

and the total flow of agents on edge $e \in E$ is

$$w_e(q) := \sum_{i\in I} f_e^i(q). \tag{9.3}$$

The travel time experienced by agents taking edge $e \in E$ is $\ell_e(w_e(q))$, where the latency function $\ell_e \colon \mathbb{R}_+ \to \mathbb{R}_+$ is *continuous, strictly increasing, and convex.* Consequently, the total travel time experienced by agents from o-d pair $k \in K$ who use route $r \in R^k$ is given by $\ell_r(q) := \sum_{e\in r} \ell_e(w_e(q))$. With slight abuse of notation, we use $\ell_r(q)$ and $\ell_r(w)$ interchangeably to represent the latency of route r where w is the edge flow vector corresponding to the strategy distribution q. In addition to the travel time, the total cost experienced by each individual agent also includes the congestion price imposed by the planner, and the gas cost required to travel on the route the agent chooses. In particular, let p_e^i be the toll

price imposed on travelers of type $i \in I$ for using edge $e \in E$, and g_e be the gas cost of using an edge $e \in E$. Note that we allow for the toll price to be type-specific in the general setting. We will later discuss different scenarios for setting the toll prices. Given the tolls $p = (p_e^i)_{e\in E, i\in I}$, the cost experienced by travelers of type $i \in I$ associate with o-d pair $k \in K$ and taking route $r \in R^k$ is given by

$$c_r^i(q, p) := \ell_r(q) + \frac{1}{\theta^i} \sum_{e\in r} (p_e^i + g_e). \tag{9.4}$$

Crucially, a key feature of our model is that the toll and gas costs experienced by each agent are modulated by the value-of-time θ^i of that agent. This allows us to model the heterogeneity present in the types of travelers. Given this setup, we define Nash equilibrium to be the strategy distribution such that no traveler has incentive to deviate from their chosen route. That is,

Definition 9.1.1. *Given tolls p, a strategy profile $q^*(p)$ is a Nash equilibrium if*

$$\begin{aligned} &\forall i \in I, k \in K, r \in R^k, \quad q_r^{ik*}(p) > 0 \\ &\Rightarrow \quad c_r^i(q^*(p), p) \leqslant c_{r'}^i(q^*(p), p) \quad \forall r' \in R^k. \end{aligned}$$

The objective of the planner is to minimize the network congestion, measured by the total travel time experienced by all travelers. For any strategy distribution q, we denote the planner's cost function as follows:

$$S(q) := \quad \sum_{e\in E} w_e(q)\ell_e(w_e(q)), \tag{9.5}$$

where $w_e(q)$ is given by (9.3). We denote the set of socially optimal strategy distributions as $q^\dagger := \arg\min_{q\in\mathcal{Q}(D)} S(q)$, and the induced socially optimal edge flows as $w^\dagger = (w_e^\dagger)_{e\in E}$, where $w_e^\dagger = w_e(q^\dagger)$ given by (9.3).

Congestion pricing

We now introduce two practical considerations for toll implementation. The first consideration is whether or not the toll is type-specific. In particular, a congestion pricing scheme is *homogeneous* if the toll is uniform across all population types, and *heterogeneous* if the toll varies with population types (formally, whether p_e^i is allowed to depend on i or not, on each edge). The challenge of implementing a heterogeneous scheme is that the population type (i.e. value-of-time) is a latent variable that is privately known only by the individual traveler. In practice, an individual's value-of-time is often closely correlated with their income level, i.e. higher-income groups are typically associated with a higher value-of-time, while lower-income groups correlate with a lower value-of-time [19, 167, 423, 406]. Therefore, one way to implement heterogeneous tolling is to set tolls based on the income level of travelers.

For example, low income groups, which have significant overlap with the population of low value-of-time travelers, may receive a subsidy or a toll rebate in certain areas. Such toll relief programs have been established in several states in the United States, e.g. California [1] , Virginia [2] , New York [3] etc.

The second consideration is whether or not tolls can be set on all the edges of the network or only on a subset (formally, whether or not p_e^i is allowed to be strictly positive on all $e \in E$). In practical terms, congestion pricing often requires the installation of toll collection facilities, which might not be feasible on all road segments. Thus, a congestion pricing scheme has no support constraints if tolls can be imposed on all edges, or has support constraints if tolls can only be imposed on a subset of edges, denoted as E_T. We note that congestion pricing schemes with (resp. without) support constraints are also referred as first-best (resp. second-best) tolling schemes in literature.

Building on the above two considerations, we define four types of tolling schemes: *(i) Homogeneous tolls with no support constraints* (hom): $p_e^i \geqslant 0$ and $p_e^i = p_e^j$ for all $e \in E$ and all $i, j \in I$; *(ii) Heterogeneous tolls with no support constraints* (het): $p_e^i \geqslant 0$ for all $e \in E, i \in I$; *Homogeneous tolls with support constraints* (hom_sc): $p_e^i = p_e^j$ for all $i, j \in I$ and all $e \in E$. Additionally, $p_e^i = 0$ for all $e \in E \backslash E_T$, and $p_e^i \geqslant 0$ for all $e \in E_T$.*(iii)* Building on the above two considerations, we define four types of tolling schemes: *(i) Homogeneous tolls with no support constraints* (hom): $p_e^i \geqslant 0$ and $p_e^i = p_e^j$ for all $e \in E$ and all $i, j \in I$; *(ii) Heterogeneous tolls with no support constraints* (het): $p_e^i \geqslant 0$ for all $e \in E, i \in I$; *(iii) Homogeneous tolls with support constraints* (hom_sc): $p_e^i = p_e^j$ for all $i, j \in I$ and all $e \in E$. Additionally, $p_e^i = 0$ for all $e \in E \backslash E_T$, and $p_e^i \geqslant 0$ for all $e \in E_T$; *(iv) Heterogeneous tolls with support constraints* (het_sc): $p_e^i = 0$ for all $e \in E \backslash E_T$, and $p_e^i \geqslant 0$ for all $e \in E_T$.

9.2 Computation Methods

In this section, we outline methods for computing equilibrium routing strategies and the four congestion pricing schemes. We first establish that, given any fixed toll values, the equilibrium outcome can be derived as the optimal solution to a convex optimization problem. We then demonstrate that the set of homogeneous tolls (hom) and heterogeneous tolls (het) without support constraints that realize the socially optimal edge flows can be characterized as the set of optimal solutions of linear programs. Next, we present a multi-step approach for calculating the toll prices that strikes a balance between equity, as measured by the cost disparity between travelers from different populations, and at the same time, maximizing the welfare of all traveler populations. For congestion pricing schemes with support constraints, we adapt our approach to provide a heuristic for calculating hom_sc and het_sc, acknowledging that such solutions may not guarantee the implementation of the socially optimal edge flows.

[1] https://mtc.ca.gov/news/new-year-brings-new-toll-payment-assistance-programs

[2] https://www.vdottollrelief.com/

[3] https://new.mta.info/fares-and-tolls/bridges-and-tunnels/resident-programs

Proposition 9.2.1. *Given toll prices p, a strategy distribution $q^*(p)$ is a Nash equilibrium if and only if it is a solution to the following convex optimization problem:*

$$\min_{q\in\mathcal{Q}(D)} \quad \Phi(q,p,\theta) = \sum_{e\in E}\int_0^{w_e(q)} \ell_e(z)\ \mathrm{d}z + \sum_{i\in I}\sum_{e\in E}\frac{(p_e^i+g_e)}{\theta^i}f_e^i(q), \tag{9.6}$$

where $w_e^i(q)$, $w_e(q)$ are given by (9.2) and (9.3), respectively. Moreover, given any toll price vector p, the equilibrium edge flow vector $w^*(p) := w(q^*(p))$ is unique. Additionally, the socially optimal edge flow vector $w^\dagger$ is unique.

[169] showed the same result as Proposition 9.2.1 without the gas price. The proof follows directly from [169], and is available in the Appendix H.

Proposition 9.2.1 shows that $w^\dagger$ is unique. However, we note that such a $w^\dagger$ may be induced by multiple type-specific flow vectors $f^\dagger$. Although these different type-specific flow vectors all induce the same aggregate edge load, and thus minimize the total cost, they may lead to different travel times experienced by different populations.

The following proposition shows that the set of prices hom (resp. het) that implements the socially optimal edge load can be characterized each by a linear program.

Proposition 9.2.2. *(1) A homogeneous congestion pricing scheme $p^\dagger = (p_e^\dagger)_{e\in E}$ implements the socially optimal edge flow $w^\dagger = (w_e^\dagger)_{e\in E}$ if and only if there exists $z^\dagger$ such that $(p^\dagger, z^\dagger)$ is a solution to the following linear program:*

$$\begin{aligned} T^*_{\mathsf{hom}} = \max_{p,z}\ & \sum_{i\in I}\sum_{k\in K} D^{ik}z^{ik} - \sum_{e\in E} p_e w_e^\dagger, \\ s.t.\quad & z^{ik} - \sum_{e\in r}(p_e+g_e) \leqslant \theta^i \ell_r(w^\dagger), \\ & \forall k\in K, r\in R^k, i\in I, \\ & p_e \geqslant 0, \quad \forall e\in E. \end{aligned} \tag{$\mathcal{P}_{\mathsf{hom}}$}$$

(2) A heterogeneous congestion pricing scheme $p^\dagger = (p_e^{i\dagger})_{e\in E, i\in I}$ implements a type-specific socially optimal edge flow $f^{\dagger,} = (f_e^{\dagger i,})_{e\in E, i\in I}$ if and only if there exists a $z^\dagger$ such that $(p^\dagger, z^\dagger)$ is a solution to the following linear program:

$$\begin{aligned} T^*_{\mathsf{het}}(f^\dagger) = \max_{p,z}\ & \sum_{i\in I}\sum_{k\in K} D^{ik}z^{ik} - \sum_{e\in E}\sum_{i\in I} p_e^i f_e^{\dagger i}, \\ s.t.\ & z^{ik} - \sum_{e\in r}(p_e^i+g_e) \leqslant \theta^i \ell_r(w^\dagger), \\ & \forall k\in K, r\in R^k, i\in I, \\ & p_e^i \geqslant 0, \quad \forall e\in E, i\in I. \end{aligned} \tag{$\mathcal{P}_{\mathsf{het}}$}$$

Proposition 9.2.2 follows the results in [142, 206, 285, 435, 175]. The proof builds on the two linear programs $(\mathcal{P}_{\mathsf{hom}})$ – $(\mathcal{P}_{\mathsf{het}})$ and their dual programs $(\mathcal{D}_{\mathsf{hom}})$ and $(\mathcal{D}_{\mathsf{het}})$ as follows:

$$\min_{q} \quad \sum_{i\in I}\sum_{k\in K}\sum_{r\in R^k}(\theta^i \ell_r(w^\dagger) + \sum_{e\in r} g_e) q_r^{ik} \qquad (\mathcal{D}_{\mathsf{hom}})$$

$$\text{s.t.} \quad \sum_{i\in I}\sum_{k\in K}\sum_{r\in R^k: e\in r} q_r^{ik} \leqslant w_e^\dagger, \quad \forall e \in E, \qquad (\mathcal{D}_{\mathsf{hom}.a})$$

$$\sum_{r\in R^k} q_r^{ik} = D^{ik}, \quad \forall i \in I, k \in K, \qquad (\mathcal{D}_{\mathsf{hom}.b})$$

$$q_r^{ik} \geqslant 0 \quad \forall i \in I, k \in K, r \in R^k. \qquad (\mathcal{D}_{\mathsf{hom}.c})$$

$$\min_{q} \quad \sum_{i\in I}\sum_{k\in K}\sum_{r\in R^k}(\theta^i \ell_r(w^\dagger) + \sum_{e\in r} g_e) q_r^{ik} \qquad (\mathcal{D}_{\mathsf{het}})$$

$$\text{s.t.} \quad \sum_{k\in K}\sum_{r\in R^k: e\in r} q_r^{ik} \leqslant f_e^{\dagger i}, \quad \forall e \in E, i \in I, \qquad (\mathcal{D}_{\mathsf{het}.a})$$

$$\sum_{r\in R^k} q_r^{ik} = D^{ik}, \quad \forall i \in I, k \in K, \qquad (\mathcal{D}_{\mathsf{het}.b})$$

$$q_r^{ik} \geqslant 0, \quad \forall i \in I, k \in K, r \in R^k. \qquad (\mathcal{D}_{\mathsf{het}.c})$$

Under both hom and het, the feasibility constraints of the associated primal and dual programs as well as the complementary slackness conditions are equivalent to the equilibrium condition where only routes with the minimum cost are taken by travelers. Moreover, constraints $(\mathcal{D}_{\mathsf{hom}.a})$ and $(\mathcal{D}_{\mathsf{het}.a})$ must be tight at optimality, indicating that the induced flow vector in equilibrium is indeed $w^\dagger$, which minimizes total travel time. Therefore, the set of optimal solutions of $(\mathcal{P}_{\mathsf{hom}})$ and $(\mathcal{P}_{\mathsf{het}})$ are the set of toll vectors that induce $w^\dagger$ under hom and het, respectively.

We denote $P^\dagger_{\text{hom}}$ as the set of socially optimal toll prices for hom, and $P^\dagger_{\mathsf{het}}(f^\dagger)$ as the set of socially optimal toll price for het that induces a type-specific socially optimal edge flow $f^\dagger$. Proposition 9.2.2 demonstrates that both sets can be computed as the optimal solution set of linear programs. We note that the set $P^\dagger_{\mathsf{het}}(f^\dagger)$ depends on which type-specific socially optimal flow $f^\dagger$ is induced since the objective function $(\mathcal{P}_{\mathsf{het}})$ depends on $f^\dagger$.

Furthermore, $P^\dagger_{\text{hom}}$ and $P^\dagger_{\mathsf{het}}(f^\dagger)$ may not be singleton. This presents an opportunity for the planner to decide which specific toll price from the optimal solution set to implement. While all tolls in $P^\dagger_{\text{hom}}$ and $P^\dagger_{\mathsf{het}}(f^\dagger)$ achieve the minimum social cost, they do so by impacting travelers differently given their individual origin-destination pair and value-of-time. We

consider that the central planner aims at solving the following problem:

$$\begin{aligned}
&\min_{p} \; L(p) := \\
&\underbrace{\max_{i,i' \in I} \left| \frac{1}{D^i} \sum_{k \in K} D^{ik} \frac{c^{ik\dagger}(p)}{c^{ik\dagger}(0)} - \frac{1}{D^{i'}} \sum_{k' \in K} D^{i'k'} \frac{c^{i'k'\dagger}(p)}{c^{i'k'\dagger}(0)} \right|}_{(i)} \\
&\qquad + \lambda \underbrace{\frac{1}{D} \sum_{i \in I} \sum_{k \in K} D^{ik} \frac{c^{ik\dagger}(p)}{c^{ik\dagger}(0)}}_{(ii)}, \qquad (9.8) \\
&s.t. \qquad p \in \begin{cases} P^{\dagger}_{\text{hom}}, & \text{in } \mathsf{hom}, \\ P^{\dagger}_{\mathsf{het}}(f^{\dagger}), & \text{in } \mathsf{het} \text{ with route flow } f^{\dagger}, \end{cases}
\end{aligned}$$

where $D^i = \sum_{k \in K} D^{ik}$, $D = \sum_{i \in I} D^i$, $\lambda \geqslant 0$,

$$c^{ik\dagger}(p) = \min_{r \in R^k} \left\{ \ell_r(w^{\dagger}) + \frac{1}{\theta^i} \sum_{e \in r} (p_e + g_e) \right\} \qquad (9.9)$$

is the equilibrium cost of individuals with o-d pair k and type i given the toll price p and socially optimal edge load $w^{\dagger}$, and $c^{ik\dagger}(0)$ is the equilibrium cost of individuals with o-d pair k and type i given no (or zero) tolls. We emphasize that the cost $c^{ik\dagger}(p)$ is the minimum cost of choosing a route given the socially optimal load vector $w^{\dagger}$, the toll price, and the gas fee. This is indeed an equilibrium cost of traveler population with type i and o-d pair k since any $p \in P^{\dagger}_{\text{hom}}$ or $p \in P^{\dagger}_{\mathsf{het}}(f^{\dagger})$ guarantees that the equilibrium edge vector is $w^{\dagger}$.

The objective function in (9.8) indicates that the central planner selects the toll price that not only minimizes the total travel time but also balances the equity among populations with different value-of-time, and the average welfare that accounts for the travel time as well as the toll price and gas fee. In particular, in (9.8) *(i)* reflects an equity objective by assessing the maximum disparity in the relative change in travel costs experienced by different types of travelers following the implementation of tolls, compared to a scenario without tolls, and *(ii)* reflects the an average welfare objective that is the average of the relative change in travel costs experienced by all of the travelers following the implementation of tolls, compared to a scenario without tolls. Balancing welfare maximization with cost disparity minimization avoids the potential problem with just minimizing cost disparity: charging excessively high tolls to every type of travelers. Moreover, $\lambda \geqslant 0$ is a parameter that governs the relative weight between the equity objective and the welfare objective.

We denote the socially optimal homogeneous congestion pricing scheme that solves the central planner's problem (9.8) as p^*_{hom}. The next proposition shows that we can solve the central planner's problem (9.8) for hom by another linear program.

Proposition 9.2.3. *For the* hom *tolling scheme,* p^*_{hom} *is an optimal solution of the following linear program:*

$$\min_{p,z,y} \quad y + \frac{\lambda}{D}\sum_{i\in I}\sum_{k\in K} D^{ik}\frac{z^{ik}}{\theta^i c^{ik\dagger}(0)} \tag{$\mathcal{P}^*_{\mathsf{hom}}$}$$

$$\text{s.t.} \quad y \geqslant \frac{1}{D^i}\sum_{k\in K} D^{ik}\frac{z^{ik}}{\theta^i c^{ik\dagger}(0)} - \frac{1}{D^{i'}}\sum_{k'\in K} D^{i'k'}\frac{z^{i'k'}}{\theta^{i'} c^{i'k'\dagger}(0)},$$

$$\forall i, i' \in I, \tag{$\mathcal{P}^*_{\mathsf{hom}.a}$}$$

$$\sum_{i\in I}\sum_{k\in K} D^{ik} z^{ik} - \sum_{e\in E} p_e w^\dagger_e \geqslant T^*_{hom}, \tag{$\mathcal{P}^*_{\mathsf{hom}.b}$}$$

$$z^{ik} - \sum_{e\in r}(p_e + g_e) \leqslant \theta^i \ell_r(w^\dagger),\ \forall k \in K, r \in R^k, i \in I, \tag{$\mathcal{P}^*_{\mathsf{hom}.c}$}$$

$$p_e \geqslant 0 \quad \forall e \in E, \tag{$\mathcal{P}^*_{\mathsf{hom}.d}$}$$

where T^*_{hom} *is the optimal value of* $(\mathcal{P}_{\mathsf{hom}})$.

In $(\mathcal{P}^*_{\mathsf{hom}})$, constraints $(\mathcal{P}^*_{\mathsf{hom}.c})$ and $(\mathcal{P}^*_{\mathsf{hom}.d})$ ensure that variables (p, z) are in the feasible set of $(\mathcal{P}_{\mathsf{hom}})$, and constraint $(\mathcal{P}^*_{\mathsf{hom}.b})$ further restrict that the set of (p, z) in $(\mathcal{P}^*_{\mathsf{hom}})$ to be the set of optimal solutions of $(\mathcal{P}_{\mathsf{hom}})$. Thus, following Proposition 9.2.2, any feasible p in $(\mathcal{P}^*_{\mathsf{hom}})$ must be a toll vector that induces the socially optimal edge flow $w^\dagger$. Moreover, the proof of Proposition 9.2.2 further ensures that for every $i \in I, k \in K$ there exists $r \in R^k$ such that the corresponding constraint in $(\mathcal{P}^*_{\mathsf{hom}.c})$ must be tight at optimum, which indicates that any z^{ik} in $(\mathcal{P}^*_{\mathsf{hom}})$ equals to $\theta^i \cdot c^{ik\dagger}(p)$. Additionally, constraints $(\mathcal{P}^*_{\mathsf{hom}.a})$ guarantee that at optimality $y = \max_{i,i'\in I}\left|\frac{1}{D^i}\sum_{k\in K} D^{ik}\frac{c^{ik\dagger}(p)}{c^{ik\dagger}(0)} - \frac{1}{D^{i'}}\sum_{k'\in K} D^{i'k'}\frac{c^{i'k'\dagger}(p)}{c^{i'k'\dagger}(0)}\right|$. Thus, the linear program $(\mathcal{P}^*_{\mathsf{hom}})$ computes the homogeneous toll prices that minimize the total travel time and optimize the equity and welfare objectives with a relative weight λ.

To summarize, the programs $(\mathcal{P}_{\mathsf{hom}})$ and $(\mathcal{P}^*_{\mathsf{hom}})$ provide a *two-step approach* for computing p^*_{hom}: first, compute T^*_{hom} by solving the linear program $(\mathcal{P}_{\mathsf{hom}})$ given the unique edge flow $w^\dagger$. Second, compute p^*_{hom} by solving the linear program $(\mathcal{P}^*_{\mathsf{hom}})$ using T^*_{hom}.

Next, we show that the central planner can compute the heterogeneous toll price vector (het) that minimize the total travel time and optimize the equity-welfare objectives (9.8), denoted as p^*_{het}, using a similar approach as described above. However, in het, one additional issue arises as the set $P^\dagger_{\mathsf{het}}(f^\dagger)$ and consequently p^*_{het} depend on the selection of the type-specific flow vector $f^\dagger$, which may not be unique. Here, we propose to select the type-specific flow vector $f^\dagger$ as the one that induces the edge flow vector $w^\dagger$ (which minimizes total travel time) while also minimizing the disparity in the total travel time experienced across all traveler populations. To compute such a $f^\dagger$, we first find a feasible routing strategy profile $q^\dagger$ that induces $w^\dagger$ and minimizes the average cost difference among traveler populations.

Such a $q^\dagger$ can be solved by the following linear program:

$$\begin{aligned}
\min_{q} \quad & x, \\
\text{s.t.} \quad & x \geqslant \sum_{k\in K}\sum_{r\in R^k}\left(q_r^{ik}\ell_r(w^\dagger) - q_r^{i'k}\ell_r(w^\dagger)\right), \forall i, i' \in I, \\
& \sum_{r\in R^k} q_r^{ik} = D^{ik}, \forall\ i \in I, k \in K, \\
& \sum_{i\in I}\sum_{k\in K}\sum_{r\in R^k: e\in r} q_r^{ik} = w_e^\dagger, \forall e \in E, \\
& q_r^{ik} \geqslant 0, \forall\ i \in I, k \in K, r \in R^k.
\end{aligned}$$

Then, the induced population-specific flow vector $f^\dagger$ associated with $q^\dagger$ is given by (9.2). Based on $f^\dagger$, we compute p^*_{het} as the optimal solution of a linear program.

Proposition 9.2.4. *For the* het *tolling scheme, given $f^\dagger$, p^*_{het} is an optimal solution of the following linear program:*

$$\begin{aligned}
\min_{p,z,y} \quad & y + \frac{\lambda}{D}\sum_{i\in I}\sum_{k\in K} D^{ik}\frac{z^{ik}}{\theta^i c^{ik\dagger}(0)} && (\mathcal{P}^*_{\mathsf{het}}) \\
s.t. \quad & y \geqslant \frac{1}{D^i}\sum_{k\in K} D^{ik}\frac{z^{ik}}{\theta^i c^{ik\dagger}(0)} - \frac{1}{D^{i'}}\sum_{k'\in K} D^{i'k'}\frac{z^{i'k'}}{\theta^{i'} c^{i'k'\dagger}(0)}, \\
& \qquad \forall i, i' \in I, && (\mathcal{P}^*_{\mathsf{het}.a}) \\
& \sum_{i\in I}\sum_{k\in K} D^{ik} z^{ik} - \sum_{e\in E} p_e w_e^\dagger \geqslant T^*_{\mathsf{het}}(f^\dagger), && (\mathcal{P}^*_{\mathsf{het}.b}) \\
& z^{ik} - \sum_{e\in r}(p_e^i + g_e) \leqslant \theta^i \ell_r(w^\dagger), \forall k \in K, r \in R^k, i \in I, && (\mathcal{P}^*_{\mathsf{het}.c}) \\
& p_e^i \geqslant 0, \forall e \in E, i \in I, && (\mathcal{P}^*_{\mathsf{het}.d})
\end{aligned}$$

where $T^*_{\mathsf{het}}(f^\dagger)$ is the optimal value of the objective function of $(\mathcal{P}_{\mathsf{het}})$ associated with $f^\dagger$.

Propositions 9.2.2 and 9.2.4 show that p^*_{het} can be computed using a *three-step approach*: first, we compute the type-specific flow vector $f^\dagger$ that induces the edge flow $w^\dagger$ while also minimizing the travel time difference among all traveler populations using (9.2). Second, we compute $T^*_{\mathsf{het}}(f^\dagger)$ using $(\mathcal{P}_{\mathsf{het}})$ given $f^\dagger$. Third, we compute p^*_{het} using $(\mathcal{P}^*_{\mathsf{het}})$.

Finally, we discuss how to extend our approaches of computing p^*_{hom} and p^*_{het} to incorporate the support constraints of the toll price. Previous studies [182, 56] showed that the problem of computing toll prices that satisfy support set constraints and also minimize the total travel time is NP hard even without considering heterogeneous value-of-time of travelers or equity objectives. Here, we provide heuristics for computing the toll prices with support constraints. We evaluate the performance of our heuristics in terms of total travel time, equity, and welfare on the San Francisco Bay Area network in Sec. 9.2.

Heuristics for computing $p^*_{\mathsf{hom\_sc}}$ We propose a two-step heuristic to compute $\mathsf{hom\_sc}$ by appropriately modifying the two-step method to compute hom.

We first solve the following linear program that adds the support constraints to $(\mathcal{P}_{\mathsf{hom}})$:

$$\begin{aligned} T^*_{\mathsf{hom\_sc}} = \max_{p,z} \ & \sum_{i\in I}\sum_{k\in K} D^{ik} z^{ik} - \sum_{e\in E} p_e w^\dagger_e, \\ \text{s.t.} \quad & z^{ik} - \sum_{e\in r}(p_e + g_e) \leqslant \theta^i \ell_r(w^\dagger), \\ & \qquad \forall k \in K, r \in R^k, i \in I, \\ & p_e \geqslant 0, \quad \forall e \in E_T, \quad p_e = 0, \quad \forall e \in E \setminus E_T. \end{aligned} \tag{$\mathcal{P}_{\mathsf{hom\_sc}}$}$$

We note that the equilibrium edge load associated with any optimal solution of $(\mathcal{P}_{\mathsf{hom\_sc}})$, say $\hat{w}$, may not be equal to the socially optimal edge load $w^\dagger$. This is because the constraints that edges in $E \setminus E_T$ having zero tolls remove the dual constraints in $(\mathcal{D}_{\mathsf{hom}.a})$ for edges in E_T. As a result, the primal and dual argument in the proof of Proposition 9.2.2 no longer holds, and thus the induced edge flow $\hat{w}$ may not be equal to $w^\dagger$.

Note that the optimal solution to $(\mathcal{P}_{\mathsf{hom\_sc}})$ will be non-unique. Therefore, inspired by $(\mathcal{P}^*_{\mathsf{hom}})$, we consider the following heuristic to incorporate both equity and welfare metric while also accounting for support constraints. Note that simply adding the support constraints in $(\mathcal{P}^*_{\mathsf{hom}})$ could render the optimization problem infeasible as the optimal set of homogeneous tolls P^*_{hom} need not have a solution that satisfies the support constraints. Particularly the constraint $(\mathcal{P}^*_{\mathsf{hom}.b})$ would get violated. This is because $T^*_{\mathsf{hom}} \geqslant T^*_{\mathsf{hom\_sc}}$ as the constraint set of $(\mathcal{P}^*_{\mathsf{hom\_sc}})$ is contained in that of $(\mathcal{P}^*_{\mathsf{hom}})$. Therefore, we compute $p^*_{\mathsf{hom\_sc}}$ as the optimal solution of the following linear program which adds support constraints to $(\mathcal{P}^*_{\mathsf{hom}})$ while relaxing the constraint $(\mathcal{P}^*_{\mathsf{hom}.b})$ by using $T^*_{\mathsf{hom\_sc}}$ instead of T^*_{hom}:

$$\min_{p,z,y} \quad y + \frac{\lambda}{D}\sum_{i\in I}\sum_{k\in K} D^{ik}\frac{z^{ik}}{\theta^i c^{ik\dagger}(0)} \tag{$\mathcal{P}^*_{\mathsf{hom\_sc}}$}$$

$$\text{s.t.} \quad y \geqslant \frac{1}{D^i}\sum_{k\in K} D^{ik}\frac{z^{ik}}{\theta^i c^{ik\dagger}(0)} - \frac{1}{D^{i'}}\sum_{k'\in K} D^{i'k'}\frac{z^{i'k'}}{\theta^{i'} c^{i'k'\dagger}(0)},$$

$$\forall i, i' \in I, \tag{$\mathcal{P}^*_{\mathsf{hom\_sc}.a}$}$$

$$\sum_{i\in I}\sum_{k\in K} D^{ik} z^{ik} - \sum_{e\in E} p_e w^\dagger_e \geqslant T^*_{\mathsf{hom\_sc}}, \tag{$\mathcal{P}^*_{\mathsf{hom\_sc}.b}$}$$

$$z^{ik} - \sum_{e\in r}(p_e + g_e) \leqslant \theta^i \ell_r(w^\dagger), \forall k \in K, r \in R^k, i \in I, \tag{$\mathcal{P}^*_{\mathsf{hom\_sc}.c}$}$$

$$p_e \geqslant 0, \quad \forall e \in E_T, p_e = 0, \quad \forall e \in E \setminus E_T. \tag{$\mathcal{P}^*_{\mathsf{hom\_sc}.d}$}$$

Heuristics for computing $p^*_{\mathsf{het\_sc}}$ The computation of $p^*_{\mathsf{het\_sc}}$ follows a three-step procedure, similar to that of p^*_{het} but restricting the set of allowable tolls to be zero on non-tollable edges, as done in $\mathsf{hom\_sc}$. First, we compute the population-specific flow vector $f^\dagger$ that induces the congestion minimizing edge flow $w^\dagger$ while also minimizes the average difference of

travel time among all traveler populations using (9.2). Next, we add the support constraints to $(\mathcal{P}_{\mathsf{het}})$ to compute the optimal value $T^*_{\mathsf{het\_sc}}(f^\dagger)$ as follows:

$$\begin{aligned} T^*_{\mathsf{het\_sc}}(f^\dagger) = \max_{p,z} \ & \sum_{i\in I}\sum_{k\in K} D^{ik} z^{ik} - \sum_{e\in E}\sum_{i\in I} p_e^i f_e^{\dagger i}, \\ \text{s.t. } & z^{ik} - \sum_{e\in r}(p_e^i + g_e) \leqslant \theta^i \ell_r(w^\dagger), \\ & \qquad \forall k \in K, r \in R^k, i \in I, \\ & p_e^i \geqslant 0, \quad \forall e \in E_T \ \forall i \in I, \\ & p_e = 0, \quad \forall e \in E \setminus E_T \ \forall i \in I. \end{aligned} \tag{$\mathcal{P}_{\mathsf{het\_sc}}$}$$

Analogous to the case hom_sc, the equilibrium edge load associated with the optimal solution of $(\mathcal{P}_{\mathsf{het\_sc}})$, $\hat{w}$, may not be equal to the socially optimal edge load $w^\dagger$ due to the added support constraints. We compute $p^*_{\mathsf{het\_sc}}$ as the optimal solution of the following linear program which adds support constraints to $(\mathcal{P}^*_{\mathsf{het}})$ while relaxing the constraint $(\mathcal{P}^*_{\mathsf{het}.b})$ by using $T^*_{\mathsf{het\_sc}}$ instead of T^*_{het}:

$$\min_{p,z,y} \quad y + \frac{\lambda}{D}\sum_{i\in I}\sum_{k\in K} D^{ik} z^{ik}, \tag{$\mathcal{P}^*_{\mathsf{het\_sc}}$}$$

$$\text{s.t.} \quad y \geqslant \frac{1}{D^i}\sum_{k\in K} D^{ik}\frac{z^{ik}}{\theta^i c^{ik\dagger}(0)} - \frac{1}{D^{i'}}\sum_{k'\in K} D^{i'k'}\frac{z^{i'k'}}{\theta^{i'} c^{i'k'\dagger}(0)},$$

$$\forall i, i' \in I, \tag{$\mathcal{P}^*_{\mathsf{het\_sc}.a}$}$$

$$\sum_{i\in I}\sum_{k\in K} D^{ik} z^{ik} - \sum_{e\in E} p_e^i w_e^\dagger \geqslant T^*_{\mathsf{het\_sc}}, \tag{$\mathcal{P}^*_{\mathsf{het\_sc}.b}$}$$

$$z^{ik} - \sum_{e\in r}(p_e^i + g_e) \leqslant \theta^i \ell_r(w^\dagger),$$

$$\forall k \in K, r \in R^k, i \in I, \tag{$\mathcal{P}^*_{\mathsf{het\_sc}.c}$}$$

$$p_e^i \geqslant 0, \ \forall e \in E_T, i \in I, \quad p_e^i = 0, \ \forall e \in E \setminus E_T, i \in I. \tag{$\mathcal{P}^*_{\mathsf{het\_sc}.d}$}$$

9.3 Model Calibration for the San Francisco Bay Area Freeway Network

In this section, we calibrate the non-atomic routing game model for the San Francisco Bay Area freeway network using the Caltrans Performance Measurement System (PeMS) dataset [4], American Community Survey (ACS) dataset [5] and Safegraph neighborhood patterns

[4] available at `https://pems.dot.ca.gov/`

[5] available at `https://www.census.gov/programs-surveys/acs`

dataset from 2019 [6]. In Sec. 9.3, we briefly describe each dataset. We subsequently present the calibration of the Bay Area transportation network, the demand of each population type in Sec. 9.3, and the value-of-time parameters in Sec. 9.3.

Datasets

Caltrans PeMS Dataset The Caltrans PeMS dataset is based on measurements taken from loop detectors placed on a network of freeways and bridges in California. Our dataset is taken from district 4, which covers the entire San Francisco Bay Area. This dataset provides hourly flow counts and average vehicle speeds measured by each loop detector placed along the freeways. We use this dataset to calibrate the latency functions of edges (See Sec. 9.3 for detailed discussion).

American Community Survey (ACS) Dataset The ACS dataset is collected by the US Census Bureau to record demographic and socioeconomic information. We use the information from Means of Transportation (2019) entry in ACS, which provides information of commuters' mode choices (percentage of driving population), employment, and household income. The dataset is collected at the zip-code level for the entire United States.

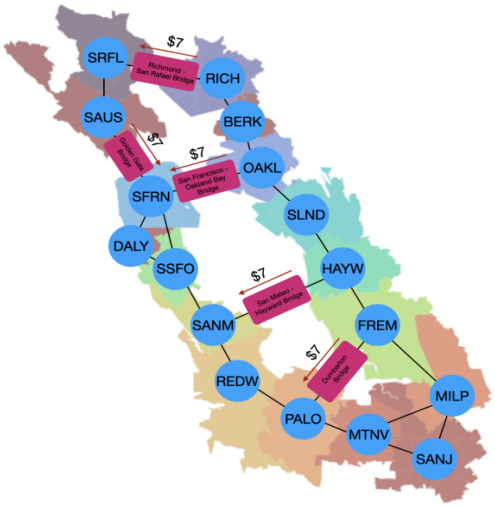
Safegraph Neighborhood Patterns Dataset This dataset records the aggregate mobility pattern using the data collected from 40 million mobile devices in the US. This dataset estimates the commuting pattern by counting the number of mobile devices that travel from one census block group (CBG) to another CBG and dwell for at least 6 hours between 7:30 am and 5:30 pm Monday through Friday. We use this dataset in conjunction with ACS and the Safegraph datasets to estimate the demand of driving commuters between each o-d pair in the network within each income level (See Sec. 9.3 for detailed discussion).

The San Francisco Bay Area freeway network

We represent the San Francisco Bay Area using a network with 17 nodes (see Fig. 9.3). Each node represents a major city, and the edges are the major freeways connecting these cities. Among these edges, five of them are bridges: the Golden Gate Bridge, the Richmond-San Rafael Bridge, the San Francisco-Oakland Bay Bridge, the San Mateo-Hayward Bridge, and the Dumbarton Bridge. They are represented as the magenta boxes in Figure 9.3. In 2019, a flat toll of $7 is imposed for a single crossing on each bridge in the direction denoted in Figure 9.3.

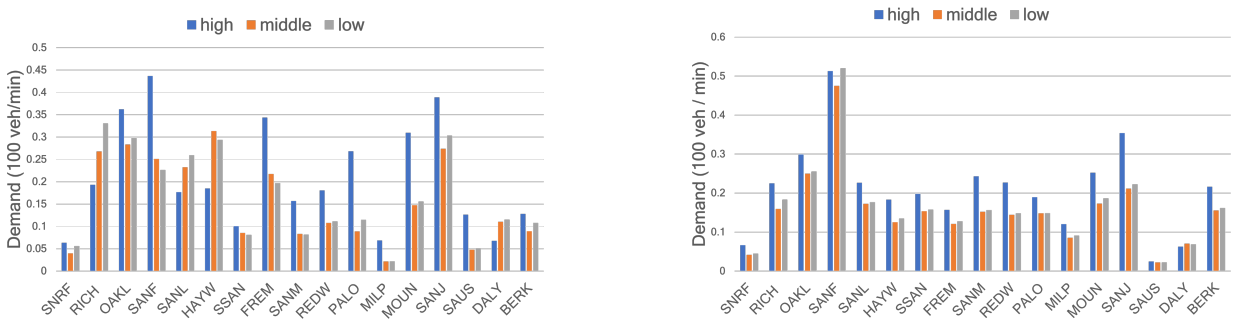
Demand estimate We categorize the driving population into three distinct segments based on their value-of-time, namely *low, middle, and high* value-of-time. The determination

[6]This dataset was available for public use at `https://www.safegraph.com` till 2021 and is now commercially available



Node	Abbr
San Rafael	SRFL
Richmond	RICH
Oakland	OAKL
San Francisco	SFRN
San Leandro	SLND
Hayward	HAYW
South SF	SSFO
Fremont	FREM
San Mateo	SANM
Redwood City	REDW
Palo Alto	PALO
Milpitas	MILP
Mountain View	MTNV
San Jose	SANJ
Sausalito	SAUS
Daly City	DALY
Berkeley	BERK

Figure 9.3: Bay Area transportation network with tolled bridge segments. Different colors on the map represent the boundaries of cities. The table contains the names of the nodes in map along with abbreviations.



(a) Distribution of traveler types based on their origin (home) nodes

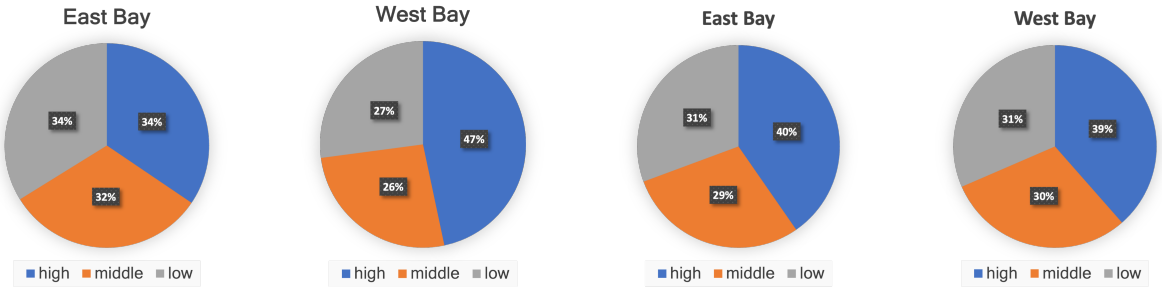
(b) Distribution of traveler types based on their destination (work) nodes.

Figure 9.4: Distribution of origin and destination traveler demands.

of the fraction of driving population in each of these categories relies on the *Means of Transportation* dataset from ACS. Specifically, we assign a traveler to the (a) low value-of-time category if their annual individual income is less than $\$25,000$, to the (b) middle value-of-time category if their annual individual income falls within the range of $\$25,000$ to $\$65,000$, and to the (c) high value-of-time category if their annual individual income exceeds $\$65,000$.

Figure 9.4 provides a visual representation of the distribution of traveler demand to and from each node in the network, stratified by value-of-time. Note that this demand specifically pertains to inter-node travel, with within-node demand excluded from the analysis. We find that approximately 40% of travelers are high willingess-to-pay, and 30% of travelers are of middle and low value-of-time, each. In Figure 9.4a (resp. Figure 9.4b), we present the distribution of traveler demand based on their home (resp. work) location. Around 55% of traffic emerges from relatively few nodes on the East Bay such as RICH, OAKL, SLND, HAYW, FREM, SANJ. Moreover, around 40% of traffic has a work destination in one of the four nodes SFRN, PALO, MTNV, and OAKL. Notably, there exists substantial heterogeneity in both the home and work locations of different traveler types, as can be observed by comparing the distribution of demands in Figure 9.4 to the distribution of median income found in Figure 9.1. For instance, nodes such as RICH, HAYW, SLND, and DALY are predominantly inhabited by a higher number of low value-of-time travelers, while nodes such as PALO, OAKL, SFRN, FREM, and SAUS are predominantly inhabited by high value-of-time travelers. It is interesting to note that on most of the nodes the demographics of incoming traffic predominantly comprise high value-of-time travelers. Additionally, as can be seen in Figure 9.5, high-income travelers make up a large fraction demand that originates in the West Bay, as well as of the work location demand on both the East and West Bay.

Next, we describe the approach used to compute the daily demand of different types of travelers traveling between different o-d pairs during January 2019-June 2019. There are three main steps to our approach: first, we obtain an estimate of the relative demand of travelers traveling between different zip-codes in the Bay Area by using the Safegraph



(a) Home Location (b) Home Location (c) Work Location (d) Work Location

Figure 9.5: Distribution of home and work demands across high, middle, and low income levels aggregated over the two sides of the bay area.

dataset. Particularly, for every month, the Neighborhood Patterns data in the Safegraph dataset provides the average daily count of mobile devices that travel between different census block groups (CBGs) during the work day, which is then aggregated to obtain the relative demand of travelers traveling between different zip codes. After accounting for the sampling bias induced due to the randomly sampled population across the United States, we calibrate demands by using the ACS dataset which provides the income-stratified driving population in every zip code. Finally, to obtain an estimate of daily variability in demand we further augment the demand data with the PeMS dataset by adjusting for daily variation in the total flow on the network in every month. The details of demand estimation are included in [269].

Calibrating the edge latency functions We calibrate the latency functions of each edge of the Bay Area freeway network shown in Figure 9.3. We adopt the Bureau of Public Roads (BPR) function proposed by the Federal Highway Administration (FHA) [282],defined as $\ell_e(w_e) = a_e + b_e w_e^4$, for every $e \in E$, where a_e represents the free-flow travel time (i.e. latency with zero flow) of edge e and b_e is the slope of congestion.

We compute the average driving time of each edge during the morning rush hour (6am to 12pm) on each workday from January 1, 2019 to June 30, 2019 using the speed and distance data from the PeMS dataset. We denote the set of all days as $\mathcal{T}$, the travel time and traffic flow of each edge $e \in E$ on day $t \in \mathcal{T}$ as $\hat{\ell}_e^t$ and $\hat{w}_e^t$, respectively. The details of computing $\left(\hat{\ell}_e^t, \hat{w}_e^t\right)_{t\in\mathcal{T}}$ are provided in [269]. We estimate the free-flow travel time a_e of each $e \in E$ using the average travel time of edge e computed from the PeMS dataset at 3am, when the traffic flow is approaching zero. We denote the estimated value of a_e as $\hat{a}_e$ for each $e \in E$. We next estimate the slope b_e of each edge $e \in E$ using an ordinary least squares regression. In particular, the estimate $\hat{b}_e$ is solved as the minimizer of the following convex program: for every $e \in E$,

$$\hat{b}_e = \arg\min_{b_e \in \mathbb{R}} \sum_{t\in\mathcal{T}} \|\hat{\ell}_e^t - \hat{a}_e - b_e \cdot (\hat{w}_e^t)^4\|^2.$$

Estimating the value-of-time parameters

We formulate the problem of estimating the value-of-time parameters as an inverse optimization problem. Specifically, the optimal estimate of value-of-time parameters corresponding to the three types of travelers, $\theta^{H*}, \theta^{M*}, \theta^{L*}$, are the ones that minimize the difference between the observed flows on each edge of the network and the corresponding equilibrium edge flows. That is,

$$\begin{aligned} \theta_H^*, \theta_M^*, \theta_L^* = \underset{\theta^H, \theta^M, \theta^L}{\arg\min} \quad & \sum_{t \in \mathcal{T}} \sum_{e \in E} (\hat{w}_e^t - w_e(q^t))^2 \\ \text{s.t.} \quad & q^t \in \underset{q \in \mathcal{Q}(D^t)}{\arg\min}\, \Phi(q, p, \theta) \quad \forall t \in \mathcal{T}, && \text{(9.14a)} \\ & w_e(q^t) \text{ is given by (9.3)}, && \text{(9.14b)} \\ & \mathcal{Q}(D^t) \text{ is given by (9.1)}, && \text{(9.14c)} \end{aligned}$$

where p is the toll price vector in 2019 (i.e. \$7 on each bridge, and \$0 for the remaining edges), $\hat{w}_e^t$ is the observed edge flow on each edge $e \in E$ and each day $t \in \mathcal{T}$ computed using the PeMS dataset, and D^t is the estimated demand vector of each day t computed using the ACS and Safegraph datasets.

Directly solving (9.14) is challenging due to the non-linearity of the edge latency function and the potential function in (9.14a). We compute the estimates using grid search: we construct a grid of value-of-time, where the granularity of each of $\theta^H, \theta^M, \theta^L$ is \$5 per hour. We also assume that the maximum value of value-of-time is \$100 per hour and the minimum is \$0 per hour. Therefore, we define the set of all possible parameter values as $\Theta := \{0, 5, 10, 15, \ldots, 100\}^3$. For each $\theta = (\theta^H, \theta^M, \theta^L) \in \Theta$, we compute the equilibrium flow q_t for every $t \in \mathcal{T}$ and compute the total squared error as in the objective function of (9.14). The optimal parameter θ^* is the one that minimizes the total squared error. We obtain:

$$\theta^* = (\theta^{L*}, \theta^{M*}, \theta^{H*}) = (\$10/\text{hour},\ \$30/\text{hour},\ \$70/\text{hour}).$$

Our estimate θ^* is consistent with the observations reported in prior works, which show that the value-of-time values typically lie between $60\% - 100\%$ of the average hourly income of the population ([327, 19, 299]).

Furthermore, as a robustness check, we plot the equilibrium edge flow $w_e(q^{t*})$ and observed edge flow $\hat{w}_e^t$ for every $e \in E, t \in \mathcal{T}$ in Figure 9.6. Each dot in this figure represents the flow on an edge $e \in E$ on a single day $t \in \mathcal{T}$. Overall, the dots are distributed along the diagonal of the plot indicating that the our computed equilibrium edge flow are relatively consistent with the observed edge flow subject to noise in time costs and demand fluctuations.

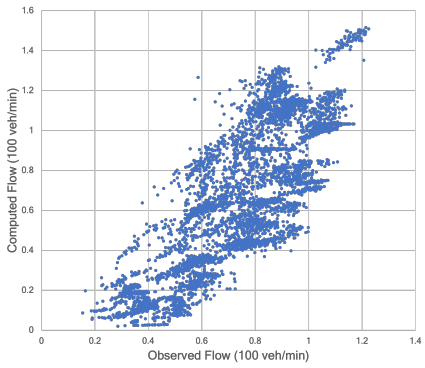


Figure 9.6: Observed and computed equilibrium edge flow.

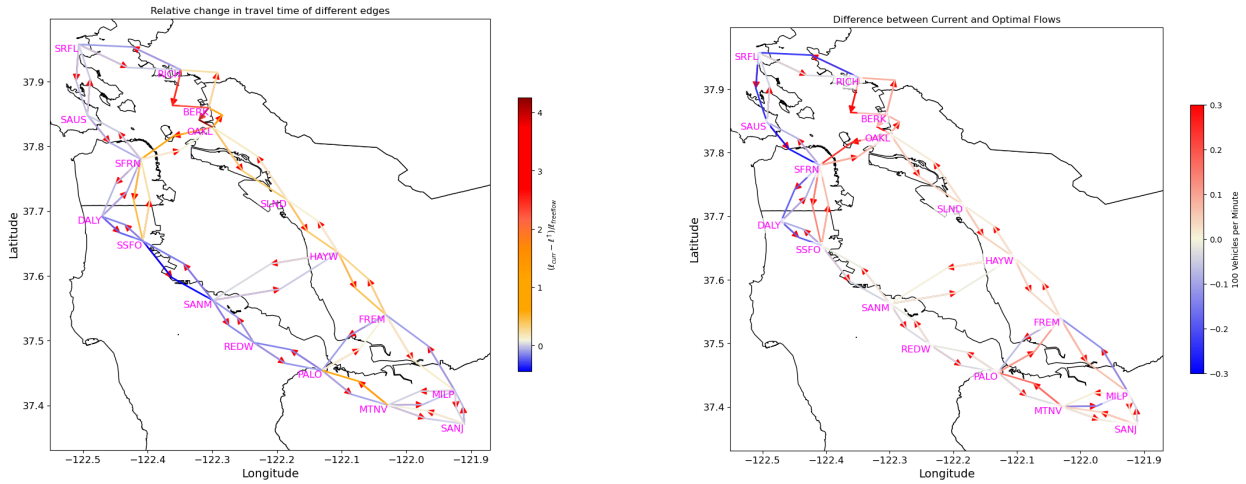
9.4 Efficiency and Equity Analysis of Congestion Pricing schemes

Our goal in this section is three fold. First, we analyze the congestion levels induced at equilibrium due to current congestion pricing scheme, curr, and identify corridors in the Bay Area which are congested. Next, using the computational method introduced in Section 9.2 and the calibrated model of San Francisco Bay area freeway network in Section 9.3, we compute the toll values under the congestion pricing schemes hom, het, hom_sc, and het_sc. Finally, we compare different congestion pricing schemes in terms of efficiency and equity of travel cost, and also in terms of overall revenue generated at equilibrium.

Congestion under the current congestion pricing scheme (curr)

Here, we analyze the congestion levels induced at equilibrium under the current congestion pricing scheme, curr, which imposes a uniform toll of $7 on each of the five bridges in the Bay Area, namely on the Richmond-San Rafael Bridge (RICH-SRFL), San Francisco-Oakland Bay Bridge (OAKL-SFRN), Golden Gate Bridge (SAUS-SFRN), San Mateo-Hayward Bridge (HAYW-SANM), and Dumbarton Bridge (FREM-PALO).

Figure 9.7a depicts the difference between the equilibrium travel time given curr and the congestion minimizing travel time (normalized by free flow travel time on every edge). We observe that edges on the eastern corridor (connecting nodes RICH-BERK-OAKL-SLND-HAYW-FREM) are over-congested. Meanwhile, the edges on the western corridor (connecting nodes SRFL-SAUS-SFRN-DALY-SSFO-SANM-REDW) are relatively less congested. Furthermore, we observe that amongst all bridges the Bay Bridge (OAKL-SFRN) is also most congested, which is consistent with several prior studies [311, 31, 159]. Additionally, Figure 9.7b presents the difference in the edge flows induced at equilibrium with that of socially optimal edge flows. We observe that in order to reduce the overall congestion we need to ensure that



(a) Proportional travel time increase under curr (normalized by free flow travel time).

(b) Difference between equilibrium flow induced by curr and optimal flow.

Figure 9.7: Current congestion pricing scheme

(R1) the travelers using the edges in the corridor RICH-BERK-OAKL-SFRN (resp. SFRN-OAKL-BERK-RICH) are incentivized to use the edges in the corridor RICH-SRFL-SAUS-SFRN (resp. SFRN-SAUS-SRFL-RICH).

(R2) the travelers using the edges in the corridor SFRN-SSFO are incentivized to use the corridor SFRN-DALY-SSFO.

(R3) the travelers using the eastern corridor MILP-FREM-HAYW-SLND-OAKL are diverted to use the western corridor MTNV-PALO-REDW-SANM-SSFO by suitably incentivizing them to use the Dumbarton Bridge or the San Mateo-Hayward Bridge.

Furthermore, we note that the average travel cost[7] (the sum of the travel time cost and the equivalent time cost of the monetary expense as in (9.4)) experienced by different types of travelers at equilibrium is unequal in curr. Specifically, low value-of-time travelers bear the travel cost of approximately 91 minutes, while high and middle value-of-time travelers face costs of 61 and 68 minutes, respectively. Moreover, as indicated in Table 9.4.1, this unequal distribution of travel time persists not only on average but also when examined across different threshold levels of travel cost.

To summarize, we observe that the current congestion pricing scheme implemented the Bay area does not result in efficient allocation of traffic on the network. Additionally, it also leads to unequal distribution of travel cost across different types of travelers.

[7]It can be shown that the average travel cost experienced by travelers is independent of the route flows on the network and is only dependent on the equilibrium edge flows, which are unique as shown in Proposition 9.2.1.

Travel Cost	Low (%)	Middle (%)	High (%)
$\geqslant$ 60 minutes	69	55	46
$\geqslant$ 90 minutes	51	31	28
$\geqslant$ 120 minutes	32	13	12
$\geqslant$ 150 minutes	17	1	1

Table 9.4.1: Fraction of low, middle and high value-of-time travelers that incur total cost (in minutes) more than stated threshold at equilibrium.

Toll values under different congestion pricing schemes

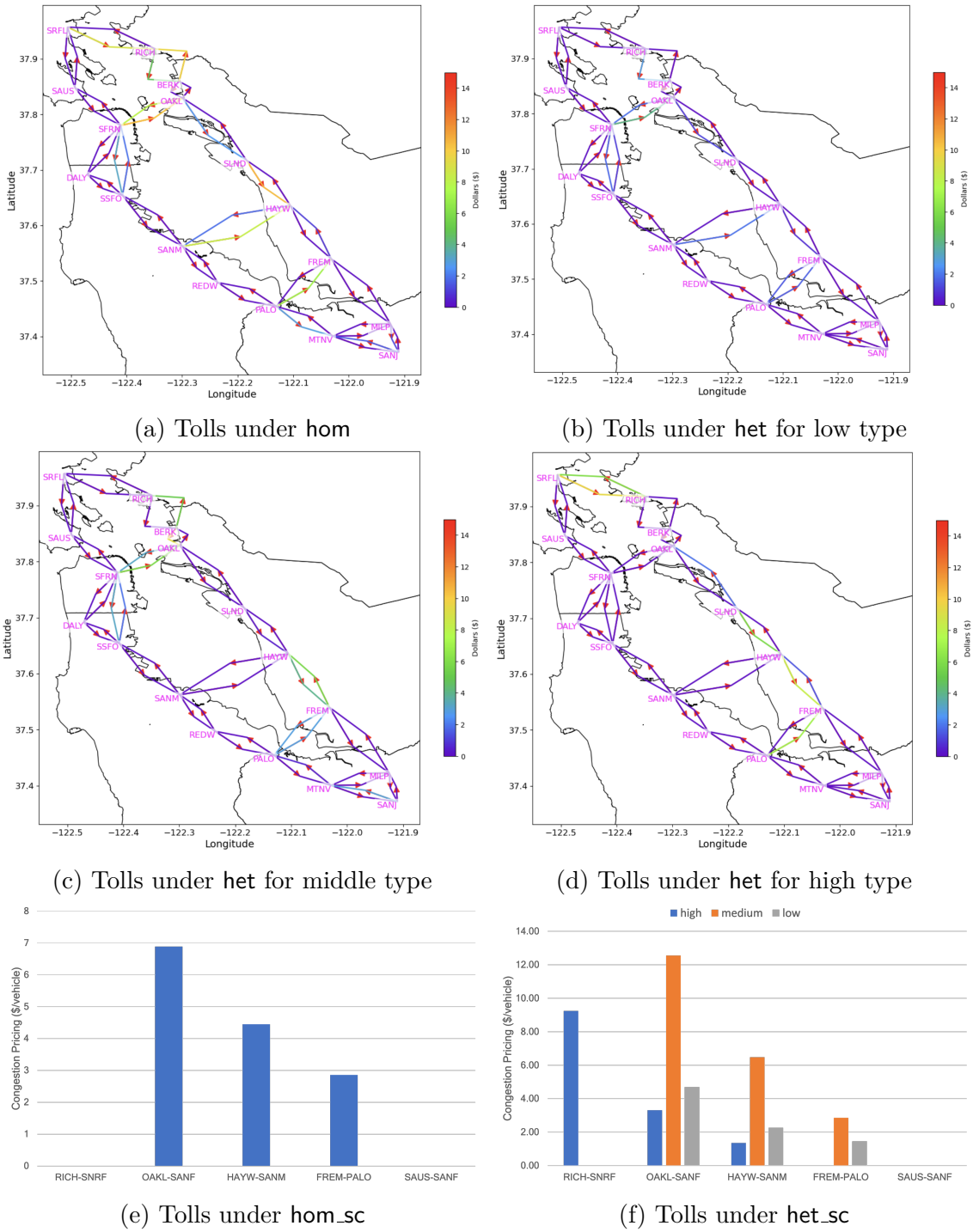
Here, using the calibrated model of the Bay area obtained in Section 9.3, we present the computed values of tolls on various edges of the Bay area network under different congestion pricing schemes (namely, hom, het, hom_sc, het_sc) obtained using the computational methodology presented in Section 9.2.

Figure 9.8a presents the toll values computed under hom by solving ($\mathcal{P}^*_{\mathsf{hom}}$). Figures 9.8b-9.8d present the toll values for low, middle, and high value-of-time travelers under het by solving ($\mathcal{P}^*_{\mathsf{het}}$). Figure 9.8e presents the toll values computed under hom_sc by solving ($\mathcal{P}^*_{\mathsf{hom\_sc}}$). Figure 9.8f further presents the toll values for low, middle, and high value-of-time travelers under het_sc by solving ($\mathcal{P}^*_{\mathsf{het\_sc}}$). To compute all of these toll values, we choose $\lambda = 20$ in ($\mathcal{P}^*_{\mathsf{hom}}$), ($\mathcal{P}^*_{\mathsf{het}}$), ($\mathcal{P}^*_{\mathsf{hom\_sc}}$), and ($\mathcal{P}^*_{\mathsf{het\_sc}}$). This choice of parameter λ ensures that the numerical value of the average welfare metric and the equity metric in these optimization problems are of the same order of magnitude.

Note that in hom and het, on all the bridges, tolls in the east-to-west direction are lower than tolls in the west-to-east direction. This is in contrast to curr, where the west-to-east direction is not tolled at all on any bridge and only the east-to-west direction is tolled at a flat rate of \$7 (refer Figure 9.3). Given that the western corridor is less congested than the eastern corridor in curr (refer Figure 9.7a), such tolling is useful to efficiently redistribute traffic in the network. Furthermore, note that in all of the congestion pricing schemes we compute, unlike curr, the Golden Gate Bridge (SAUS-SFRN) is not tolled at all. This choice ensures that more travelers in the eastern corridor, particularly in nodes such as RICH and BERK are able to reach nodes in the west, particularly SFRN, through Golden-Gate bridge instead of Bay-bridge (OAKL-SFRN).

Discussion on efficiency, equity and revenue generation

In this subsection, we compare the effectiveness of curr, hom, hom_sc, het and het_sc in terms of efficiency (the average travel time per traveler), equity (average increase in travel cost in comparison to no tolls), and revenue generation (the total toll revenue generated by these



(a) Tolls under hom

(b) Tolls under het for low type

(c) Tolls under het for middle type

(d) Tolls under het for high type

(e) Tolls under hom_sc

(f) Tolls under het_sc

Figure 9.8: Toll values under congestion pricing schemes hom, het, hom_sc, and het_sc.

schemes). Additionally, we also compare these pricing schemes with the scenario when no toll is implemented (denoted zero).

Efficiency and Equity Considerations.

Figure 9.9 represents the average travel time experienced by travelers under different congestion pricing schemes.

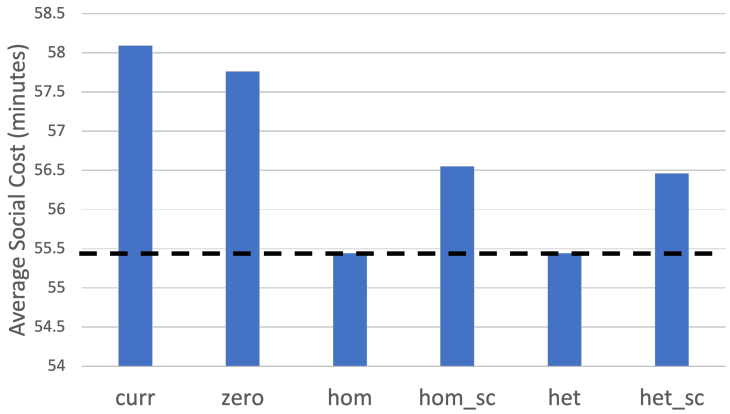


Figure 9.9: Comparison of average social cost per traveler for curr, zero, hom, hom_sc, het, and het_sc. Here, the dashed line represent congestion minimizing cost computed by solving (9.5).

As expected from Proposition 9.2.2, the congestion pricing schemes hom and het achieves the minimum congestion levels on the network. Additionally, we note that hom_sc and het_sc do not achieve the minimum congestion level due to the support constraints. Furthermore, it's noteworthy that het_sc results in a slightly improved average travel time compared to hom_sc. This improvement can be attributed to the flexibility of heterogeneous pricing schemes, which allow for type-specific tolls.

From Figure 9.9, we observe that the Price of Anarchy (PoA) – which is the ratio of the social cost of equilibrium congestion levels induced under no tolls with that of opt– is 1.04 for the Bay area transportation network. This is likely due to the high congestion level of the network during the morning rush hour. Indeed, theoretical studies [102, 103] have proved that the PoA approaches to 1 as the total demand of travelers increases. Moreover, empirical studies [441, 321, 307] have also shown that the PoA in the transportation networks of London, Boston, New York city and Singapore are also close to 1. Furthermore, from Figure 9.9, we find that all congestion pricing schemes hom, hom_sc, het, het_sc outperform curr in terms of the average travel time. Surprisingly, it is also marginally outperformed by zero. A key reason is that curr imposes the same tolls on all of the bridges which does not result in effective re-distribution of traffic from eastern corridor to western corridor. While a reduced toll price or zero toll price may increase the total demand of travelers, but its impact

is likely to be not significant due to (1) the high expense of car ownership and parking fee ([119] estimates that US average annual car ownership cost is $12182 in 2023), and (2) the low coverage of public transportation in the Bay Area.

Figure 9.10 illustrates the average travel cost experienced by type of travelers under different pricing schemes. We observe that the difference of average cost across the three traveler types is lower in het, het_sc, hom_sc, and zero, in comparison to curr. Moreover, we observe that for all type of travelers, the average travel cost is lower in het, het_sc, hom_sc, and zero, in comparison to curr. Furthermore, we note that this observation not only holds in the averaged sense but also in a distributional sense as illustrated in Table 9.4.2-9.4.4, which presents the proportion of travelers of a particular type experiencing travel costs surpassing a predetermined threshold. We observe that, regardless of the value of threshold and the type of travelers, the proportion of travelers experiencing cost higher than a threshold is higher in curr in comparison to het, het_sc, hom_sc, and zero. This clearly shows that curr is not preferred by any type of traveler. The pricing scheme hom results in higher travel cost because it cannot differentiate between type of traveler and charges higher tolls to travelers in order to ensure minimum average travel time.

In homogeneous congestion pricing schemes, regardless of the threshold and the type of traveler, a higher percentage of travelers incur travel costs exceeding a set threshold compared to heterogeneous pricing. This is due to type-specific tolls in heterogeneous schemes resulting in lower tolls for low income travelers. Additionally, pricing schemes with support constraints reduce the percentage of travelers exceeding a threshold. While the differences are marginal between het and het_sc, such differences are more prominent between hom and hom_sc.

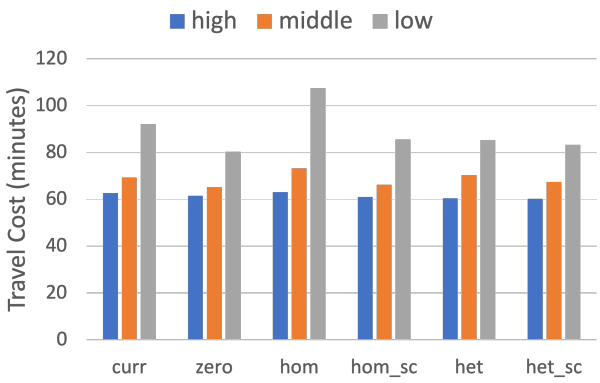


Figure 9.10: Average travel cost experienced by different types of travelers under different tolling schemes.

Next, in Figure 9.11, we compare different pricing schemes using two metrics: average travel time and the equity metric (as defined in (9.8)-(i)). Our results show that all pricing schemes, except for hom , outperform curr on both metrics. Additionally, we present a Pareto front (dotted line) that illustrates the trade-off between minimizing average travel

Travel Cost	curr	zero	het_sc	hom_sc	hom	het
$\geqslant$ 60 minutes	69%	56%	64%	67%	76%	66%
$\geqslant$ 90 minutes	51%	39%	42%	43%	55%	42%
$\geqslant$ 120 minutes	32%	22%	25%	27%	41%	25%
$\geqslant$ 150 minutes	17%	10%	11%	12%	26%	13%

Table 9.4.2: low value-of-time travelers

Travel Cost	curr	zero	het_sc	hom_sc	hom	het
$\geqslant$ 60 minutes	55%	49%	50%	50%	54%	52%
$\geqslant$ 90 minutes	31%	28%	31%	30%	35%	30%
$\geqslant$ 120 minutes	13%	12%	11%	11%	17%	14%
$\geqslant$ 150 minutes	1%	1%	1%	1%	3%	2%

Table 9.4.3: middle value-of-time travelers

Travel Cost	curr	zero	het_sc	hom_sc	hom	het
$\geqslant$ 60 minutes	46%	46%	46%	46%	48%	46%
$\geqslant$ 90 minutes	28%	27%	26%	27%	30%	27%
$\geqslant$ 120 minutes	12%	7%	7%	8%	10%	7%
$\geqslant$ 150 minutes	1%	0%	0%	0%	1%	1%

Table 9.4.4: high value-of-time travelers

time and reducing inequity. The method used to compute this trade-off curve is detailed in the Appendix.

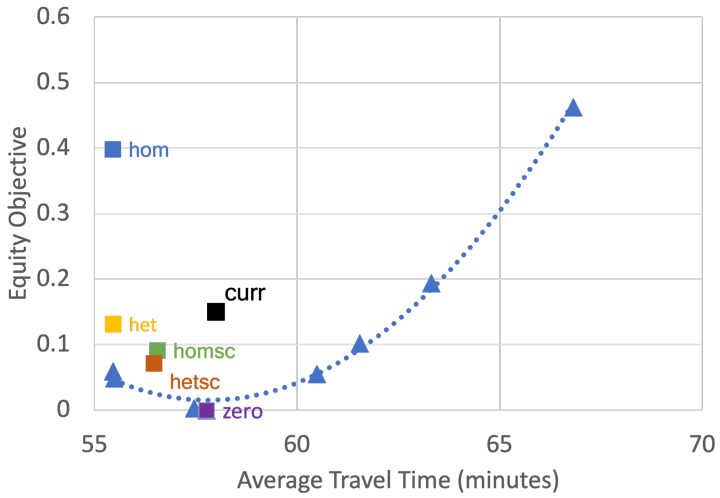


Figure 9.11: Trade-off between average travel time and equity: The blue triangles represent different pricing schemes, positioned near the Pareto curve (a polynomial best-fit curve through the triangle points), based on computations detailed in the Appendix.

Revenue considerations

Another important aspect of determining the congestion pricing scheme is the revenue it generates, which could be used for maintenance of existing transportation infrastructure, enhancing public transit options, amongst other things. Figure 9.12 presents a comparison of different congestion pricing scheme in terms of total revenue. As per the data released by Metropolitan Transportation Commission (MTC) [8] a total toll revenue of $\$633,932,206$ was collected in the Bay Area during the year 2019-2020. Our calibrated model in curr predicts toll revenues on the same order of magnitude but slightly lower than MTC data. The mismatch between our prediction and MTC data is attributed to the fact that *(i)* our analysis only focuses on morning rush hour but MTC data also include tolls collected beyond morning rush hour as well, *(ii)* MTC data also includes tolls on HOV (High Occupancy Vehicle) lanes which are currently not added in our analysis, *(iii)* there is some additional demand incoming from other nearby cities not included in our analysis, and *(iv)* higher tolls are charged to multi-axle vehicles, with tolls charged as high as \$36 in 2019.[9]

[8] available at `https://mtc.ca.gov/about-mtc/authorities/bay-area-toll-authority/historic-toll-paid-vehicle-counts-toll-revenue`

[9] refer `http://tinyurl.com/MTC-Multi-Axle`)

Notably, hom generates the highest revenue as it applies uniformly higher prices across all edges, irrespective of traveler types, with the goal of achieving a minimum congestion congestion level. Moreover, the revenue of the other three pricing schemes hom_sc, het and het_sc are comparable to that of curr with het being slightly higher and hom_sc and het_sc being slightly lower.

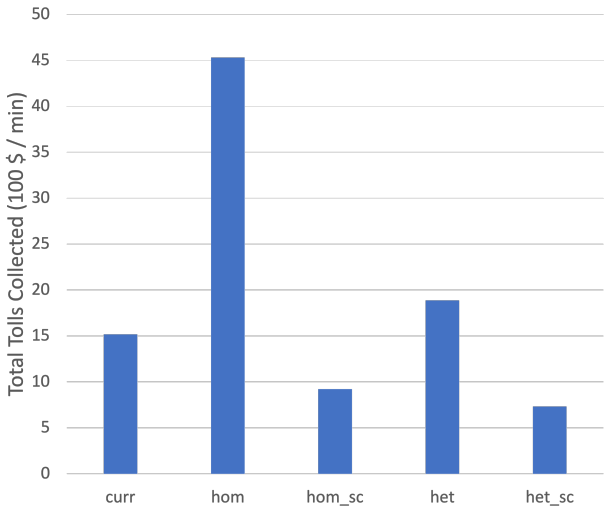


Figure 9.12: Comparison of total revenue collected for current,hom,hom-c,het,het-c.

9.5 Concluding Remarks

We study the problem of designing congestion pricing schemes which not only minimize the overall congestion but also reduce the disparate impact of congestion pricing schemes on the basis of socioeconomic and geographic diversity of travelers. We present a multi-step linear programming based approach to design four kinds of congestion pricing schemes varying in terms of their implementation depending on whether (a) they can toll travelers on the basis of their willingness-to-pay, and (b) they can toll every edge of the network or only a subset of it. The evaluation and comparison of these congestion pricing schemes on the San Francisco Bay Area highway network reveal several significant insights. The proposed schemes outperform the currently implemented scheme in terms of overall congestion reduction and exhibit improvements in equity by providing better travel costs to each type of traveler. The analysis also highlights the revenue generation potential of different pricing schemes. Furthermore, heterogeneous pricing schemes can yield more equitable distribution of travel cost between different types of travelers, paving the way for future research to explore effective implementation strategies.

Chapter 10

Data-driven Method for Distributionally Robust Strategic Classification

Real-world deployment of machine learning models often triggers feedback effects, where the model influences user behavior, which in turn alters the data distribution. In high-stakes domains such as credit scoring, hiring, or spam detection, individuals may strategically modify their features to receive more favorable predictions. Strategic classification is an emerging paradigm that attempts to close this loop—designing classifiers that account for such reactive behavior during training. Ignoring this feedback loop by deploying a naïvely trained model can result in substantial performance degradation or even catastrophic failure.

Modeling these interactions poses significant challenges: the learner typically lacks access to agents' private objectives or preferences and thus cannot observe or encode their best-response functions directly. To address this, we adopt a natural model of agent behavior introduced by [124], where agents make optimal manipulations subject to individual costs. This model captures key aspects of strategic behavior, but still suffers from a major limitation: it assumes perfect knowledge of how agents respond. In practice, this assumption rarely holds.

Prior work has studied this setting through the lens of risk minimization with decision-dependent data distributions, identifying conditions under which the learner can optimize performance [300]. However, these analyses generally do not account for model misspecification. When the assumed agent response model deviates from reality, the performance of standard learning approaches may degrade significantly—an issue we explore empirically in this chapter.

To address this, we propose a data-driven approach that ensures robustness to such misspecification by adopting a distributionally robust optimization (DRO) framework. Specifically, we place an ambiguity set over possible, strategically influenced data distributions and optimize for the worst-case risk over this set. We show that, under mild assumptions, this DRO formulation of the decision-dependent learning problem, which is an infinite dimen-

sional problem, is equivalent to a finite dimensional convex-concave min-max optimization problem.

To solve the resulting min-max problem, we develop a novel zeroth-order (gradient-free) algorithm. This method is especially well-suited to settings with strategic agents, where the decision-maker cannot observe or differentiate through the data-generating process. This is because the decision maker may not know the exact preferences of users. Our algorithm relies solely on function evaluations, enabling robust learning even when gradients of the loss with respect to strategic responses are inaccessible or undefined.

Contributions. This chapter makes the following contributions:

1. **Problem Formulation.** We formulate the Wasserstein distributionally robust strategic classification problem as a constrained, finite-dimensional, smooth convex-concave min-max optimization problem:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} L(x,y), \tag{10.1}$$

   where $L(x,y) = \frac{1}{n}\sum_{i=1}^{n} L_i(x,y)$ is a finite-sum loss, and $\mathcal{X}, \mathcal{Y}$ are compact sets. This formulation extends classical strategic classification to account for uncertainty in the data-generating process via a distributional ambiguity set.

2. **Algorithmic Development.** We propose a *zeroth-order randomized algorithm* for solving this min-max problem that operates without assuming strong convexity or concavity. The algorithm is well-suited to settings where gradients are unavailable due to unobserved agent response mechanisms.

3. **Theoretical Guarantees.** We provide the *first non-asymptotic convergence analysis* of the Optimistic Gradient Descent Ascent (OGDA) method under *random reshuffling and zeroth-order oracle access* for general convex-concave min-max problems.

Together, these results establish a robust and practical framework for learning in strategic environments under model mis-specification.

Related Work

Our work builds upon and integrates insights from four key research areas: distributionally robust optimization, strategic classification and performative prediction, zeroth-order methods for min-max optimization. Below, we summarize the most relevant developments in each.

Distributionally Robust Optimization (DRO). DRO aims to train models that are robust to distributional shifts between training and deployment. These shifts may result from sample selection bias, class imbalance, or adversarial perturbations [79, 266]. A common approach is to formulate the learning problem as a min-max optimization, where the learner minimizes loss under the worst-case distribution within an ambiguity set, typically defined using f-divergence or Wasserstein distance [24, 46, 312, 41, 377, 442, 186]. While these frameworks address robustness to arbitrary or adversarial noise, they generally do not capture settings in which the data distribution shifts in response to the learner's decision rule—a key challenge in strategic environments.

Strategic Classification and Performative Prediction. Strategic classification [173, 124, 373] and performative prediction [336, 300, 126] study learning settings where the data-generating process is influenced by the deployed model. Such feedback loops occur when users respond strategically to deployed decision rules—for example, when applicants alter their features to receive a favorable credit decision, or when bank clients withdraw funds in response to perceived instability. In these models, the learner observes only the strategically altered (best-response) features and lacks access to the original data-generating process [124]. While this literature has introduced frameworks for learning in such interactive settings, it has largely ignored robustness to misspecification in the assumed response behavior—a gap that our work addresses.

Zeroth-Order Methods for Min-Max Optimization. Zeroth-order (gradient-free) methods are attractive in applications where gradient information is unavailable or expensive to compute, such as black-box adversarial attacks [91, 187, 409]. Recent work has established non-asymptotic convergence guarantees for such methods in convex optimization and strongly-convex/concave min-max settings [262, 152, 419, 157, 316]. However, these guarantees typically rely on strong convexity or concavity assumptions that are often violated in practice. In contrast, our work provides the first convergence analysis for a zeroth-order algorithm under the more general and practically relevant convex-concave setting.

10.1 Primer on Distributionally Robust Generalized Linear Problem

Consider a *generalized linear problem*, which generalizes many classification tasks, where the goal is to estimate the parameter $\theta \in \Theta$, with Θ assumed to be a compact set, by solving the following convex optimization problem:

$$\inf_{\theta\in\Theta} \mathbb{E}_{(\bar{x},\bar{y})\sim\mathcal{D}} \left[\phi\Big(\langle \bar{x}, \theta\rangle\Big) - \bar{y}\langle \bar{x}, \theta\rangle\right], \tag{10.2}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a smooth convex function, and $(\bar{x}, \bar{y}) \in \mathbb{R}^d \times \{-1, +1\}$ is a random feature-label pair drawn from an unknown distribution $\mathcal{D}$, which is typically approximated

by the empirical distribution of observed data. The generalized linear model framework captures a wide range of classical and modern machine learning models [292].

A *distributionally robust generalized linear problem* extends this formulation by minimizing the worst-case expected loss over an ambiguity set $\mathcal{P}$ of probability distributions close to the nominal data distribution. Before introducing this robust formulation, we recall the definition of the Wasserstein metric, which we will use to define the ambiguity set.

Definition 10.1.1 (**Wasserstein Distance between Distributions on $\mathcal{Z}$ with Cost Function** c)**.** *Let μ, ν be probability distributions over $\mathcal{Z} := \mathbb{R}^d \times \{+1, -1\}$ with finite second moments. Denote by $\Pi(\mu, \nu)$ the set of all couplings (joint distributions) with marginals μ and ν. Given a metric $c : \mathcal{Z} \times \mathcal{Z} \to [0, \infty)$, the Wasserstein distance is defined as*

$$\mathcal{W}_c(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{(Z,Z') \sim \pi} \Big[c(Z, Z') \Big].$$

Assumption 10.1.1. *We use the cost function*

$$c(z, z') := \|x - x'\|_2^2 + \kappa \cdot |y - y'|,$$

for $z = (x, y), z' = (x', y') \in \mathcal{Z}$, with a fixed constant $\kappa > 0$.

This setup can be interpreted as a game between a learning algorithm and an adversary. Given the parameter θ chosen by the learning algorithm, the adversary selects a probability measure from the uncertainty set $\mathcal{P}$ to maximize the expected risk for that parameter choice:

$$\inf_{\theta \in \Theta} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{(\bar{x}, \bar{y}) \sim \mathbb{P}} \left[\phi\Big(\langle \bar{x}, \theta \rangle \Big) - \bar{y} \langle \bar{x}, \theta \rangle \right]. \tag{10.3}$$

Here, $(\bar{x}, \bar{y}) \sim \mathbb{P} \in \mathcal{P}$.

Typically, $\mathcal{P}$ is chosen as a Wasserstein ball around the empirical distribution $\tilde{\mathcal{D}}_n$ constructed from a dataset of n independent observations $\{(\tilde{x}_i, \tilde{y}_i) \in \mathbb{R}^d \times \{-1, +1\}\}_{i=1}^n$, sampled from the true data distribution $\mathcal{D}$. For any radius $\delta > 0$, the ambiguity set $\mathcal{P}$ is defined to be a subset of ball in Wasserstein metric

$$\mathbb{B}_\delta(\tilde{\mathcal{D}}_n) := \Big\{ \mathbb{P} : \mathcal{W}_c(\mathbb{P}, \tilde{\mathcal{D}}_n) \leqslant \delta \Big\}. \tag{10.4}$$

We refer the reader to [377] for algorithmic and theoretical properties of this formulation.

A key limitation of the formulations in (10.2) and (10.3) is the assumption that the underlying data distribution $\mathcal{D}$ is fixed and independent of the decision variable θ. In many strategic settings, however, the data distribution depends on the chosen parameter θ. *Decision-dependent supervised learning* aims to address such problems. When specialized to the generalized linear model, this formulation becomes

$$\inf_{\theta \in \Theta} \mathbb{E}_{(\bar{x}, \bar{y}) \sim \mathcal{D}(\theta)} \left[\phi\Big(\langle \bar{x}, \theta \rangle \Big) - \bar{y} \langle \bar{x}, \theta \rangle \right],$$

where the data distribution $\mathcal{D}(\theta)$ explicitly depends on the decision θ.

In this work, we take a step further and consider a special case of the following *distributionally robust decision-dependent generalized linear model*:

$$\inf_{\theta\in\Theta} \sup_{\mathbb{P}\in\mathcal{P}(\theta)} \mathbb{E}_{(\bar{x},\bar{y})\sim\mathbb{P}} \left[\phi\Big(\langle \bar{x},\theta\rangle\Big) - \bar{y}\langle \bar{x},\theta\rangle\right], \tag{10.5}$$

where $\mathcal{P}(\theta) \subseteq \mathbb{B}_\delta(\tilde{\mathcal{D}}_n(\theta))$. The dependence of $\mathbb{P}$ on θ is explicitly captured by the ambiguity set $\mathcal{P}(\theta)$, which is a Wasserstein ball centered at the empirical distribution $\tilde{\mathcal{D}}_n(\theta)$ induced by θ.

To model the decision-dependent distribution shifts $\tilde{\mathcal{D}}_n(\theta)$, we focus on the setting of strategic classification. The following section formalizes the modeling assumptions underlying this framework.

10.2 Model

The distributionally robust decision-dependent learning problem considered in this chapter comprises two main components: the *strategic component*, which captures the distribution shift $\mathcal{D}(\theta)$ induced by the choice of classifier θ; and the *adversarial component*, which models ambiguity in the data distribution via the uncertainty set $\mathcal{P}(\theta)$. We describe each of these components in detail below.

Strategic Component

We denote data points sampled from the *true distribution* by $(\tilde{x}_i, \tilde{y}_i) \sim \mathcal{D}$, where $\mathcal{D}$ is an unknown underlying distribution. For convenience, we associate each data point index i with a distinct agent. Each agent $i \in [n]$ is assumed to act strategically, choosing a reported feature vector in response to the deployed classifier parameter $\theta \in \mathbb{R}^d$. Specifically, given a utility function $u_i(x; \theta, \tilde{x}_i, \tilde{y}_i) \in \mathbb{R}$, the agent selects a response $b_i(\theta, \tilde{x}_i, \tilde{y}_i)$ satisfying:

$$b_i(\theta, \tilde{x}_i, \tilde{y}_i) \in \arg\max_x u_i(x; \theta, \tilde{x}_i, \tilde{y}_i).$$

We allow each agent to have a distinct utility function.

We now impose the following assumptions on these utility functions, which will be crucial for our subsequent analysis.

Assumption 10.2.1. *For each agent $i \in [n]$, the utility function is defined as:*

$$u_i(x; \theta, \tilde{x}_i, \tilde{y}_i) := \frac{1-\tilde{y}_i}{2}\langle x, \theta\rangle - g_i(x - \tilde{x}_i), \tag{10.6}$$

where $g_i : \mathbb{R}^d \to \mathbb{R}$ satisfies the following properties:

1. *$g_i(x) > 0$ for all $x \neq 0$;*

2. *g_i is convex on $\mathbb{R}^d$;*
3. *g_i is positively homogeneous*[1] *of degree $p > 1$;*
4. *The convex conjugate $g_i^*(\theta) \coloneqq \sup_{x\in\mathbb{R}^d} \{\langle x, \theta\rangle - g_i(x)\}$ is G_i-Lipschitz and $\bar{G}_i$-smooth on Θ.*

A direct consequence of this setup is that for any agent i, if $\tilde{y}_i = +1$, then the agent reports truthfully: $b_i(\theta, \tilde{x}_i, +1) = \tilde{x}_i$. Strategic behavior arises only when $\tilde{y}_i = -1$, as only such agents benefit from misreporting. This asymmetry reflects many real-world scenarios—for example, in algorithmic loan approval systems, where only applicants at risk of rejection (i.e., with negative labels) have an incentive to strategically modify their features.

A broad class of functions satisfies Assumption 10.2.1. For instance, for any norm $\|\cdot\|$ and any $p > 1$, the function $g(x) = \frac{1}{p}\|x\|^p$ is a valid choice [124].

The following lemma, which plays a key role in our analysis, characterizes the structure of agents' best responses.

Lemma 10.2.1 ([124])**.** *Under Assumption 10.2.1, for each agent $i \in [n]$, the set of best responses* $\arg\max_x u_i(x; \theta, \tilde{x}_i, \tilde{y}_i)$ *is nonempty, finite, and bounded. Furthermore, the function $\theta \mapsto \langle b_i(\theta, \tilde{x}_i, \tilde{y}_i), \theta\rangle$ is convex. Specifically, for all $i \in [n]$,*

$$\langle b_i(\theta, \tilde{x}_i, \tilde{y}_i), \theta\rangle = \langle \tilde{x}_i, \theta\rangle + \frac{1 - \tilde{y}_i}{2} q g_i^*(\theta),$$

where $q > 1$ is such that $\frac{1}{p} + \frac{1}{q} = 1$.

Adversarial Component

In this subsection, we formally define our model for the adversary and the uncertainty set of distributions resulting from both strategic and adversarial perturbations of the data.

Following the standard formulation of distributionally robust optimization, we restrict $\mathcal{P}(\theta)$ to be a Wasserstein neighborhood of $\tilde{\mathcal{D}}_n(\theta)$, the empirical distribution of the strategically manipulated dataset $\{(b_i(\theta), y_i)\}_{i=1}^n$. That is,

$$\mathcal{P}(\theta) \subseteq \mathbb{B}_\delta\left(\tilde{\mathcal{D}}_n(\theta)\right)$$

for some $\delta > 0$. However, to ensure that the resulting min-max formulation of the WDRSC problem is convex-concave, we impose an additional restriction on the adversary: it may modify the *feature* $b_i(\theta)$ of any data point i, but may modify the *label* y_i only if $y_i = +1$. That is, the adversary cannot flip negative labels to positive.

This imposes a constraint on the conditional distribution $\mathbb{P}_\theta^i$ of (dx, y) generated by the adversary for each sample i:

$$\mathbb{P}_\theta^i(dx, +1 \mid b_i(\theta), -1) = 0 \quad \forall i \in [n].$$

[1] A function $f : \mathbb{R}^d \to \mathbb{R}$ is *positively homogeneous of degree r* if for all $\alpha > 0$ and $x \in \mathbb{R}^d$, $f(\alpha x) = \alpha^r f(x)$.

By the definition of empirical distributions, any $\mathbb{P} \in \mathcal{P}(\theta)$ must be a mixture of these conditionals:

$$\mathbb{P}(dx, y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}_\theta^i(dx, y \mid b_i(\theta), \tilde{y}_i).$$

We now formalize the above constraint.

Assumption 10.2.2. *We assume that for all $i \in [n]$, $\mathbb{P}_\theta^i(dx, +1 \mid b_i(\theta), -1) = 0$, and $\mathbb{P} \in \mathbb{B}_\delta(\tilde{\mathcal{D}}_n(\theta))$. Thus, the uncertainty set is given by:*

$$\begin{aligned} &\mathcal{P}(\theta) \\ &= \mathbb{B}_\delta(\tilde{\mathcal{D}}_n(\theta)) \cap \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}_\theta^i(dx, y \mid b_i(\theta), \tilde{y}_i) \,\middle|\, \mathbb{P}_\theta^i(dx, +1 \mid b_i(\theta), -1) = 0 \text{ for all } i \in [n] \right\}. \end{aligned} \tag{10.7}$$

10.3 Reformulation to Finite Dimensional Convex-Concave Min-max Optimization

In this section, we present the main result that reformulates the WDRSC problem (10.5) into a finite-dimensional min-max optimization problem, under the modeling assumptions described in Section 10.2.

Theorem 10.3.1. *Suppose Assumptions 10.1.1, 10.2.1, and 10.2.2 hold. Additionally, let $\phi : \mathbb{R} \to \mathbb{R}$ be convex, β-smooth, and assume the function $x \mapsto \phi(x) + x$ is non-decreasing. Then, the WDRSC problem* (10.5) *admits the following convex-concave min-max reformulation:*

$$\begin{aligned} \min_{(\theta,\alpha)} \max_{\gamma \in \mathbb{R}^n} \Bigg\{ &\alpha(\delta - \kappa) + \frac{1}{n} \sum_{i=1}^{n} \frac{1 + y_i}{2} \left[\phi\left(\langle b_i(\theta), \theta \rangle \right) + \gamma_i \left(\langle b_i(\theta), \theta \rangle - \alpha\kappa \right) \right] \\ &+ \frac{1}{n} \sum_{i=1}^{n} \frac{1 - y_i}{2} \left[\phi\left(\langle b_i(\theta), \theta \rangle \right) + \langle b_i(\theta), \theta \rangle \right] \Bigg\} \end{aligned} \tag{10.8}$$

$$s.t. \quad \|\theta\| \leqslant \frac{\alpha}{\beta + 1}, \quad \|\gamma\|_\infty \leqslant 1,$$

where for any $i \in [n]$, we use the shorthand $b_i(\theta) := b_i(\theta, x_i, y_i)$.

The proof of Theorem 10.3.1 is provided in Appendix I.

Remark 10.3.1. *The condition that the mapping $x \mapsto \phi(x) + x$ is non-decreasing is satisfied by many loss functions, including the logistic loss commonly used in supervised learning.*

Remark 10.3.2. *Note that we can convert the smooth convex-concave minmax problem* (10.8) *into a non-smooth convex minimization problem by explictly taking maximization over* γ*. But we refrain from doing as it has been observed [442] that solving the smooth minimax optimization problem is faster than solving the non-smooth problem.*

Proof of Theorem 10.3.1

The proof takes inspirations from [377, Theorem 1]. First, we define the *Wasserstein distance between distributions on* $\mathcal{Z}$ *with cost function* c (Definition 10.1.1).

Proof. (**Proof of Theorem 10.3.1**) Fix a $\theta \in \Theta$. Note that $b_i(\theta, \tilde{x}_i, +1) = \tilde{x}_i$. For any $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$, let $\ell((x, y), \theta) \coloneqq \phi(\langle x, \theta \rangle) - y \langle x, \theta \rangle$. We first analyze the inner supremum term, i.e.

$$\begin{aligned}
&\sup_{\mathbb{P} \in \mathcal{P}(\theta)} \mathbb{E}_{\mathbb{P}}[\phi(\langle x, \theta \rangle) - y\langle x, \theta \rangle] \\
&= \sup_{\mathbb{P} \in \mathcal{P}(\theta)} \int_{\mathcal{Z}} \ell(z, \theta) \mathbb{P}(z) dz \\
&= \begin{cases} \sup\limits_{\pi_\theta \in \Pi(\mathbb{P}, \tilde{\mathcal{D}}_n(\theta))} & \int_{\mathcal{Z}} \ell(z, \theta) \pi_\theta(dz, \mathcal{Z}), \\ \text{s.t.} & \int_{\mathcal{Z} \times \mathcal{Z}} \|z - \tilde{z}\| \pi_\theta(dz, d\tilde{z}) \leqslant \delta. \end{cases}
\end{aligned}$$

Here, $\Pi(\mathbb{P}, \tilde{\mathcal{D}}_n(\theta))$ denotes the set of all joint distributions that couple $\mathbb{P} \in \mathcal{P}(\theta)$ and $\tilde{\mathcal{D}}_n(\theta)$. Since the marginal distribution $\tilde{\mathcal{D}}_n(\theta)$ of $\tilde{z}$ is discrete, such couplings π_θ are completely determined by the conditional distribution $\mathbb{P}^i_\theta$ of z given $\tilde{z}_i = (\tilde{x}_i(\theta), \tilde{y}_i)$ for each $i \in \{1, \ldots, n\}$. That is:

$$\pi_\theta(dz, d\tilde{z}) = \frac{1}{n} \sum_{i \in [n]} \vartheta_{(b_i(\theta), \tilde{y}_i)}(d\tilde{z}) \mathbb{P}^i_\theta(dz),$$

where for any $(x, y) \in \mathcal{Z}$, $\vartheta_{(x,y)}$ is a *Dirac delta* distribution with its support at point (x, y).

We introduce some notations. Let $\mathcal{I}_{+1} = \{i \in [n] : \tilde{y}_i = +1\}$ and $\mathcal{I}_{-1} = \{i \in [n] : \tilde{y}_i = -1\}$. Let's introduce two distributions μ^i_θ and ν^i_θ such that

$$\mathbb{P}^i_\theta = \begin{cases} \mu^i_\theta & \text{if } i \in \mathcal{I}_{+1}, \\ \nu^i_\theta & \text{if } i \in \mathcal{I}_{-1}. \end{cases}$$

Due to Assumption (10.2.2), we have $\nu^i_\theta(dx, +1) = 0$ at every x. This implies:

$$\pi_\theta(dz, d\tilde{z}) = \frac{1}{n} \left(\sum_{i \in \mathcal{I}_{+1}} \vartheta_{(b_i(\theta), 1)}(d\tilde{z}) \mu^i_\theta(dz) + \sum_{i \in \mathcal{I}_{-1}} \vartheta_{(b_i(\theta), -1)}(d\tilde{z}) \nu^i_\theta(dz) \right).$$

With a slight abuse of notation, we denote $\mu^i_{\theta,+1}(dx) = \mu^i_\theta(dx,+1)$, $\mu^i_{\theta,-1}(dx) = \mu^i_\theta(dx,-1)$ and $\nu^i_\theta(dx) = \nu^i_\theta(dx,-1)$. The optimization problem of concern then simplifies to:

$$\begin{aligned}
\sup_{\mu^i_{\theta,\pm1},\nu^i_\theta} \quad & \frac{1}{n}\sum_{i\in\mathcal{I}_{+1}}\int_{\mathbb{R}^d}\ell((x,+1),\theta)\mu^i_{\theta,+1}(dx) + \frac{1}{n}\sum_{i\in\mathcal{I}_{+1}}\int_{\mathbb{R}^d}\ell((x,-1),\theta)\mu^i_{\theta,-1}(dx) \\
& \qquad + \frac{1}{n}\sum_{i\in\mathcal{I}_{-1}}\int_{\mathbb{R}^d}\ell((x,-1),\theta)\nu^i_\theta(dx) \\
\text{s.t.} \quad & \frac{1}{n}\sum_{i:\tilde{y}_i=+1}\int_{\mathbb{R}^d}\|(x,+1)-(b_i(\theta),\tilde{y}_i)\|\mu^i_{\theta,+1}(dx) \\
& \qquad + \frac{1}{n}\sum_{i:\tilde{y}_i=+1}\int_{\mathbb{R}^d}\|(x,-1)-(b_i(\theta),\tilde{y}_i)\|\mu^i_{\theta,-1}(dx) \\
& \int_{\mathbb{R}^d}\mu^i_{\theta,+1}(dx) + \int_{\mathbb{R}^d}\mu^i_{\theta,-1}(dx) = 1, \quad \forall \quad i\in\mathcal{I}_{+1} \\
& \int_{\mathbb{R}^d}\nu^i_\theta(dx) = 1, \quad \forall \quad i\in\mathcal{I}_{-1}.
\end{aligned}$$

First, we rewrite the inequality constraint above as follows. Recall that:

$$\begin{aligned}
& \frac{2\kappa}{n}\int_{\mathbb{R}^d}\sum_{i\in\mathcal{I}_{+1}}\mu^i_{\theta,-1}(dx) + \frac{1}{n}\int_{\mathbb{R}^d}\sum_{i\in\mathcal{I}_{+1}}\|x-b_i(\theta)\|\mu^i_{\theta,+1}(dx) \\
& \quad + \frac{1}{n}\int_{\mathbb{R}^d}\sum_{i\in\mathcal{I}_{+1}}\|x-b_i(\theta)\|\mu^i_{\theta,-1}(dx) + \frac{1}{n}\int_{\mathbb{R}^d}\sum_{i\in\mathcal{I}_{-1}}\|x-b_i(\theta)\|\nu^i_\theta(dx) \leqslant \delta.
\end{aligned}$$

Hence,

$$\begin{aligned}
\sup_{\mu^i_{\theta,\pm1},\nu^i_\theta} \quad & \frac{1}{n}\sum_{i\in\mathcal{I}_{+1}}\int_{\mathbb{R}^d}\ell((x,+1),\theta)\mu^i_{\theta,+1}(dx) + \frac{1}{n}\sum_{i\in\mathcal{I}_{+1}}\int_{\mathbb{R}^d}\ell((x,-1),\theta)\mu^i_{\theta,-1}(dx) \\
& \qquad + \frac{1}{n}\sum_{\tilde{y}_i=-1}\int_{\mathbb{R}^d}\ell((x,-1),\theta)\nu^i_\theta(dx) \\
\text{s.t.} \quad & \frac{2\kappa}{n}\int_{\mathbb{R}^d}\sum_{i\in\mathcal{I}_{+1}}\mu^i_{\theta,-1}(dx) + \frac{1}{n}\int_{\mathbb{R}^d}\sum_{i\in\mathcal{I}_{+1}}\|x-b_i(\theta)\|\mu^i_{\theta,+1}(dx) \\
& \qquad + \frac{1}{n}\int_{\mathbb{R}^d}\sum_{i\in\mathcal{I}_{+1}}\|x-b_i(\theta)\|\mu^i_{\theta,-1}(dx) + \frac{1}{n}\int_{\mathbb{R}^d}\sum_{i\in\mathcal{I}_{-1}}\|x-b_i(\theta)\|\nu^i_\theta(dx) \leqslant \delta \\
& \int_{\mathbb{R}^d}\mu^i_{\theta,+1}(dx) + \int_{\mathbb{R}^d}\mu^i_{\theta,-1}(dx) = 1, \quad \forall \quad i\in\mathcal{I}_{+1} \\
& \int_{\mathbb{R}^d}\nu^i_\theta(dx) = 1, \quad \forall \quad i\in\mathcal{I}_{-1}.
\end{aligned}$$

Now, we can use duality to reformulate the infinite-dimensional optimization problem

into a finite-dimensional problem:

$$\sup_{\mathbb{P}\in\mathcal{P}(\theta)} \mathbb{E}_{\mathbb{P}}[\phi(\langle x,\theta\rangle) - y\langle x,\theta\rangle]$$
$$= \begin{cases} \inf_{\alpha,s_i} & \alpha\delta + \frac{1}{n}\sum_{i\in\mathcal{I}_{+1}} s_i + \frac{1}{n}\sum_{i\in\mathcal{I}_{-1}} t_i \\ \text{s.t.} & \sup_x \ell((x,+1),\theta) - \alpha\cdot\frac{1+\tilde{y}_i}{2}\|x-b_i(\theta)\| \leqslant s_i \quad \forall\, i\in\mathcal{I}_{+1} \\ & \sup_x \ell((x,-1),\theta) - \alpha\cdot\frac{1+\tilde{y}_i}{2}\|x-b_i(\theta)\| - \alpha\kappa(1+\tilde{y}_i) \leqslant s_i \quad \forall\, i\in\mathcal{I}_{+1} \\ & \sup_x \ell((x,-1),\theta) - \alpha\cdot\frac{1-\tilde{y}_i}{2}\|x-b_i(\theta)\| \leqslant t_i \quad \forall\, i\in\mathcal{I}_{-1} \\ & \alpha \geqslant 0, \end{cases}$$

which is equivalent to:

$$\sup_{\mathbb{P}\in\mathcal{P}(\theta)} \mathbb{E}_{\mathbb{P}}[\phi(\langle x,\theta\rangle) - y\langle x,\theta\rangle]$$
$$= \begin{cases} \inf_{\alpha,s_i} & \alpha\delta + \frac{1}{n}\sum_{i\in\mathcal{I}_{+1}} s_i + \frac{1}{n}\sum_{i\in\mathcal{I}_{-1}} t_i \\ \text{s.t.} & \sup_x \ell((x,+1),\theta) - \alpha\|x-b_i(\theta)\| \leqslant s_i \quad \forall\, i\in\mathcal{I}_{+1} \\ & \sup_x \ell((x,-1),\theta) - \alpha\|x-b_i(\theta)\| - 2\alpha\kappa \leqslant s_i \quad \forall\, i\in\mathcal{I}_{+1} \\ & \sup_x \ell((x,-1),\theta) - \alpha\|x-b_i(\theta)\| \leqslant t_i \quad \forall\, i\in\mathcal{I}_{-1} \\ & \alpha \geqslant 0, \end{cases}$$

We now invoke [442, Lemma A.1], which claims that for any $\tilde{y}\in\{+1,-1\}$ and $\tilde{x}\in\mathbb{R}^d$:

$$\sup_x \ell((x,\tilde{y}),\theta) - \alpha\|x-\tilde{x}\| = \begin{cases} \ell((\tilde{x},\tilde{y}),\theta) & \text{if } \|\theta\| \leqslant \alpha/(L+1), \\ -\infty & \text{otherwise.} \end{cases}$$

We now have:

$$\sup_{\mathbb{P}\in\mathcal{P}(\theta)} \mathbb{E}_{\mathbb{P}}[\phi(\langle x,\theta\rangle) - y\langle x,\theta\rangle]$$
$$= \begin{cases} \inf_{\alpha,s_i} & \alpha\delta + \frac{1}{n}\sum_{i\in\mathcal{I}_{+1}} s_i + \frac{1}{n}\sum_{i\in\mathcal{I}_{-1}} t_i \\ \text{s.t.} & \ell((b_i(\theta),+1),\theta) \leqslant s_i \quad \forall\, i\in\mathcal{I}_{+1} \\ & \ell((b_i(\theta),-1),\theta) - 2\alpha\kappa \leqslant s_i \quad \forall\, i\in\mathcal{I}_{+1} \\ & \ell((b_i(\theta),-1),\theta) \leqslant t_i \quad \forall\, i\in\mathcal{I}_{-1} \\ & \alpha \geqslant 0 \\ & \|\theta\| \leqslant \alpha/(L+1). \end{cases}$$

In the above presented optimization problem we can conclude that:

$$t_i = \phi(\langle b_i(\theta),\theta\rangle) + \langle b_i(\theta),\theta\rangle \qquad \forall i\in\mathcal{I}_{-1}$$
$$s_i = \max\{\ell((b_i(\theta),+1),\theta), \ell((b_i(\theta),-1),\theta) - 2\alpha\kappa\} \qquad \forall i\in\mathcal{I}_{+1}.$$

To further simplify the s_i expression, note that:

$$\begin{aligned} s_i &= \max\{\phi(\langle b_i(\theta),\theta\rangle) - \langle b_i(\theta),\theta\rangle\,, \phi(\langle b_i(\theta),\theta\rangle) + \langle b_i(\theta),\theta\rangle - 2\alpha\kappa\} \\ &= \phi(\langle b_i(\theta),\theta\rangle) - \langle b_i(\theta),\theta\rangle + \max\{0, 2\,\langle b_i(\theta),\theta\rangle - 2\alpha\kappa\} \\ &= \phi(\langle b_i(\theta),\theta\rangle) - \alpha\kappa + \max_{\gamma_i:|\gamma_i|\leqslant 1} \gamma_i\,(\langle b_i(\theta),\theta\rangle - \alpha\kappa)\,, \end{aligned}$$

so the overall objective can be written as:

$$\begin{aligned} &\sup_{\mathbb{P}\in\mathcal{P}(\theta)} \mathbb{E}_{\mathbb{P}}[\phi(\langle x,\theta\rangle) - y\langle x,\theta\rangle] \\ &= \begin{cases} \inf_\alpha \max_{\gamma:\|\gamma\|_\infty\leqslant 1} & \alpha(\delta-\kappa) + \frac{1}{n}\sum_i \frac{1+\tilde{y}_i}{2}\left(\phi(\langle b_i(\theta),\theta\rangle) + \gamma_i\left(\langle b_i(\theta),\theta\rangle - \alpha\kappa\right)\right) \\ & \qquad + \frac{1}{n}\sum_i \frac{1-\tilde{y}_i}{2}\left(\phi(\langle b_i(\theta),\theta\rangle) + \langle b_i(\theta),\theta\rangle\right) \\ \text{s.t.} & \|\theta\| \leqslant \alpha/(L+1). \end{cases} \end{aligned}$$

We claim that the minimax objective above is convex is θ. There are mainly two cases to analyze:

1. **Case I ($i \in \mathcal{I}_{+1}$)**: We have $b_i(\theta) = \tilde{x}_i$ as per the strategic classification model. Therefore $\langle b_i(\theta),\theta\rangle$ is a linear function. For every γ, α, we claim that the mapping $\theta \mapsto \phi(\langle b_i(\theta),\theta\rangle) + \gamma_i(\langle b_i(\theta),\theta\rangle - \alpha\kappa)$ is convex. Indeed, the assumption that ϕ is convex and the observation that $\langle b_i(\theta),\theta\rangle$ is affine in θ ensures the convexity.

2. **Case II ($i \in \mathcal{I}_{-1}$)**: We know from Lemma 10.2.1 that $\langle b_i(\theta),\theta\rangle$ is convex in θ. Moreover, the convexity of ϕ and the assumption that $z \mapsto \phi(z) + z$ is non-decreasing ensures that $\phi(\langle b_i(\theta),\theta\rangle) + \langle b_i(\theta),\theta\rangle$ is convex for every i.

This concludes the proof.

□

10.4 A New Gradient-free Algorithm for Convex Concave Min-max Optimization

In this section, we introduce a novel gradient-free version of the well-studied Optimistic Gradient Descent Ascent (OGDA) algorithm to solve convex-concave min-max optimization problem, and provide non-asymptotic rates showing that it can efficiently find the saddle point in constrained convex-concave problems. In the following section, we use this algorithm to numerically solve (10.8).

Preliminaries on Min-max Optimization

Here, we review the following form of min-max optimization problem:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} L(x,y), \tag{10.9}$$

where $\mathcal{X} \subset \mathbb{R}^{d_x}$, $\mathcal{Y} \subset \mathbb{R}^{d_y}$, and $L := \frac{1}{n}\sum_{i=1}^{n} L_i$, where $L_1, \ldots, L_n : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ denote n individual loss functions. For convenience, we denote $d := d_x + d_y$.

Assumption 10.4.1. *The following statements hold:*

1. *The sets $\mathcal{X} \subset \mathbb{R}^{d_x}$ and $\mathcal{Y} \subset \mathbb{R}^{d_y}$ are convex and compact.*
2. *The functions $L_1, \ldots, L_n : \mathbb{R}^d \to \mathbb{R}$ are convex in $x \in \mathbb{R}^{d_x}$ for each $y \in \mathbb{R}^{d_y}$, concave in $y \in \mathbb{R}^{d_y}$ for each $x \in \mathbb{R}^{d_x}$, and G-Lipschitz[2] and ℓ-smooth[3] in $(x, y) \in \mathbb{R}^d$ (which implies that $L : \mathbb{R}^d \to \mathbb{R}$, by definition, also possesses the same properties).*

For ease of exposition, we denote

$$u := (x, y), \quad M_L := \sup_{u \in \mathcal{X} \times \mathcal{Y}} |L(u)|, \quad D := \sup_{u, u' \in \mathcal{X} \times \mathcal{Y}} \|u - u'\|_2,$$

and define the operators $F, F_i : \mathbb{R}^d \to \mathbb{R}^d$, for each $i \in [n]$, by:

$$F(u) := \begin{bmatrix} \nabla_x L(u) \\ -\nabla_y L(u) \end{bmatrix}, \quad F_i(u) := \begin{bmatrix} \nabla_x L_i(u) \\ -\nabla_y L_i(u) \end{bmatrix}. \tag{10.10}$$

Observe that under Assumption 10.4.1, $M_L, D < \infty$, and F and each F_i are monotone[4]. Finally, we define the *gap function* $\Delta : \mathbb{R}^d \to [0, \infty)$ associated with the loss L by

$$\Delta(x, y) := L(x, y^\star) - L(x^\star, y) \geqslant 0, \tag{10.11}$$

where $u^\star := (x^\star, y^\star) \in \mathcal{X} \times \mathcal{Y}$ denotes the min-max saddle point of the overall loss $L(x, y)$, and $(x, y) \in \mathcal{X} \times \mathcal{Y}$ denotes any feasible point. This gap function allows us to measure the convergence rate of our proposed algorithm. To this end, we define the ϵ-optimal saddle-point of (10.9) as follows.

Definition 10.4.1 (**ϵ-optimal saddle point solution**). *A feasible point $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is said to be an ϵ-optimal saddle-point solution of* (10.9) *if*

$$\Delta(x, y) = L(x, y^\star) - L(x^\star, y) \leqslant \epsilon.$$

[2] A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be G-Lipschitz for some $G > 0$ if, for each $u, u' \in \mathbb{R}^d$,

$$|f(u) - f(u')| \leqslant G\|u - u'\|_2.$$

[3] A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be ℓ-smooth for some $\ell > 0$ if, for each $u, u' \in \mathbb{R}^d$:

$$\|\nabla f(u) - \nabla f(u')\|_2 \leqslant \ell \cdot \|u - u'\|_2.$$

[4] A function $F : \mathbb{R}^d \to \mathbb{R}^d$ is called *monotone* if $\langle F(x) - F(y), x - y \rangle \geqslant 0$ for all $x, y \in \mathbb{R}^d$.

Zeroth-Order Gradient Estimates

In our zeroth-order, random reshuffling-based variant of the OGDA algorithms, we use the one-shot randomized gradient estimator [394, 141]. In particular, given the current iterate $u \in \mathbb{R}^d$ and a *query radius* $\delta > 0$, we sample a vector v uniformly from unit sphere $\mathcal{S}^{d-1}$ (i.e. $v \sim \mathsf{Unif}(\mathcal{S}^{d-1})$), and define the zeroth-order estimator $\hat{F}(u;\delta,v) \in \mathbb{R}^d$ of the min-max loss $L(u)$ to be:

$$\hat{F}(u;\delta,v) := \frac{d}{\delta} L(u+\delta v)v. \tag{10.12}$$

Several important properties of this estimator is reviewed in Appendix I.

Optimistic Gradient Descent Ascent with Random Reshuffling (OGDA-RR)

In this subsection, we formulate our main algorithm, Optimistic Gradient Descent Ascent with Random Reshuffling (OGDA-RR). In each *epoch* $t \in \{0, 1, \cdots, T-1\}$, the algorithm generates a uniformly random permutation $\sigma^t := (\sigma_1^t, \cdots, \sigma_n^t)$ of $[n] := \{1, \cdots, n\}$ independently of any other randomness. This is what is referred as random reshuffling (or sampling without replacement) where within every epoch we do not re-sample and this naturally gives rise to correlations between different iterations within an epoch. Furthermore, the algorithm fixes a query radius $\delta^t > 0$ and search direction $v_i^t \in \mathbb{R}^d$ in every epoch t. Note that query radii only depends on epoch indices t, and not on sample indices $\{\sigma_i^t\}_{i=1}^n$. For each $i \in [n], t \in [T-1]$, we compute the OGDA-RR update as follows:

$$u_{i+1}^t = \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}\Big(u_i^t - \eta^t \hat{F}_{\sigma_i^t}(u_i^t;\delta^t,v_i^t) - \eta^t \hat{F}_{\sigma_{i-1}^t}(u_i^t;\delta^t,v_i^t) + \eta^t \hat{F}_{\sigma_{i-1}^t}(u_{i-1}^t;\delta^t,v_{i-1}^t)\Big), \tag{10.13}$$

where the terms $\hat{F}_{\sigma_i^t}$ and $\hat{F}_{\sigma_{i-1}^t}$ are the zeroth-order estimators of gradients $F_{\sigma_i^t}$ and $F_{\sigma_{i-1}^t}$ (defined in (10.10)).

After repeating this process for T epochs, the algorithm returns the step-size weighted average of the iterates, $\tilde{u}^T := \frac{1}{n\cdot\sum_{t=0}^{T-1}\eta^t}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\eta^t u_i^t$. The following theorem states that if we run Algorithm 9 long enough then $\tilde{u}^T$ will be close to the saddle point.

Theorem 10.4.1. *Let $L(u)$ denote the objective function in the constrained min-max optimization problem given by (10.9), and let $u^\star = (x^\star, y^\star) \in \mathcal{X}\times\mathcal{Y}$ denote any saddle point of $L(u)$. Fix $\epsilon > 0$. Suppose Assumption 10.4.1 holds, and the number of epochs T, step sizes*

Algorithm 9 OGDA-RR Algorithm

Require: Stepsizes η^t, δ^t, data points $\{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}$, initial value $u_0^{(0)}$, time horizon T

1: **for** $t = 0$ **to** $T-1$ **do**
2: $\quad \sigma^t = (\sigma_1^t, \cdots, \sigma_n^t) \leftarrow$ a random permutation of the set $[n]$
3: $\quad$ **for** $i = 0$ **to** $n-1$ **do**
4: $\quad\quad$ Sample $v_i^t \sim \mathsf{Unif}(\mathcal{S}^{d-1})$
5: $\quad\quad u_{i+1}^t \leftarrow$ update from equation (10.13)
6: $\quad$ **end for**
7: $\quad u_0^{(t+1)} \leftarrow u_n^t$
8: $\quad u_{-1}^{(t+1)} \leftarrow u_{n-1}^t$
9: **end for**

Ensure: $\tilde{u}^T := \frac{1}{n \cdot \sum_{t=0}^{T-1} \eta^t} \sum_{t=0}^{T-1} \sum_{i=1}^{n} \eta^t u_i^t$

sequence $\{\eta^t\}_{t=0}^{T-1}$*, and query radius sequence* $\{\delta^t\}_{t=0}^{T-1}$ *satisfy:*

$$\eta^t := \eta^0 \cdot (t+1)^{-3/4+\chi}, \qquad \forall t \in \{0, 1, \cdots, T-1\},$$
$$\delta^t := \delta^0 \cdot (t+1)^{-1/4}, \qquad \forall t \in \{0, 1, \cdots, T-1\},$$
$$T > \frac{1}{\epsilon^4}\left(\frac{3}{16n}D + \frac{5}{4}C \cdot \max\left\{\delta^0, \eta^0, \eta^0\delta^0, \frac{\eta^0}{\delta^0}, \frac{\eta^0}{(\delta^0)^2}\right\}\left(1 + \frac{1}{\chi}\right)\right)^{\frac{4}{1-4\chi}},$$

for some initial step size $\eta^0 \in \left(0, \frac{1}{2\ell}\right)$*, initial query radius* $\delta^0 > 0$*, parameter* $\chi \in (0, 1/4)$*, and constant* $C = \mathcal{O}(nd^2D)$*. Then the iterates* $\{u_i^t\}$ *generated by the OGDA-RR Algorithm (Alg. 9) satisfy:*

$$\mathbb{E}\Big[\Delta(\tilde{u}^T)\Big] < \epsilon.$$

There are three main components to the proof of Theorem 10.4.1: First, we bound the bias introduced due to random reshuffling (or sampling without replacement) by Wasserstein distance between two appropriate distributions that characterize the correlations introduced between iterates because of random reshuffling. Second, we bound the aforementioned Wasserstein distance by constructing an appropriate coupling between iterates generated with and without random reshuffling [191]. The coupled iterates thus obtained are then bounded by exploiting the recent connections between OGDA method and proximal point methods [303], which is one of the main contributions of our proof technique. Third, we balance the bias and variance introduced due to zeroth-order gradient estimator by suitably choosing the step size sequence $\{\eta^t\}$ and the perturbation radius sequence $\{\delta^t\}$.

Remark 10.4.1. *Note that one can obtain better convergence rates if we use a multi-point zeroth-order estimator as opposed to the single-point zeroth-order estimator* (10.12)*. For*

instance, if we use the following two-point gradient estimator:

$$\hat{F}(u;\delta,v) = \frac{d}{2\delta}(L(u+\delta v) - L(u-\delta v))v$$

then it follows easily from our analysis that the epochs required to obtain an $\epsilon-$optimal saddle point decreases from $\tilde{\mathcal{O}}\left(\frac{n^4d^8}{\epsilon^4}\right)$ to $\tilde{\mathcal{O}}\left(\frac{n^2d^4}{\epsilon^2}\right)$. But we restrict our presentation to single point estimators, as the distributionally robust strategic classification problem discussed in previous section demands that we query the objective function as minimally as possible. It is an interesting future research direction to study the OGDA-RR algorithm with more advanced zeroth-order methods. Detailed proof is provided in Appendix I.

Remark 10.4.2. *The analysis of OGDA algorithm with random reshuffling and exact gradient information is an immediate feature of our proof technique. For such algorithms, the number of epochs required to obtain an $\epsilon-$optimal saddle point is $\tilde{\mathcal{O}}\left(\frac{n^2}{\epsilon^2}\right)$. Note that there is no dependence on d with exact gradient based methods.*

Remark 10.4.3. *Note that the OGDA-RR algorithm is computationally more efficient than [442, Algorithm 2],* if *one replaces the gradient estimates with true gradient values. This is because that algorithm requires $\mathcal{O}(\log(n))$ inner loop iterations to approximate a proximal point update. Here, we overcome extra computations by exploiting the recent perspective that the OGDA update is a perturbed proximal point update [304].*

10.5 Empirical Results

In this section we deploy zeroth-order OGDA algorithm with random reshuffling to solve the convex concave reformulation of WDRSC as presented in (10.8). Throughout the rest of this chapter, we denote the min-max objective in (10.8) by $L(\alpha, \theta, \gamma)$.

We point out that in order to solve (10.8), the zeroth-order method should only be applied to estimate the gradient with respect to θ. This is because the gradient with respect to other variables, namely (α, γ), can be exactly computed. Specifically, to compute derivative with respect to θ the designer must know the best response function which is often not available and it can only be queried.

We now present some illustrations of the empirical performance of our proposed algorithm, as well as empirical justification for solving the WDRSC problem over existing prior approaches to strategic classification.

Experimental Setup

Our first set of empirical results uses synthetic data to illustrate the effectiveness of our algorithms. The datasets used in this section are constructed as follows: the ground truth classifier $\theta^\star$ and features $\tilde{x}_i$ are sampled as $\theta^\star \sim \mathcal{N}(0, I_d)$ and $\tilde{x}_i \sim$ i.i.d. $\mathcal{N}(0, I_d)$, for each

$i \in [n]$, while the ground truth labels $\tilde{y}_i$ are given by $\tilde{y}_i = \mathsf{sign}(\langle \tilde{x}_i, \theta^\star \rangle + z_i)$ for each $i \in [n]$, where $z_i \sim$ i.i.d. $\mathcal{N}(0, 0.1 \cdot I_d)$. We use $n \in \{500, 1000\}$ with $d = 10$. The first five of the $d = 10$ features are chosen to be strategic. In all experiments, we take $\kappa = 0.5$ and $\delta = 0.4$. Each strategic agent $i \in [n]$ has a utility function given by:

$$u_i(x; \theta, \tilde{x}_i, \tilde{y}_i, \zeta_i) = \frac{1 - \tilde{y}_i}{2} \langle x, \theta \rangle - \frac{1}{2\zeta_i} \|x - \tilde{x}_i\|^2, \tag{10.14}$$

where ζ_i denote the perturbation "power" of agent i. For simplicity, we assume all agents are homogeneous, in the sense that $\zeta_i = \zeta > 0$ for all $i \in [n]$; in practice, one need not impose this assumption. Given this utility function, the best response of agents takes the form:

$$b_i(\theta, \tilde{x}_i, \tilde{y}_i; \zeta) = \begin{cases} \tilde{x}_i & \text{if } \tilde{y}_i = +1, \\ \tilde{x}_i + \zeta\theta & \text{if } \tilde{y}_i = -1, \end{cases} \tag{10.15}$$

where, in our simulations, we fix $\zeta = 0.05$. We reemphasize that our algorithm does not use the value of ζ in any of its computations. For purposes of illustration, we focus on the performance of the following algorithms:

1. Zeroth-order optimistic-GDA *with* random reshuffling (see Algorithm 9),
2. Zeroth-order optimistic-GDA *without* random reshuffling (see Algorithm 17 in Appendix I),
3. Zeroth-order stochastic-GDA *with* random reshuffling (see Algorithm 18 in Appendix I),
4. Zeroth-order stochastic-GDA *without* random reshuffling (see Algorithm 19 in Appendix I).

and we evaluate the proposed algorithms and model formulation on two criteria:

1. *Suboptimality*: To measure suboptimality, we use the *gap function*

   $$\Delta(\alpha, \theta, \gamma) = L(\alpha, \theta, \gamma^\star) - L(\alpha^\star, \theta^\star, \gamma),$$

   defined in Definition 10.11, where $(\alpha^\star, \theta^\star, \gamma^\star)$ is a solution of the min-max reformulation (10.8) of the WDRSC problem. If the objective $L(\cdot)$ is convex-concave, $\Delta(\cdot)$ is non-negative, and equals zero at (and only at) saddle points.

2. *Accuracy*: Given a data set $\{(\tilde{x}_i, \tilde{y}_i)\}_{i \in [n]}$, the accuracy of a classifier θ is measured as $\frac{1}{n} \sum_{i \in [n]} \tilde{y}_i \langle b_i(\theta, \tilde{x}_i, \tilde{y}_i; \zeta), \theta \rangle$. Under this criterion we compare the accuracy under different perturbations for different classifiers θ;

To compute suboptimality, we first compute a true min-max saddle point $(\alpha^\star, \theta^\star, \gamma^\star)$ via a first order gradient based algorithm (namely, GDA). All experiments were run using Python 3.7 on a standard MacBook Pro laptop (2.6 GHz Intel Core i7 and 16 GB of RAM).

Results

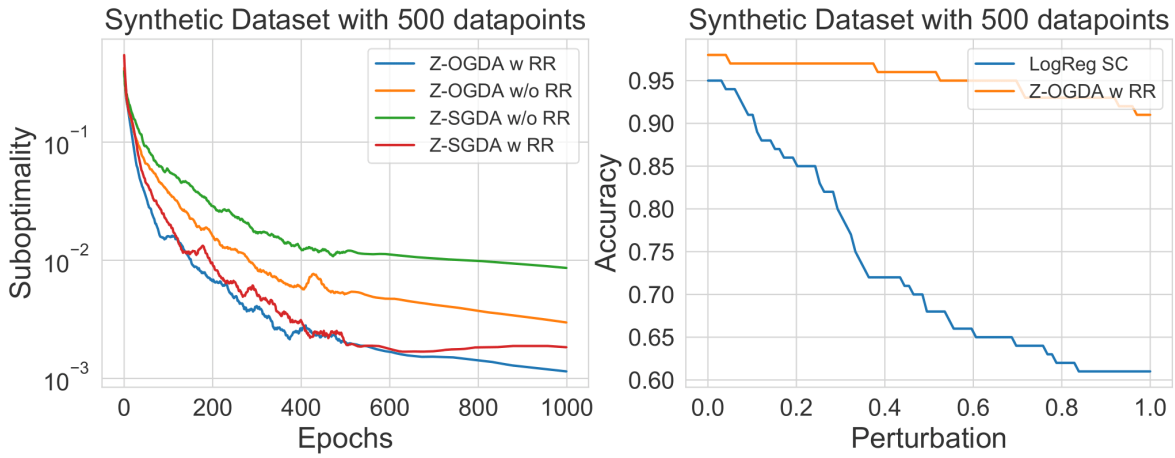
Simulation results presented in Figure (10.1a)-(10.1b) show that our proposed algorithm (i.e. Algorithm 1) outperforms algorithms without reshuffling (i.e. 2 and 4). However, its performance resembles that of zeroth-order stochastic-GDA with random reshuffling. More experimental studies need to be conducted to more conclusively determine whether 1 outperforms 3, or vice versa. In fact, there has been no theoretical investigations even for the *first order* stochastic-GDA algorithm with random reshuffling.

In Figure 10.1, we also compare the robustness of the classifier obtained by using Algorithm 1 with that obtained from prior work on solving probems of strategic classification trained with $\zeta = 0.05$ (referred as *LogReg SC* in Figure 10.1). As expected, due to the formulation, the performance of the classifier obtained via 1 degrades gracefully even when subject to large perturbations, while the performance of existing approaches to strategic classification degrades rapidly. Further numerical results on synthetically generated and real world datasets are given in Appendix I.

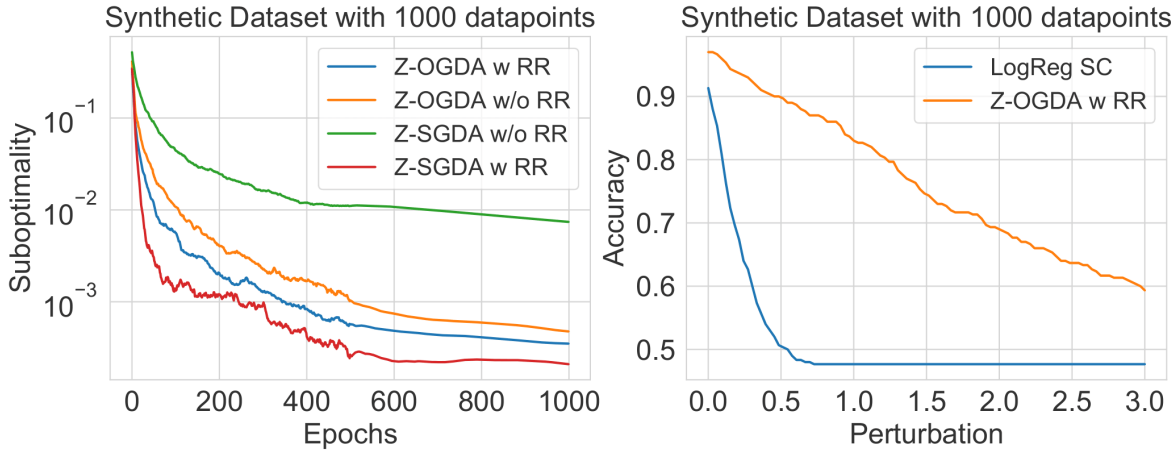
10.6 Concluding Remarks

This chapter addresses the challenge of learning in strategic environments where agent behavior is influenced by the deployed model, and where the learner's assumptions about agent responses may be misspecified. To overcome the limitations of prior work that assumes perfect knowledge of agent behavior, we propose a novel formulation based on a Wasserstein distributionally robust optimization (DRO) framework. This approach explicitly accounts for model uncertainty by optimizing for worst-case performance over an ambiguity set that captures strategically perturbed data distributions. We show that the resulting infinite-dimensional DRO problem can be reformulated as a finite-dimensional, smooth convex-concave min-max problem.

Building on this reformulation, we develop a gradient-free (zeroth-order) optimization algorithm tailored for settings in which gradients of the min-max objective are unavailable. This is particularly important in the DRO setting, where the gradient with respect to users' strategic responses is not directly accessible. The proposed algorithm enables learning purely from observed outcomes, without requiring access to the underlying response mechanisms. Furthermore, the method is applicable to general convex-concave min-max problems. Our theoretical analysis establishes non-asymptotic convergence guarantees for the algorithm, even under random reshuffling of data-points.



(a) $n = 500$



(b) $n = 1000$

Figure 10.1: Experimental results for a synthetic dataset with $n = 500$ and $n = 1000$. (Left panes of (10.1a), (10.1b))) Suboptimality iterates generated by the four algorithms 1, 2, 3, 4, respectively denoted as *Z-OGDA w RR*, *Z-OGDA w/o RR*, *Z-SGDA w RR*, *Z-SGDA w/o RR*. (Right panes of (10.1a), (10.1b))) Comparison between decay in accuracy of strategic classification with logistic regression (trained with $\zeta = 0.05$) and Algorithm 1 with change in perturbation.

Chapter 11

Follower Agnostic Learning in Stackelberg Games

Incentive mechanisms play a crucial role in many societal systems, where outcomes are governed by the interactions of a large number of self-interested users (or algorithms acting on their behalf). The outcome of such strategic interactions, characterized by the Nash equilibrium, is often suboptimal because individual players typically do not account for the *externality* of their actions (i.e., how their actions affect the costs of others) when minimizing their own costs. An important way to address this suboptimality is to provide players with incentives that align their individual goal of cost minimization with the goal of minimizing the total cost of the societal system [242, 33]. However, this problem becomes more challenging as the system operator often needs to account for the learning behavior of players, who repeatedly update their strategies in response to the incentive mechanism, especially when the physical system experiences a random shock and players are learning to reach a new equilibrium [32, 106, 270, 97].

In this chapter, we study the problem of designing incentive mechanisms in settings where the operator (also referred to as the leader) has no knowledge of the utility functions or learning algorithms used by the agents (referred to as followers). These strategic interactions can be naturally modeled using Stackelberg games, which encompass a wide range of practical applications, including incentive design, Bayesian persuasion, inverse optimization, bilevel optimization, cybersecurity, and adversarial learning [253, 18, 258, 244, 443, 268], to name a few.

Mathematically, Stackelberg games are represented as follows:

$$\begin{aligned} \min_{x\in X, y\in Y} \quad & f(x,y) \\ \text{s.t.} \quad & y \in S(x) := \mathsf{SOL}(Y, G(x,\cdot))), \end{aligned} \tag{11.1}$$

where X is the leader's strategy set, $Y \subseteq \mathbb{R}^d$ is the followers' (joint) strategy set, $f : X \times Y \to \mathbb{R}$ is the utility of the *leader*, $G : X \times Y \to \mathbb{R}^d$ is the *game Jacobian* of *followers* and $\mathsf{SOL}(Y, G(x,\cdot))$ is a *variational inequality problem* that denotes the equilibrium response

of followers, given the strategy of leader be $x \in X$. Assuming that the set $\mathsf{br}(x)$ is singleton for every $x \in X$ (commonly referred as *lower-level singleton assumption*), (11.1) is equivalent to optimizing the following *hyper-objective*:

$$\min_{x \in X} \tilde{f}(x) := f(x, \mathsf{br}(x)). \tag{11.2}$$

Note that in general (11.2) is non-convex optimization problem. Thus, the goal in Stackelberg games is to find a stationary point / local optima of (11.2) ([139]).

In numerous practical scenarios, it is unrealistic to presume that the leader possesses any information regarding the variational inequality problem at the lower-level, including the mapping $G(x, \cdot)$ and even their strategy set Y – information traditionally assumed in prior research on solving Stackelberg games. Thus, the key question we ask in this chapter is:

> **Q:** *Can we design efficient algorithms for Stackelberg games where the leader does not require any explicit knowledge of the game played between followers?*

In this chapter, we affirmatively answer the above question in the setting where the leader can only probe the followers with different strategies and receive estimates of their (approximated) equilibrium responses. This is in contrast to the common assumption in the literature on Stackelberg games, where it is assumed that the leader has access to an equilibrium or best-response of followers either by knowledge of the utility function of followers or through an oracle. In particular, we consider that followers are rational in the sense that they employ an adaptation/learning algorithm, which asymptotically converges to the equilibrium [149].

We propose a *two-loop* algorithm where, in the outer loop, the leader fixes its strategy (i.e., the value of x) and announces it to the followers. Between two updates of the leader's strategy, the followers employ an adaptation algorithm, for a finite number of steps, so that they converge to an *approximate* equilibrium (or best-response). Upon observing the followers' behavior, the leader constructs an approximate estimator of the gradient of the hyper-objective (11.2) and updates its strategy via gradient descent using the estimator.

We show that the proposed algorithm converges to a stationary point of (11.2) at a rate $\mathcal{O}(T^{-1/2})$. Moreover, we show that if the hyper-objective satisfies the *strict-saddle property*, i.e. the minimum eigenvalue at any saddle point is strictly negative, then the iterates asymptotically avoid saddle points (which include local maxima) and converge to a local minima of the hyper-objective (aka local Stackelberg equilibrium [139]).

We corroborate the theoretical results by conducting a simulated study of the proposed algorithm to design tolls over the Sioux Falls (South Dakota, US) transportation network. In this setup, we assume that the leader does not know the origin-destination (o-d) demand of travelers moving between different o-d pairs, which is sensitive information.

Related works

Learning in Stackelberg games: Learning in Stackelberg games with *finite* actions is an active area of research ([55, 241, 334, 25, 374]), where the leader has access to either a noisy

or exact best response oracle. Furthermore, a dominant paradigm in this literature is to consider two-player games with finite strategy sets or linearly parametrized utility functions, with the exception of [139, 165, 164, 253]. In [139], the authors study the convergence of a two-timescale algorithm to the Stackelberg equilibrium, requiring knowledge of the Hessian of followers' utility functions for leader updates. In [165, 164], the authors require the followers to follow a specific (i.e. gradient type) learning algorithm in order to ensure convergence. Finally, in [253], the authors impose strong convexity assumption on the hyper-objective which is restrictive (as shown in [272]). In this chapter, we aim to design follower-agnostic learning in a general-sum Stackelberg game in continuous spaces with *no* knowledge of the followers' utility functions or learning algorithms and not imposing restrictive assumptions about convexity of hyper-objective.

Bilevel optimzation. Bilevel optimization, a subset of problem (11.1), is extensively studied in literature, resembling a Stackelberg game with a single leader and follower. Existing research on bilevel optimization pursues three main approaches. The first utilizes a value function-based approach, converting the problem into a constrained single-level optimization problem with convergence guarantees to approximate Karush-Kuhn-Tucker (KKT) points [393, 437]. However, such points may not capture locally optimal solutions [90]. Another line of research focuses on asymptotic convergence of solutions of simpler bilevel problems than (11.1) under various assumptions on the lower-level objective function structure [261, 259, 260]. The third line explores solving the non-convex optimization problem (11.2) using gradient descent, requiring the computation of the gradient of the solution mapping, denoted as $\nabla \mathsf{br}(x)$. While many methods exist for approximating $\nabla \mathsf{br}(x)$, including *Automatic Implicit Differentiation* (AID) ([163, 145, 333, 196, 146]), or *Iterative Differentiation* ([145, 155, 162, 376]), this chapter is closely related to zeroth-order methods, specifically avoiding the computation of the Hessian ([90]). Our proposed algorithm shares similarities with [90], but we eliminate the need for oracle access to a lower-level optimal solution, leveraging two-timescale stochastic approximation to analyze accumulated errors [90].

11.1 Problem Formulation

Consider the following Stackelberg game

$$\begin{aligned} \min_{x\in X, y\in Y} \quad & f(x,y) \\ \text{such that} \quad & y \in S(x) := \mathsf{SOL}(Y, G(x,\cdot))), \end{aligned} \tag{SG}$$

where *(i)* $X = \mathbb{R}^d$ and $Y \subset \mathbb{R}^{d'}$ is assumed to be convex and compact set; *(ii)* $f : X \times Y \to \mathbb{R}$ and $G : X \times Y \to \mathbb{R}^{d'}$ are twice continuously differentiable functions; *(iii)* $\mathsf{SOL}(Y, G(x,\cdot))$ denotes the solution to variational inequality characterized by functional $G(x,\cdot)$. That is, $\mathsf{SOL}(Y, G(x,\cdot)) = \{y \in Y : \langle y' - y, G(x,y)\rangle \geqslant 0, \quad \forall\ y' \in Y\}$. Under mild conditions on the monotonicity of $G(x,\cdot)$, it is ensured that $S(x)$ is non-empty and convex ([134]).

In what follows, we call a continuously differentiable function $\tilde{f} : \mathbb{R}^d \to \mathbb{R}$ to be $L-$Lipschitz if for every $x, x' \in \mathbb{R}^d$, $\|\tilde{f}(x) - \tilde{f}(x')\| \leqslant L\|x - x'\|$. Furthermore, we call it to be $\ell-$smooth if for every $x, x' \in \mathbb{R}^d$, $\|\nabla\tilde{f}(x) - \nabla\tilde{f}(x')\| \leqslant \ell\|x - x'\|$.

Next, we introduce the main assumptions on the parameters of (SG) made throughout this chapter.

Assumption 11.1.1. *(1) For every $y \in Y$, the function $f(\cdot, y)$ is L_1-Lipschitz. Additionally, for every $x \in X$, the function $f(x, \cdot)$ is L_2-Lipschitz and ℓ_2-smooth.*

(2) For every $x \in X$, the set $S(x)$ is singleton and function $S(x)$ is L_S-Lipschitz.

(3) The function $\tilde{f}(x) = f(x, S(x))$ is twice-continuously differentiable, $\tilde{L}$-Lipschitz and $\tilde{\ell}-$smooth.

Assumption 11.1.1-(1) is a common assumption employed in literature to derive rates of convergence [253, 164]. Assumption 11.1.1-(2), which requires that the set $S(x)$ exists, is singleton and Lipscthiz continuous for every x, holds for strongly monotone games at lower level [112]. Furthermore, it also applies to the incentive design problem in routing games, as discussed in Section 11.2. Assumption 11.1.1-(3) is a technical condition we impose on the hyper-objective to use Taylor's series expansion in the proof of convergence. Notably, this assumption is less restrictive than those imposed on the hyper-objective in [253]. We believe this assumption can be further relaxed, but we leave this as an interesting direction for future work.

11.2 Motivating Example: Incentive Design in Routing Games

Consider a transportation network $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ comprised of set of nodes $\mathcal{N}$ and set of edges $\mathcal{E}$, used by self-interested (infinitesimal)travelers. Each traveler is traveling between some origin-destination (o-d) pair on the network. The set of all o-d pairs be denoted by $\mathcal{Z}$. For each o-d pair $z \in \mathcal{Z}$, let $\mathcal{R}_z$ be the set of routes connecting the o-d pair z. Let D_z be the demand of travelers traveling between o-d pair $z \in \mathcal{Z}$ and y_{rz} be the flow of travelers from o-d pair $z \in \mathcal{Z}$ that choose route $r \in \mathcal{R}_z$. Naturally, $\sum_{r\in\mathcal{R}_z} y_{rz} = D_z$, for every $z \in \mathcal{Z}$. We denote the set of all feasible route flows by $Y = \prod_{z\in\mathcal{Z}} Y_z$, where $Y_z := D_z \cdot \Delta(\mathbb{R}^{|\mathcal{R}_z|})$ is a simplex. The route flow gives rise of congestion on the edges of the network. Given a route flow $y \in Y$, the resulting congestion on edges is denoted by $w(y) = (w_e(y))_{e\in\mathcal{E}}$, where

$$w_e(y) = \sum_{z\in\mathcal{Z}} \sum_{r\in\mathcal{R}_z} y_{rz}\mathbf{1}(e \in r), \quad \forall\, e \in \mathcal{E}. \tag{11.3}$$

Higher congestion leads to higher travel time on any edge. More formally, let $\ell_e(\cdot)$ be a strictly increasing smooth function which denotes the travel time of using edge $e \in \mathcal{E}$ as a function of congestion. A social planner can alter the congestion levels on the network by

imposing tolls on the edges of the network which changes the preferences of travelers for different routes. Let $x_e \in \mathbb{R}$ denote the tolls imposed on edge $e \in \mathcal{E}$. [1] Under the network tolls $x = (x_e)_{e\in\mathcal{E}} \in \mathbb{R}^{|\mathcal{E}|}$ and route flow $y \in Y$, the overall cost experienced by travelers from o-d pair $z \in \mathcal{Z}$ taking a route $r \in \mathcal{R}_z$ is

$$c_r(y, x) = \sum_{e\in r} \ell_e(w_e(y)) + x_e. \tag{11.4}$$

Given a fixed network tolls x, the resulting congestion – *Wardrop equilibrium* – can be obtained by solving the following strictly convex optimization problem ([331])

$$S(x) = \arg\min_{y\in Y} \Phi(y, x) = \sum_{e\in\mathcal{E}} \int_0^{w_e(y)} \ell_e(\theta)d\theta + x_e w_e(y). \tag{11.5}$$

Under the setting presented in this section, it can be verified that the set $S(x)$ exists, is singleton, and is Lipschitz continuous mapping [112].

The goal of social planner is to minimize the overall congestion on the network while also minimizing the tolls levied on travelers. More formally, the planner's objective function is given by $f(x, y) = \sum_{e\in\mathcal{E}} w_e(y)\ell_e(w_e(y)) + \lambda\|x\|^2$, where the first term corresponds to the average congestion on the network and second term is a regularization term with parameter $\lambda > 0$, which ensures low values of tolls[2]. Thus, the problem of toll design is as follows

$$\begin{aligned} \min_{x\in\mathbb{R}^{|\mathcal{E}|}, y\in Y} \quad & f(x, y) \\ \text{s.t.} \quad & y \in S(x) = \arg\min_{y'\in Y} \Phi(y, x), \end{aligned} \tag{11.6}$$

which is an instantiation of (SG).

Remark 11.2.1. *In order to compute $S(x)$ in* (11.5) *the planner needs to know the set Y that requires knowledge of the demand of travelers between various o-d pairs, which is a sensitive information. In Section 11.4, we use the proposed approach to solve* (11.6) *where the designer does not know the demand of travelers and can only observe the congestion levels $(w_e)_{e\in\mathcal{E}}$ on the network in response to the set tolls.*

11.3 Algorithms and Analysis

In this section, we present a follower agnostic algorithm for solving (SG). Following which, we present the convergence guarantees of the proposed algorithm to a stationary point. Additionally, we show that the algorithm will eventually converge to a local optima by avoiding the saddle points and local maximum.

[1] Here, we allow for tolls to take negative values. Such tolling scheme can be implemented by considering revenue-refunding schemes.

[2] Note that λ can in-general be zero, i.e. we do not require strong convexity of leader's objective function in its decision variable for our theoretical results to hold. We choose $\lambda > 0$ to impose a "soft-constraint" on the amount of tolls.

Algorithmic structure

The algorithm is based on alternatively moving towards solution to the variational inequality at lower level and descending along the upper-level objective function. Specifically, between two updates of leader (upper-level), the followers (lower level) employ an iterative adaptive rule, aimed to solve the variational inequality $\mathsf{SOL}(\cdot)$, for a fixed number of steps. Following which, the upper level iterates descend along an "approximated" gradient estimator, inspired from zeroth-order optimization ([395, 140]), evaluated at the lower-level iteration in current round.

Leader's strategy update The leader's update rule is as follows:

$$x_{t+1} = x_t - \eta_t \hat{F}(x_t; \delta_t, v_t), \tag{UL}$$

where $\hat{F}(x; \delta, v)$ denotes a gradient estimator of function $\tilde{f}(\cdot) := f(\cdot, \mathsf{br}(\cdot))$, evaluated at x with parameters δ, v. We shall describe the estimator in detail below.

Gradient estimator In order to compute the gradient of $\tilde{f}(x)$, we need to compute the derivative through the solution to the variational inequality in (SG), i.e. $S(x)$, which may involve higher order gradient computations and at times is not computable in closed form due to constraints. In this work, we consider a gradient estimator inspired from [395, 140]. Specifically, we consider the following estimator

$$\hat{F}(x; \delta, v) := \frac{d}{\delta} \left(f(\hat{x}, y^{(K)}(\hat{x})) - f(x, y^{(K)}(x)) \right) v, \tag{11.7}$$

such that *(i)* $\hat{x} = x + \delta v$, where $v \in \mathcal{S}(\mathbb{R}^d) := \{z \in \mathbb{R}^d : \|z\|_2 = 1\}$ and $\delta > 0$, are referred as *perturbation* and *perturbation radius* respectively; *(ii)* K is a positive integer capturing the number of rounds of adaptation rule employed by followers between two updates of leader's strategy; *(iii)* for any $x \in X$, $k \in [K-1]$ consider a iterative solver for variational inequality denoted by H such that

$$y^{(k+1)}(x) = H_k(y^{(k)}(x); x), \quad \forall\ k \in [K-1], \tag{LL}$$

where $y^{(0)}$ is some initialization for the iterative solver of variational inequality. For example, when the lower level problem is just a convex optimization problem with objective function $g(x, \cdot)$, a typical choice of H_k is projected gradient descent, i.e. $H_k(y; x) = \mathcal{P}_Y(y - \gamma_k \nabla_y g(x, y))$, where $\mathcal{P}_Y$ denotes the orthogonal projection on Y and γ_k is the step size. Note that, in order to construct the gradient estimator in (11.7), the leader *need not* know the exact description of update rule H_k. For most of the chapter, we shall concisely denote $y^{(k)}(x)$ and $y^{(k)}(\hat{x})$ as $\tilde{y}^{(k)}$ and $y^{(k)}$ respectively.

Remark 11.3.1. *Direct application of zeroth-order gradient estimator from [395, 140] would result in following estimator*

$$\tilde{F}(x;\delta,v) = \frac{d}{\delta}\left(\tilde{f}(\hat{x}) - \tilde{f}(x)\right)v, \tag{11.8}$$

where $\tilde{f}$ is defined in (11.2)*. Observe that the gradient estimators $\hat{F}$ and $\tilde{F}$ differ because in* (11.7) *we evaluate $f(x,\cdot)$ at $y^{(K)}(x)$ while in* (11.8) *we evaluated it at $\mathsf{br}(x)$ for any $x \in \mathbb{R}^d$. This induces additional bias in the gradient estimator that needs to be appropriately accounted while establishing convergence results.*

Algorithm The algorithms runs for T rounds. In every round $t \in [T-1]$ the leader queries the followers with two strategies x_t and $\hat{x}_t = x_t + \delta_t v_t$ where $v_t \sim \mathsf{Unif}(\mathcal{S}(\mathbb{R}^d))$ is a vector sampled uniformly randomly from the unit sphere in $\mathbb{R}^d$ and δ_t is the *perturbation radius* (refer line **2**-**3** in Algorithm 10). The followers respond to the leader's strategies by using

Algorithm 10 Follower Agnostic Stackelberg Optimization Algorithm

1: **Input:** Time horizon T, initial conditions $y_0^{(0)} \in Y$, $\tilde{y}_0^{(0)} \in Y$, $x_0 \in X$, step sizes (η_t), perturbation radius (δ_t)
2: **for** $t = 0$ to $T-1$ **do**
3: Sample $v_t \sim \mathsf{Unif}(\mathcal{S}(\mathbb{R}^d))$
4: Set $\hat{x}_t \leftarrow x_t + \delta_t v_t$
5: **for** $k = 0$ to $K-1$ **do**
6: Update $y_t^{(k+1)} \leftarrow H_k(y_t^{(k)}; \hat{x}_t)$
7: Update $\tilde{y}_t^{(k+1)} \leftarrow H_k(\tilde{y}_t^{(k)}; x_t)$
8: **end for**
9: Update

$$x_{t+1} \leftarrow x_t - \eta_t \cdot \frac{d}{\delta_t}\left(f(\hat{x}_t, y_t^{(K)}) - f(x_t, \tilde{y}_t^{(K)})\right)v_t$$

10: Set $y_{t+1}^{(0)} \leftarrow \tilde{y}_{t+1}^{(0)} \leftarrow \tilde{y}_t^{(K)}$
11: **end for**

an iterative variational inequality solver for K steps to obtain $\tilde{y}_t^{(K)}$ and $y_t^{(K)}$ respectively (refer line **4** and **7** in Algorithm 10). After observing $\tilde{y}_t^{(K)}$ and $y_t^{(K)}$, the leader computes a gradient estimator as per (11.7). The leader updates its strategy for next time as per (UL) (refer line **8** in Algorithm 10). The followers initialize their strategies as per line **9** in Algorithm 10.

Convergence to stationary points

We now study the convergence properties of Algorithm 10.

Assumption 11.3.1. *For any $x, \hat{x} \in X$, the updates in* (LL) *are such that $\|y^{(K)}(x) - y^{(K)}(\hat{x})\| \leqslant C\|x - \hat{x}\|$, for some $C > 0$.*

Assumption 11.3.1 posits that the adaptation rule employed by followers is stable with respect to perturbations in the leader's strategy. This assumption is typically satisfied by many algorithms, including gradient-based algorithms.

Assumption 11.3.2. *Atleast one of the following holds:*

(1a) For any $x \in X$, the iterates (LL) *converge to equilibrium at a polynomial rate. That is, for any initial point $y^{(0)} \in Y$,$\|y^{(K)}(x) - \mathsf{br}(x)\|^2 \leqslant CK^{-\lambda}\|y^{(0)} - \mathsf{br}(x)\|^2$, where λ, C are positive scalars.*

(1b) For any $x \in X$, the iterates (LL) *converge to equilibrium at a exponential rate. That is, for any initial point $y^{(0)}$, $\|y^{(K)}(x) - \mathsf{br}(x)\| \leqslant C\rho^K\|y^{(0)} - \mathsf{br}(x)\|$, where C is a positive scalar and $\rho \in [0, 1)$.*

Remark 11.3.2. *Convergence of lower-level problem is extensively studied in literature, e.g. [315, 430], and is not the focus of this chapter. Assumption 11.3.2(1a) holds for gradient descent updates for convex functions that satisfy* quadratic growth *condition [207]. Meanwhile, Assumption 11.3.2(1b) holds for gradient descent on strongly convex functions.*

Theorem 11.3.1. *Let Assumption 11.1.1-11.3.2 hold. If we choose*

$$\eta_t = \bar{\eta}(t+1)^{-1/2}d^{-1}, \delta_t = \bar{\delta}(t+1)^{-1/4}d^{-1/2}$$

such that $\bar{\eta} \leqslant d/2\tilde{\ell}$. Then the updates $(x_t)_{t\in[T]}$ in Algorithm 10 are such that

$$\min_{t\in[T]} \mathbb{E}\left[\|\nabla \tilde{f}(x_t)\|^2\right] \leqslant \tilde{O}\left(\frac{d}{\sqrt{T}} + \frac{\alpha}{1-\alpha}d^3\left(1 + \frac{1}{\sqrt{T}}\right)\right),$$

where $\alpha = CK^{-\lambda}$ if Assumption 11.3.2(1a) hold, or $\alpha = \rho^K$ if Assumption 11.3.2(1b) hold.

Intuitively, the theorem states that if we want to converge closer to a stationary point then we need to run the Algorithm 10 with larger T or smaller α (i.e. larger K). Crucially, the term αd^3 in the bound is due to error accumulation between time steps due to non-convergence of lower-level to exact solution of variational inequality $S(x)$. Owing to such precise characterization of error accumulation across time steps, our rate is informative of the *computational complexity* of solving the bi-level problem while in other contemporary work, namely [90], it resembles *iteration complexity* of the oracle. Since α is a function of K, the number of lower level iterations in every round, we can choose K to be large enough to make sure that the algorithm converges closer to the stationary point.

Corollary 11.3.1. *Let Assumption 11.1.1-11.3.1 and Assumption 11.3.2(1a) hold. Set $\eta_t = \bar{\eta}(t+1)^{-1/2}d^{-1}, \delta_t = \bar{\delta}(t+1)^{-1/4}d^{-1/2}$ such that $\bar{\eta} \leqslant d/2\tilde{\ell}$. Additionally, set $K \geqslant T^{1/2\lambda}d^{2/\lambda}$. Then, the iterates of Algorithm 10 satisfy $\min_{t\in[T]} \mathbb{E}\left[\|\nabla \tilde{f}(x_t)\|^2\right] \leqslant \tilde{O}\left(\frac{d}{\sqrt{T}}\right)$.*

Corollary 11.3.2. *Let Assumption 11.1.1-11.3.1 and Assumption 11.3.2(1b) hold. Set* $\eta_t = \bar{\eta}(t+1)^{-1/2}d^{-1}, \delta_t = \bar{\delta}(t+1)^{-1/4}d^{-1/2}$ *such that* $\bar{\eta} \leqslant d/2\tilde{\ell}$. *Additionally, set* $K \geqslant (1/|\log(\rho)|)\left((1/2)\log(T) + 2\log(d)\right)$. *Then,* $\min_{t\in[T]} \mathbb{E}\left[\|\nabla \tilde{f}(x_t)\|^2\right] \leqslant \tilde{O}\left(\frac{d}{\sqrt{T}}\right)$.

Remark 11.3.3. *We know that for non-convex smooth functions, gradient descent converges to a stationary point (at a rate of* $\mathcal{O}(1/\sqrt{T})$*). However, the key point of departure of* (UL) *from standard gradient descent is the presence of bias in the gradient estimator. Consequently, the key component of the proof is to bound the error in the gradient estimator* (11.7)*. This is because the estimator can be decomposed as*

$$\hat{F}(x_t; \delta_t, v_t) = \nabla \tilde{f}(x_t) + \mathcal{E}_t^{(1)} + \mathcal{E}_t^{(2)} + \mathcal{E}_t^{(3)},$$

where

$$\begin{aligned}
\mathcal{E}_t^{(1)} &:= \mathbb{E}\left[\tilde{F}(x_t; \delta_t, v_t)|x_t\right] - \nabla \tilde{f}(x_t), \\
\mathcal{E}_t^{(2)} &:= \tilde{F}(x_t; \delta_t, v_t) - \mathbb{E}\left[\tilde{F}(x_t; \delta_t, v_t)|x_t\right], \\
\mathcal{E}_t^{(3)} &:= \hat{F}(x_t; \delta_t, v_t) - \tilde{F}(x_t; \delta_t, v_t).
\end{aligned}$$

The term $\mathcal{E}_t^{(1)}$ *denotes the bias introduced due to the difference between standard zeroth-order gradient estimator, as per* (11.8)*, and the true gradient. The term* $\mathcal{E}_t^{(2)}$ *denotes the randomness introduced if we were to use the standard zeroth-order gradient estimator* (11.8)*. Finally, the term* $\mathcal{E}_t^{(3)}$ *denotes the bias introduced due to difference between our gradient estimator* (11.7) *and the standard zeroth-order gradient estimator (cf. Remark 11.3.1).*

Proof of Theorem 11.3.1 The proof of Theorem 11.3.1 follows by noting that $\tilde{f}$ approximately decreases along the trajectory (UL) (Lemma 11.3.1). Note that the decrease is said to be "approximate" because of the bias introduced by (11.7) in comparison to actual gradient $\nabla \tilde{f}(\cdot)$. We then proceed to individually bound the bias terms (Lemma 11.3.2). The convergence rate follows by using the step size and perturbation radius stated in the statement of Theorem 11.3.1.

Proof of Theorem 11.3.1. From Lemma 11.3.1 we know that $\tilde{f}(\cdot)$ approximately decreases along the trajectory of (UL). That is,

$$\begin{aligned}
\mathbb{E}\left[\tilde{f}(x_{t+1})\right] &\leqslant \mathbb{E}\left[\tilde{f}(x_t)\right] - \frac{\eta_t}{2}\mathbb{E}\left[\|\nabla \tilde{f}(x_t)\|^2\right] \\
&+ \eta_t \mathbb{E}\left[\|\mathcal{E}_t^{(1)}\|^2\right] + \eta_t \mathbb{E}\left[\|\mathcal{E}_t^{(3)}\|^2\right] + \tilde{\ell}\eta_t^2 \mathbb{E}\left[\|\mathcal{E}_t^{(2)}\|^2\right]. \qquad (11.9)
\end{aligned}$$

Using the bounds on error terms from Lemma 11.3.2, we obtain

$$\begin{aligned}\mathbb{E}\left[\tilde{f}(x_{t+1})\right] \leqslant \mathbb{E}\left[\tilde{f}(x_t)\right] - \frac{\eta_t}{2}\mathbb{E}\left[\|\nabla \tilde{f}(x_t)\|^2\right] + \eta_t \frac{\tilde{\ell}^2 \delta_t^2 d^2}{4} \\ + \eta_t \left(\frac{d^2}{\delta_t^2} L_2^2 \left(2\alpha^t e_0 + 2Cd^2 \sum_{k=0}^{t-1} \alpha^{t-k}\eta_k^2 + Cd \sum_{k=0}^{t-1} \alpha^{t-k}\delta_k^2\right)\right) \\ + 4d^2 \tilde{L}^2 \tilde{\ell} \eta_t^2.\end{aligned}$$

Re-arranging the terms and adding and substracting the term $\tilde{f}(x^*) = \min_{x\in X} \tilde{f}(x)$, we obtain

$$\begin{aligned}\frac{\eta_t}{2}\mathbb{E}\left[\|\nabla \tilde{f}(x_t)\|^2\right] \leqslant \mathbb{E}\left[\tilde{f}(x_t)\right] - \tilde{f}(x^*) + \eta_t \frac{\tilde{\ell}^2 \delta_t^2 d^2}{4} \\ + \eta_t \frac{d^2}{\delta_t^2} L_2^2 \left(2\alpha^t e_0 + 2Cd^2 \sum_{k=0}^{t-1} \alpha^{t-k}\eta_k^2 + C \sum_{k=0}^{t-1} \alpha^{t-k}\delta_k^2\right) \\ - \left(\mathbb{E}\left[\tilde{f}(x_{t+1})\right] - \tilde{f}(x^*)\right) + 4d^2 \tilde{L}^2 \tilde{\ell} \eta_t^2.\end{aligned}$$

Summing the previous equation over time step t we obtain

$$\begin{aligned}\sum_{t\in[T]} \eta_t \mathbb{E}\left[\|\nabla \tilde{f}(x_t)\|^2\right] \leqslant \left(\tilde{f}(x_0) - \tilde{f}(x^*)\right) + \frac{\tilde{\ell}^2 d^2}{4} \sum_{t\in[T]} \eta_t \delta_t^2 + 2e_0 d^2 L_2^2 \sum_{t\in[T]} \frac{\eta_t}{\delta_t^2}\alpha^t \\ + \underbrace{2Cd^4 L_2^2 \sum_{t\in[T]} \frac{\eta_t}{\delta_t^2} \sum_{k=0}^{t-1} \alpha^{t-k}\eta_k^2}_{\text{Term E}} + \underbrace{CL_2^2 d^2 \sum_{t\in[T]} \frac{\eta_t}{\delta_t^2} \sum_{k=0}^{t-1} \alpha^{t-k}\delta_k^2}_{\text{Term F}} + 4d^2 \tilde{L}^2 \tilde{\ell} \sum_{t\in[T]} \eta_t^2. \qquad (11.10)\end{aligned}$$

Setting $\eta_t = \bar{\eta}(t+1)^{-1/2}d^{-1}$ and $\delta_t = \bar{\delta}(t+1)^{-1/4}d^{-1/2}$, as per the statement of Theorem 11.3.1, and dividing both sides by $\sum_{t\in[T]} \eta_t$, we obtain

$$\begin{aligned}\frac{1}{\sum_{t\in[T]}\eta_t} \sum_{t\in[T]} \eta_t \mathbb{E}\left[\|\nabla \tilde{f}(x_t)\|^2\right] \leqslant \frac{Cd}{\bar{\eta}\sqrt{T}}\left(\tilde{f}(x_0) - \tilde{f}(x^*)\right) \\ + \frac{C\tilde{\ell} d \log(T)\bar{\delta}^2}{4\sqrt{T}} + \frac{2Cd^3 L_2^2 \alpha}{(1-\alpha)\bar{\eta}\sqrt{T}} + \frac{4C\tilde{L}^2\tilde{\ell}^2\bar{\eta}\log(T)d}{\sqrt{T}}, \\ + \frac{1}{\sum_{t\in[T]}\eta_t} \underbrace{2d^4 C L_2^2 \sum_{t\in[T]} \frac{\eta_t}{\delta_t^2} \sum_{k=0}^{t-1} \alpha^{t-k}\eta_k^2}_{\text{Term E}} + \frac{1}{\sum_{t\in[T]}\eta_t} \underbrace{L_2^2 C d^2 \sum_{t\in[T]} \frac{\eta_t}{\delta_t^2} \sum_{k=0}^{t-1} \alpha^{t-k}\delta_k^2}_{\text{Term F}},\end{aligned}$$

where C is a positive scalar. Next, we bound Term E + Term F as follows

$$\begin{aligned}
&\text{Term E } + \text{Term F} \leqslant \sum_{t=1}^{T} \frac{\eta_t}{\bar{\delta}_t^2} \sum_{k=0}^{t-1} \alpha^{t-k} \left(d^4 \eta_k^2 + d^2 \delta_k^2 \right) \\
&= \frac{\bar{\eta}}{\bar{\delta}^2} \sum_{t=1}^{T} \sum_{k=0}^{t-1} \alpha^t \frac{\Theta_k}{\alpha^k} = \frac{\bar{\eta}}{\bar{\delta}^2} \sum_{k=0}^{T-1} \frac{\Theta_k}{\alpha^k} \sum_{t=k+1}^{T} \alpha^t \\
&\leqslant \frac{\bar{\eta}}{\bar{\delta}^2} \sum_{k=0}^{T-1} \frac{\Theta_k}{\alpha^k} \frac{\alpha^{k+1}}{1-\alpha} = \frac{\bar{\eta}}{\bar{\delta}^2} \frac{\alpha}{1-\alpha} \sum_{k=0}^{T-1} \Theta_k \\
&= \frac{\bar{\eta}}{\bar{\delta}^2} \frac{C\alpha}{1-\alpha} \left(d^2 \bar{\eta}^2 \log(T) + d^2 \bar{\delta}^2 \sqrt{T} \right), \qquad (11.11)
\end{aligned}$$

where in second equality $\Theta_k := (d^4 \eta_k^2 + d^2 \delta_k^2)$, and we have appropriately adjusted the constant C to account for additinal constants. Thus, combining (11.10) and (11.11), we obtain

$$\begin{aligned}
&\frac{1}{\sum_{t\in[T]} \eta_t} \sum_{t\in[T]} \eta_t \mathbb{E}\left[\|\nabla \tilde{f}(x_t)\|^2 \right] \\
&\leqslant \mathcal{O}\Bigg(\frac{d}{\sqrt{T}} \left(\tilde{f}(x_0) - \tilde{f}(x^*) \right) + \frac{d \log(T)}{\sqrt{T}} + \frac{d^3 \alpha}{(1-\alpha)\sqrt{T}} \\
&\quad + \frac{4 \log(T) d}{\sqrt{T}} + \frac{d}{\sqrt{T}} \frac{\alpha}{1-\alpha} \left(d^2 \log(T) + d^2 \sqrt{T} \right) \Bigg).
\end{aligned}$$

To conclude, we obtain

$$\begin{aligned}
&\min_{t\in[T]} \mathbb{E}\left[\|\nabla \tilde{f}(x_t)\|^2 \right] \leqslant \frac{1}{\sum_{t\in[T]} \eta_t} \sum_{t\in[T]} \eta_t \mathbb{E}\left[\|\nabla \tilde{f}(x_t)\|^2 \right] \\
&\leqslant \tilde{O}\left(\frac{d}{\sqrt{T}} + \frac{\alpha}{1-\alpha} d^3 \left(1 + \frac{1}{\sqrt{T}} \right) \right).
\end{aligned}$$

This concludes the proof. □

Now, we formally state the Lemmas used in the proof.

Lemma 11.3.1. *If $\bar{\eta} \leqslant d/(2\tilde{\ell})$ then*

$$\begin{aligned}
&\mathbb{E}\left[\tilde{f}(x_{t+1}) \right] \leqslant \mathbb{E}\left[\tilde{f}(x_t) \right] - \frac{\eta_t}{2} \mathbb{E}\left[\|\nabla \tilde{f}(x_t)\|^2 \right] \\
&+ \eta_t \mathbb{E}\left[\|\mathcal{E}_t^{(1)}\|^2 \right] + \eta_t \mathbb{E}\left[\|\mathcal{E}_t^{(3)}\|^2 \right] + \tilde{\ell} \eta_t^2 \mathbb{E}\left[\|\mathcal{E}_t^{(2)}\|^2 \right].
\end{aligned} \qquad (11.12)$$

The proof of Lemma 11.3.1 follows in two steps: First, we use second-order Taylor series expansion of $\tilde{f}$ along the iterate values. Second, we use (UL) and complete the squares using algebraic manipulations. A detailed proof is provided in Appendix J.

Lemma 11.3.2. *The errors* $\mathbb{E}\left[\|\mathcal{E}_t^{(i)}\|^2\right]$ *for* $i \in \{1,2,3\}$ *are bounded as follows:*

$$\mathbb{E}\left[\|\mathcal{E}_t^{(1)}\|^2\right] \leqslant \frac{\tilde{\ell}^2\delta_t^2 d^2}{4}, \quad \mathbb{E}\left[\|\mathcal{E}_t^{(2)}\|^2\right] \leqslant 4d^2\tilde{L}^2,$$

$$\mathbb{E}\left[\|\mathcal{E}_t^{(3)}\|^2\right] \leqslant \frac{d^2}{\delta_t^2}L_2^2\bigg(2\alpha^t e_0 + 2C\sum_{k=0}^{t-1}\alpha^{t-k}\eta_k^2 + C\sum_{k=0}^{t-1}\alpha^{t-k}\delta_k^2\bigg), \tag{11.13}$$

where C *is a scalar and* $e_0 = \|y_0^{(0)} - S(x_0)\|^2$.

The stated bounds on $\mathbb{E}\left[\|\mathcal{E}_t^{(1)}\|^2\right]$ and $\mathbb{E}\left[\|\mathcal{E}_t^{(2)}\|^2\right]$ are inspired by the literature on two-point zeroth-order gradient estimators [395, 140]. We use the Lipschitz property of $f(x,\cdot)$ to bound

$$\|\mathcal{E}_t^{(3)}\|^2 \leqslant 2\frac{d^2}{\delta_t^2}L_2^2\Big(\underbrace{\|y_t^{(K)} - \mathsf{br}(\hat{x}_t)\|^2}_{\text{Term A}} + \underbrace{\|\tilde{y}_t^{(K)} - \mathsf{br}(x_t)\|^2}_{\text{Term B}}\Big).$$

Following which, Term A and Term B are recursively bounded. A detailed proof is provided in Appendix J.

Non-convergence to saddle points

In this section, we show that the updates in (UL) does not converge to a saddle point. Towards that goal, we make the following assumption that posits that the function $\tilde{f}(\cdot)$ satisfy the strict saddle property.

Assumption 11.3.3. *For any saddle point* x^* *of* $\tilde{f}$*, it holds that* $\lambda_{\min}(\nabla^2\tilde{f}(x^*)) < 0$.

In the following theorem, we formally state the non-convergence result.

Theorem 11.3.2. *Let Assumption 11.1.1-11.3.3 hold. For* $\epsilon > 0$ *there exists a time* T_ϵ *such that for any saddle point* x^* *of* $\tilde{f}$ *it holds that* $\mathbb{E}\left[\|x_t - x_*\|^2\right] \geqslant \epsilon, \forall\ t \geqslant T_\epsilon$.

To prove Theorem 11.3.2, an asymptotic pseudo-trajectory of (UL) is constructed. We then show that the asymptotic pseudo-trajectory almost surely avoids saddle point.

Proof of Theorem 11.3.2 The proof follows by contradiction. Suppose there exists a saddle point x^* such that $\lim_{t\to\infty}\mathbb{E}\left[\|x_t - x^*\|^2\right] = 0$. This implies that for any $\epsilon > 0$ there exists an integer T_ϵ such that for all $t \geqslant T_\epsilon$ it holds that

$$\mathbb{E}\left[\|x_{t+s} - x^*\|^2\right] \leqslant \epsilon/4 \quad \forall s \geqslant 0. \tag{11.14}$$

Next, for any arbitrary point x_t along the trajectory (UL), we define a dynamics parametrized by $\hat{x}_t = x_t + \delta_t v_t$, as follows $z_{s+1}(\hat{x}_t) := z_s(\hat{x}_t) - \eta_{t+s}\nabla\tilde{f}(z_s(\hat{x}_t))$,

where $z_0(\hat{x}_t) = \hat{x}_t$. From Lemma 11.3.3, we know that for any $\epsilon > 0$ and positive integer S there exists $\tilde{T}_{\epsilon,S}$ such that

$$\sup_{s\in[0,S]} \mathbb{E}\left[\|z_s(\hat{x}_t) - x_{t+s}\|^2\right] \leqslant \epsilon/4 \quad \forall\, t \geqslant \tilde{T}_{\epsilon,S}. \tag{11.15}$$

Next, note that $\|z_s(\hat{x}_t) - x^*\|^2 \leqslant 2\|z_s(\hat{x}_t) - x_{t+s}\|^2 + 2\|x_{t+s} - x^*\|^2$. Therefore, combining (11.14)-(11.15), we observe that for every $t \geqslant \max\{T_\epsilon, T_{\epsilon,S}\}, \mathbb{E}[\|z_s(\hat{x}_t) - x^*\|^2] \leqslant \epsilon, \quad \forall\, s \in [0, S]$.

But from [235] we know that for gradient descent with random initialization almost surely avoids converging to saddle point [3] there exists S_ϵ such that for all $s \geqslant S_\epsilon$ it holds that $\|z_s(\hat{x}_t) - x^*\|^2 \geqslant 2\epsilon$. This establishes contradiction.

Lemma 11.3.3. *Let x_t be an arbitrary point along the trajectory* (UL)*. Define a dynamics parametrized by $\hat{x}_t = x_t + \delta_t v_t$, such that $z_{s+1}(\hat{x}_t) := z_s(\hat{x}_t) - \eta_{t+s}\nabla\tilde{f}(z_s(\hat{x}_t))$, where $z_0(\hat{x}_t) = \hat{x}_t$ and it holds that for any positive integer L, we have*

$$\lim_{t\to\infty} \sup_{s\in[0,L]} \mathbb{E}\left[\|x_{t+s} - z_s(\hat{x}_t)\|^2\right] = 0.$$

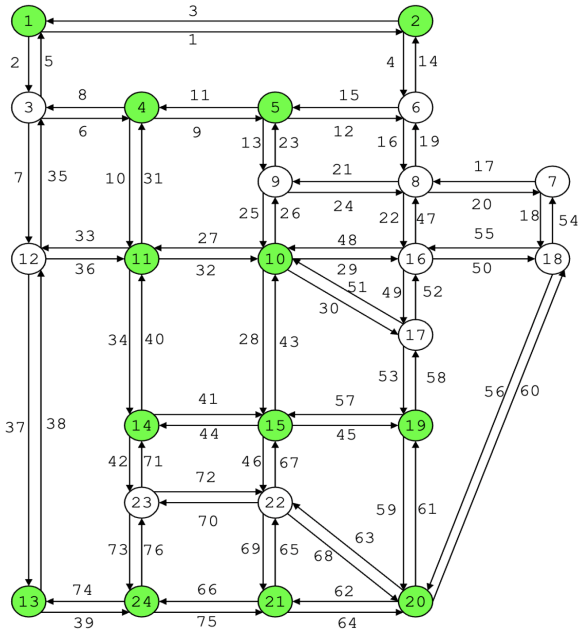
A detailed proof of Lemma 11.3.3 is provided in Appendix J.

11.4 Numerical Experiments

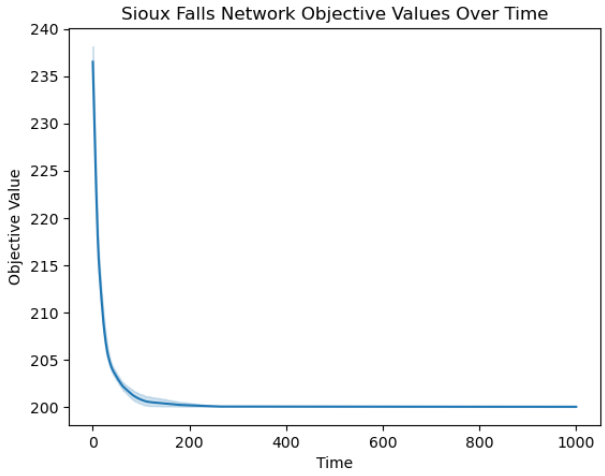
We numerically study the Algorithm 10 in the context of incentive design in routing games (described in Section 11.2). We consider the Sioux Falls transportation network, as depicted in Figure 11.1(a). The latency function and network topology are inherited from `http://tinyurl.com/y4fm4nvt`. We consider a synthetic demand of $(1, 2, 3, 2, 2, 1)$ units, respectively, between o-d pairs $\mathcal{Z} = ((1, 20), (13, 2), (20, 1), (10, 13), (11, 20), (4, 21))$.

The incentive designer can set tolls on each edge of the network. In response, unknown to the planner, the travelers alter their route selection as per a gradient rule. More formally, given a toll $x \in \mathbb{R}^{|\mathcal{E}|}$, we consider that the route choices made by the travelers are updated by descending along the gradient of the potential function $\Phi(\cdot, x)$ (cf. (11.5)). Note that, for any $x \in \mathbb{R}^{|\mathcal{E}|}, z \in \mathcal{Z}, r \in \mathcal{R}_z$, the gradient is $\frac{\partial\Phi(y,x)}{\partial y_{rz}} = \sum_{e\in\mathcal{E}} \ell_e(w_e(y))\frac{\partial w_e(y)}{\partial y_{rz}} + x_e\frac{\partial w_e(y)}{\partial y_{rz}} \underset{(i)}{=} \sum_{e\in\mathcal{E}} \ell_e(w_e(y))\mathbf{1}(e \in r) + x_e\mathbf{1}(e \in r) \underset{(ii)}{=} c_r(y, x)$, where (i) is due to (11.3) and (ii) follows from (11.4). Consequently, the gradient update takes the following form: for every $z \in \mathcal{Z}$, $y_z^{(k+1)} = \mathcal{P}_{Y_z}\left((y_{rz}^{(k)} - \gamma c_r(q^k, p))_{r\in\mathcal{R}_z}\right)$.

[3] More specifically, we use the results from [235, Proposition 8]. Even though the results in [235, Proposition 8] hold for gradient descent update with constant step-size, we can use this result for decaying step size in our context as well. This is because the proof of [235, Proposition 8] only requires each step of the gradient update to be diffeomorphism, which holds in our setting as the step-sizes are always non-negative and decaying.



(a) Schematic depiction of Sioux Falls transportation network. The numbers on the edges and nodes are identifiers.



(b) Evolution of planner's objective function. The shaded blue region denotes the confidence interval over 12 runs.

Figure 11.1: Simulation results on the Sioux Falls transportation network.

We simulate 12 runs of Algorithm 10 with $T = 1000$ and $K = 3$. The initial route flow vector $y_0^{(0)}$ and $\tilde{y}_0^{(0)}$ are randomly initialized. We set initial tolls uniformly randomly

between $[0, 0.1]$. We set the step size $\eta_t = 6(t+1)^{-1/2}$, $\delta_t = 0.3 \cdot (t+1)^{-1/4}$, $\gamma = 0.005$ and $\lambda = 0.01$. In Figure 11.1(b), we show the leader's objective value as function of time iterates $t \in [T]$. We observe that all trajectories converge to same objective value even with random initializations. This shows that the convergent point is perhaps a global optimizer.

11.5 Concluding Remarks

We propose an efficient algorithm for Stackelberg games which converges to a stationary point at a rate of $\mathcal{O}(T^{-1/2})$ and asymptotically reaches a local Stackelberg equilibrium. The algorithm is designed so that the leader does not need to know any information about the game structure at lower-level and updates its strategies by only querying for the followers response to its strategy.

Chapter 12

Externality-based Adaptive Incentive Design with Learning Agents

Similar to Chapter 11, in this chapter we study design of adaptive incentive mechanism that adjusts incentives based on the strategies of players, who repeatedly update their strategies as part of a learning process. This results in a *coupled* dynamical system that comprises both incentive and strategy updates.

The incentive mechanism studied here has four key features. Firstly, our framework applies to both atomic and non-atomic games. Secondly, the incentive update incorporates the *externality* generated by the players' current strategies, quantified as the difference between their own marginal cost and the marginal cost for the entire system. Thirdly, the incentive mechanism is agnostic to the strategy update dynamics used by players and requires only oracle access to either the gradient (in atomic games) or the value (in non-atomic games) of the cost function, given the current strategy, to evaluate the externality. Finally, the incentive update occurs on a *slower timescale* compared to the players' strategy updates. This slower evolution of incentives is a desirable characteristic because frequent incentive updates often hinder players' participation.

We prove that any fixed point of the coupled incentive and strategy updates leads to a socially optimal outcome. Specifically, at any fixed point, the incentive provided to each player equals the externality of the equilibrium strategy, ensuring that the resulting Nash equilibrium is socially optimal (Proposition 12.2.1). Additionally, we establish sufficient conditions on the underlying game that ensure the fixed point – coinciding with the socially optimal incentive mechanism – is unique (Proposition 12.2.1).

We characterize sufficient conditions for (both local and global) convergence of the coupled dynamical system to the fixed points (Proposition 12.2.3). Since the convergent strategy profile and incentive mechanism correspond to a socially optimal outcome, these sufficient conditions ensure that the coupled dynamical system induce a socially optimal outcome in the long run. Our analysis builds on the theory of two-timescale dynamical systems [59]. Due to the timescale separation between the strategy and incentive updates, we can decouple the convergence of the strategy update from that of the incentive update. First, the

convergence of the strategy update, which evolves on the faster timescale, is analyzed by treating the incentive mechanism, which evolves on the slower timescale, as static. In our work, we offload this analysis to the extensive literature on learning in games (e.g., [296, 239, 43, 361, 401]). Second, the convergence of the incentive mechanism update is examined through the corresponding continuous-time dynamical system, evaluated at the fixed point of the strategy update (i.e., the Nash equilibrium).

To demonstrate the usefulness of the adaptive incentive mechanism, we apply it to two practically relevant classes of games: *(i)* atomic aggregative games and *(ii)* non-atomic routing games. In atomic aggregative games, each player's cost function depends on their own strategy as well as the aggregate strategies of their opponents. This aggregation is performed through a linear combination of neighboring players' strategies, with weights characterized by a *network matrix*. Our proposed incentive mechanism enables the system operator to adaptively adjust incentives based on each player's externality on their neighbors while players learn their equilibrium strategies. When applied to this setting, our results provide sufficient conditions on the network matrix to ensure global convergence to a socially optimal outcome.

Furthermore, in non-atomic routing games, players (travelers) make routing decisions in a congested network with multiple origin-destination pairs. The system operator imposes incentives in the form of toll prices on network edges. Our proposed incentive mechanism is adaptively updated based solely on the observed edge flows and the gradient of the edge latency functions. Players can follow various strategy update rules that lead to the equilibrium of the routing game. We show that the adaptive incentive mechanism locally converges to the toll prices that minimize total congestion.

The chapter is organized as follows: In Section 12.1, we describe the setup for both atomic and non-atomic games and introduce the joint strategy and incentive update framework. Section 12.2 presents our results on the fixed points being socially optimal, and sufficient conditions for local and global convergence in general games. In Section 12.3, we apply these convergence results to atomic aggregative games (Section 12.3) and non-atomic routing games (Section 12.3). Finally, we conclude in Section 12.4.

Related Works

Two-timescale Learning Dynamics: Learning dynamics in which incentives are updated on a slower timescale than players' strategies have been studied in [302, 92, 343, 322, 253, 243, 11, 12]. Specifically, [302] examines Stackelberg games with a single leader and a population of followers, where the leader employs gradient-based updates while the followers adjust their strategies using replicator dynamics. Moreover, [92, 343] focus on incentive design in affine congestion games, where incentives are updated using a distributed version of gradient descent. Similarly, [322] studies incentive design for traffic control on a single highway through gradient-based incentive updates. Additionally, [253, 243] propose a two-timescale discrete-time learning dynamic in which players update their strategies using mirror descent, while the system operator adjusts the incentive parameter via a gradient-based method.

Furthermore, [11, 12] study the convergence of gradient-based incentive updates when the system operator has access to the gradient of the equilibrium strategy with respect to the incentive.

All of these works adopt gradient-based incentive updates. In such approaches, ensuring that the fixed point is socially optimal relies on the assumption that the equilibrium social cost is a convex function of the incentive parameter [302, 92, 343, 322, 253, 243] or that the gradient of the equilibrium strategy with respect to the incentive is non-singular [12, 11]. However, these assumptions are restrictive and often do not hold, even in simple games. In Chapter K.1, we provide a counterexample—a two-link routing game—in which both the convexity and non-singular gradient assumptions fail to hold.

Single-timescale Learning Dynamics: The problem of *steering* non-cooperative players toward a desired Nash equilibrium using an incentive update that operates on the same timescale as strategy updates has been studied in [378, 379, 444, 348]. Specifically, [378] examines such updates in the setting of quadratic aggregative games. In [379], the authors consider a scenario where players' costs depend only on their own actions and a price signal provided by an operator. In [444], the authors address the problem of guiding no-regret learning players toward an optimal equilibrium; however, their approach requires solving an optimization problem at each time step to compute the incentive mechanism. The work in [348] explores incentive design while simultaneously learning players' cost functions. The authors assume that both cost functions and incentive policies are linearly parameterized, with incentive updates relying on knowledge of the players' strategy update rules rather than solely on their current strategies, as in our setting.

Learning in Stackelberg Games: Our work is also related to the literature on learning in Stackelberg games, where the planner often has limited information about the interactions between players and must design an optimal mechanism by dynamically incorporating feedback from players' responses (see, e.g., [55, 241, 334, 25, 110, 192, 267]). This line of research typically imposes structural assumptions on the game among followers, such as a finite action space or linearly parameterized utility functions [55, 241, 334, 25, 110, 192]. Alternatively, some works, such as [267], focus on ensuring convergence only to a locally optimal solution.

Compared to the preceding three lines of research, we introduce a novel externality-based adaptive incentive design that applies to both atomic and nonatomic games, accommodates continuous action spaces, and allows for nonlinear utility functions. Unlike gradient-based incentive updates, externality-based updates ensure that any fixed point of the dynamics is socially optimal without requiring the equilibrium social cost function to be convex in the incentive vector or the gradient of equilibrium strategy with respect to the incentive to be non-singular. Furthermore, our incentive update is agnostic to the players' learning dynamics and relies only on oracle access to zeroth-order or first-order information about players' costs given their current strategies.

Notations

Given a function $f : \mathbb{R}^n \to \mathbb{R}$, we use $\nabla_{x_i} f(x)$ to denote the partial derivative of f with respect to x_i for any $i \in \{1, 2, ..., n\}$, and $\nabla f(x)$ to denote the gradient of the function. For any set A, we use $\mathsf{conv}(A)$ to denote its convex hull. For any set $X \subseteq \mathbb{R}^n$, a function $f : X \to \mathbb{R}$ is *Lipschitz* if there exists a positive scalar L such that $\|f(x) - f(x')\| \leqslant L\|x - x'\|$, for every $x, x' \in X$. For any vector $x \in X$ and any positive scalar $r > 0$, the set $\mathcal{B}_r(x) = \{x' \in X | \|x' - x\| < r\}$ denotes the r-radius neighborhood of the vector x. For any set X, we define $\mathsf{boundary}(X)$ and $\mathsf{int}(X)$ to be the boundary and interior of set X, respectively. Finally, for any function $f(\cdot)$, we denote the domain of the function by $\mathsf{dom}(f)$. For any vector $x \in \mathbb{R}^n$, we define $\mathsf{diag}(x) \in \mathbb{R}^{n \times n}$ to be a diagonal matrix with diagonal entries corresponding to x.

12.1 Model

We introduce both atomic and non-atomic static games in Sec. 12.1. In Sec. 12.1, we present our proposed adaptive incentive design approach.

Static Games

Atomic Games

Consider a game G with a finite set of players I. The strategy of each player $i \in I$ is denoted by $x_i \in X_i$, where X_i is a non-empty, closed interval in $\mathbb{R}$. The joint strategy profile of all players is given by $x = (x_i)_{i \in I}$, and the set of all joint strategy profiles is $X \coloneqq \prod_{i \in I} X_i$. The cost function of each player $i \in I$ is represented as $\ell_i : \mathbb{R}^{|I|} \to \mathbb{R}$.

A *system operator* designs incentives by setting a payment $p_i x_i \in \mathbb{R}$ for each player i, which is linear in their strategy x_i. Here, $p_i \in \mathbb{R}$ represents the marginal payment for every unit increase in the strategy of player i. The value of $p_i x_i$ can be either negative or positive, representing a marginal subsidy or a marginal tax, respectively. Given the incentive mechanism $p = (p_i)_{i \in I}$, the total cost for player $i \in I$ is:

$$c_i(x, p) = \ell_i(x) + p_i x_i, \quad \forall\, x \in X. \tag{12.1}$$

A strategy profile $x^*(p) \in X$ is a *Nash equilibrium* in the atomic game G with the incentive mechanism p if

$$c_i(x_i^*(p), x_{-i}^*(p), p) \leqslant c_i(x_i, x_{-i}^*(p), p),\ \forall\, x_i \in X_i,\ \forall i \in I.$$

A strategy profile $x^\dagger \in X$ is *socially optimal* if it minimizes the social cost function $\Phi : \mathbb{R}^{|I|} \to \mathbb{R}$ over X.

Assumption 12.1.1. *For any $p \in \mathbb{R}^{|I|}$, the Nash equilibrium $x^*(p)$ is unique and Lipschitz continuous in p. Moreover, the social cost function $\Phi(x)$ is continuously differentiable, has a Lipschitz gradient, and is strictly convex in x.*

Assumption 12.1.1 is widely adopted in the literature to study incentive design in atomic games (e.g., [245, 253, 243, 378]), either directly or through other conditions that guarantee this[1].

Non-atomic Games

Consider a game $\tilde{G}$ with a finite set of player populations $\tilde{I}$. Each population $i \in \tilde{I}$ is comprised of a continuum set of (infinitesimal) players with mass $\tilde{M}_i > 0$. Every (infinitesimal) player in any population can choose an action in a finite set $\tilde{S}_i$. The strategy distribution of population $i \in \tilde{I}$ is $\tilde{x}_i = (\tilde{x}_i^j)_{j\in\tilde{S}_i}$, where $\tilde{x}_i^j$ is the mass of players in population i who choose action $j \in \tilde{S}_i$. Then, the set of all strategy distributions of population i is $\tilde{X}_i = \left\{\tilde{x}_i \in \mathbb{R}^{|\tilde{S}_i|} | \sum_{j\in\tilde{S}_i} \tilde{x}_i^j = \tilde{M}_i,\ \tilde{x}_i^j \geqslant 0, \forall j \in \tilde{S}_i\right\}$. The strategy distribution of all populations is given by $\tilde{x} = (\tilde{x}_i)_{i\in\tilde{I}} \in \tilde{X} = \prod_{i\in\tilde{I}} \tilde{X}_i$. We define $\tilde{S} = \prod_{i\in\tilde{I}} \tilde{S}_i$. Given a strategy distribution $\tilde{x} \in \tilde{X}$, the cost of players in population $i \in \tilde{I}$ for choosing action $j \in \tilde{S}_i$ is $\tilde{\ell}_i^j(\tilde{x})$. We denote $\tilde{\ell}_i(\tilde{x}) = (\tilde{\ell}_i^j(\tilde{x}))_{j\in\tilde{S}_i}$ as the vector of costs for each $i \in \tilde{I}$.

A system operator designs incentives by setting a payment $\tilde{p}_i^j$ for players in population i who choose action $j \in \tilde{S}_i$. Consequently, given the incentive mechanism $\tilde{p} = (\tilde{p}_i^j)_{j\in S_i, i\in\tilde{I}}$, the total cost experienced by any player in population $i \in \tilde{I}$ who chooses action $j \in \tilde{S}_i$ is

$$\tilde{c}_i^j(\tilde{x}, \tilde{p}) = \tilde{\ell}_i^j(\tilde{x}) + \tilde{p}_i^j, \quad \forall\ \tilde{x} \in \tilde{X}. \tag{12.2}$$

A strategy distribution $\tilde{x}^*(\tilde{p}) \in \tilde{X}$ is a Nash equilibrium in the non-atomic game $\tilde{G}$ with $\tilde{p}$ if

$$\begin{aligned} \tilde{x}_i^{j*}(\tilde{p}) > 0, \quad \Rightarrow \tilde{c}_i^j(\tilde{x}^*(\tilde{p}), \tilde{p}) \leqslant \tilde{c}_i^{j'}(\tilde{x}^*(\tilde{p}), \tilde{p}), \\ \forall j, j' \in \tilde{S}_i,\ \forall i \in \tilde{I}. \end{aligned} \tag{12.3}$$

A strategy distribution $\tilde{x}^\dagger \in \tilde{X}$ is socially optimal if $\tilde{x}^\dagger$ minimizes a social cost function $\tilde{\Phi} : \mathbb{R}^{|\tilde{S}|} \to \mathbb{R}$.

Assumption 12.1.1. *For any $p \in \mathbb{R}^{|\tilde{I}|}$, the Nash equilibrium $\tilde{x}^*(\tilde{p})$ is unique and Lipschitz continuous in $\tilde{p}$. Moreover, $\tilde{\Phi}(\tilde{x})$ is continuously differentiable and strictly convex.*

Assumption 12.1.1 is widely adopted in the literature on incentive design for non-atomic games (e.g., [253, 322, 302]), either directly or through other conditions that guarantee this[2].

[1] Uniqueness and Lipschitz continuity of $x^*(p)$ hold if, for every $i \in I$ and $x_{-i} = (x_j)_{j\in I\setminus\{i\}}$, the cost function $\ell_i(x_i, x_{-i})$ is strongly convex in x_i and $\ell_i(\cdot)$ is continuously differentiable with a Lipschitz gradient [112].

[2] Uniqueness and Lipschitz continuity of $x^*(p)$ hold if $\tilde{\ell}(\cdot)$ is Lipschitz continuous and strongly monotone [361]. That is, there exists $\rho > 0$ such that $\langle \tilde{\ell}(\tilde{x}) - \tilde{\ell}(\tilde{x}'), \tilde{x} - \tilde{x}' \rangle \geqslant \rho \|\tilde{x} - \tilde{x}'\|^2$ for every $\tilde{x} \neq \tilde{x}' \in \tilde{X}$.

Coupled Strategy and Incentive Update

We consider a coupled dynamical system that jointly updates players' strategies and the incentive mechanism with discrete time-steps $k \in \mathbb{N}$. At step k, the strategy profile in the atomic game G (resp. non-atomic game $\tilde{\mathcal{G}}$) is $x_k = (x_{i,k})_{i\in I}$ (resp. $\tilde{x}_k = (\tilde{x}_{i,k})_{i\in\tilde{I}}$), where $x_{i,k}$ (resp. $\tilde{x}_{i,k}$) is the strategy of player i (population i), and the incentive mechanism is $p_k = (p_{i,k})_{i\in I}$ (resp. $\tilde{p}_k = (\tilde{p}^j_{i,k})_{j\in S_i, i\in\tilde{I}}$). The strategy updates and the incentive updates are presented below:

Strategy update.

$$x_{k+1} = (1-\gamma_k)x_k + \gamma_k f(x_k, p_k), \qquad (x\text{-update})$$
$$\tilde{x}_{k+1} = (1-\gamma_k)\tilde{x}_k + \gamma_k \tilde{f}(\tilde{x}_k, \tilde{p}_k). \qquad (\tilde{x}\text{-update})$$

In each step $k+1$, the updated strategy is a linear combination of the strategy in stage k (i.e. x_k in G and $\tilde{x}_k$ in $\tilde{\mathcal{G}}$), and a new strategy $f(x_k, p_k) \in \mathcal{X}$ in G (resp. $\tilde{f}(\tilde{x}_k, \tilde{p}_k) \in \tilde{X}$ in $\tilde{\mathcal{G}}$) that depends on the previous strategy and the incentive mechanism. The relative weight in the linear combination is determined by the step-size $\gamma_k \in (0,1)$. We require that for any p (resp. $\tilde{p}$), the fixed point associated with update (x-update) (resp. ($\tilde{x}$-update)) is a Nash equilibrium, i.e.

$$\begin{aligned} x^*(p) &= \{x : f(x,p) = x\}, \quad \forall p \in \mathbb{R}^{|\mathcal{I}|}, \\ \tilde{x}^*(\tilde{p}) &= \{\tilde{x} : \tilde{f}(\tilde{x},\tilde{p}) = \tilde{x}\}, \quad \forall\ \tilde{p} \in \mathbb{R}^{|\tilde{I}|}. \end{aligned} \tag{12.4}$$

We shall impose additional assumptions on $f(\cdot)$ and $\tilde{f}(\cdot)$ when studying the convergence of strategy and incentive updates in the next section. Some examples of commonly studied learning dynamics (x-update) and ($\tilde{x}$-update) include:

1. *Equilibrium update ([110, 192]):* The strategy update incorporates a Nash equilibrium strategy profile with respect to the incentive mechanism in step k:

$$f(x_k, p_k) = x^*(p_k), \quad \text{and } \tilde{f}(\tilde{x}_k, \tilde{p}_k) = \tilde{x}^*(\tilde{p}_k). \tag{12.5}$$

2. *Best response update ([149, 361]):* The strategy update incorporates a best response strategy with respect to the strategy and the incentive mechanism in step k:

$$\begin{aligned} f_i(x_k, p_k) &= \arg\min_{y_i \in X_i} c_i(y_i, x_{-i,k}, p_k), \\ \tilde{f}_i(\tilde{x}_k, \tilde{p}_k) &= \arg\min_{\tilde{y}_i \in \tilde{X}_i} \tilde{y}_i^\top \tilde{c}_i(\tilde{x}_k, \tilde{p}_k), \end{aligned} \tag{12.6}$$

   where the first equation is the best response update in atomic games [149], and the second is the best response update in non-atomic games [361].

3. *Gradient-based update ([251, 226, 361])*: Gradient-based strategy update commonly studied in literature takes the following form:

$$\begin{aligned} f_i(x_k, p_k) &= \arg\max_{y_i \in X_i} \; z_i(x_k, p_k) y_i - h(y_i), \\ \tilde{f}_i(\tilde{x}_k, \tilde{p}_k) &= \arg\max_{\tilde{y}_i \in \tilde{X}_i} \; \tilde{y}_i^\top \tilde{c}_i(\tilde{x}_k, \tilde{p}_k) - \tilde{h}(\tilde{y}_i), \end{aligned} \tag{12.7}$$

where $z_i(x_k, p_k) = x_k - \eta \nabla_{x_i} c_i(x_k, p_k)$, η is step size, and $h(\cdot), \tilde{h}(\cdot)$ are regularizers. If $h(\cdot)$ is a quadratic function, then the update becomes projected gradient descent [298]. Furthermore, if $\tilde{h}(\cdot)$ is the entropy function, then the update becomes a perturbed best-response update[361].

Incentive update.

$$p_{k+1} = (1 - \beta_k) p_k + \beta_k e(x_k), \tag{p-update}$$
$$\tilde{p}_{k+1} = (1 - \beta_k) \tilde{p}_k + \beta_k \tilde{e}(\tilde{x}_k), \tag{$\tilde{p}$-update}$$

where $e(x) = (e_i(x))_{i \in I}$, $\tilde{e}(\tilde{x}) = (\tilde{e}_i^j(\tilde{x}))_{j \in S_i, i \in \tilde{I}}$, and

$$e_i(x) = \nabla_{x_i} \Phi(x) - \nabla_{x_i} \ell_i(x), \quad \forall i \in I, \tag{12.8a}$$
$$\tilde{e}_i^j(\tilde{x}) = \nabla_{\tilde{x}_i^j} \tilde{\Phi}(\tilde{x}) - \tilde{\ell}_i^j(\tilde{x}), \quad \forall j \in \tilde{S}_i, \quad \forall i \in \tilde{I}. \tag{12.8b}$$

In (12.8a), $e_i(x)$ represents the difference between the marginal social cost and the marginal cost of player i given x. Similarly, $\tilde{e}_i^j(\tilde{x})$ denotes the difference between the marginal social cost and the cost experienced by players in population i who choose action j. We refer to $e_i(x)$ and $\tilde{e}_i(\tilde{x}) = (\tilde{e}_i^j(\tilde{x}))_{j \in S_i}$ as the externalities of players i and population i, respectively, since they capture the difference in the impact of their strategies on the social cost and individual cost.

The updates (p-update)-($\tilde{p}$-update) modify the incentives on the basis of the externality caused by the players. In each step $k+1$, the updated incentive mechanism is a linear combination of the incentive mechanism in step k (i.e. p_k in G and $\tilde{p}_k$ in $\tilde{\mathcal{G}}$), and the externality (i.e. $e(x_k)$ in G and $\tilde{e}(\tilde{x}_k)$ in $\tilde{\mathcal{G}}$) given the strategy in step k. The relative weight in the linear combination is determined by the step size $\beta_k \in (0, 1)$.

In summary, the joint evolution of strategy and incentive mechanism $(x_k, p_k)_{k=1}^\infty$ (resp. $(\tilde{x}_k, \tilde{p}_k)_{k=1}^\infty$) in the atomic game G (resp. non-atomic game $\tilde{\mathcal{G}}$) is governed by the learning dynamics (x-update)–(p-update) (resp. ($\tilde{x}$-update)–($\tilde{p}$-update)). The step-sizes $(\gamma_k)_{k=1}^\infty$ and $(\beta_k)_{k=1}^\infty$ determine the speed of strategy updates and incentive updates, respectively.

Information environment of incentive update. The incentive updates in (p-update) and ($\tilde{p}$-update) are based on the externalities created by players' strategies. In the absence of additional problem structure, computing externality requires oracle access to the gradient of players' costs (first-order information) in atomic games (cf. (12.8a)) or the players' cost

functions (zeroth-order information) in non-atomic games (cf. (12.8b)), both evaluated at the current strategy profile. This information requirement is less demanding compared to the gradient-based incentive updates adopted in previous literature [302, 92, 343, 322, 253, 243, 11, 12], where estimating the gradient of the social cost function with respect to the incentive vector often requires the knowledge of the game Jacobian (i.e., second-order information) [3] [253, 243], or knowledge of the gradient of equilibrium strategy with respect to incentive (i.e., $\nabla_p x^*(p)$) [12, 11]. Furthermore, our incentive updates do not require knowledge of players' entire cost function, and are agnostic to the specific strategy update dynamics (i.e., (x-update) and ($\tilde{x}$-update)) employed by the players.

In many settings, leveraging the structure of the underlying problem enables the social planner to compute externalities with less information. For instance, in non-atomic routing games (see Section 12.3), the social planner can compute externality using only the travel time costs of edges (road segments in the network) instead of the cost of each path taken by each population given their origin-destination pair. Additionally, in energy system applications (e.g., [245]), player's cost function $\ell_i(x) = g_i(x_i)$ often only depends on their own energy consumption x_i, and the social cost function $\Phi(x) = r(x) + \sum_{i\in\mathcal{I}} g_i(x_i)$ is modeled as the sum of the public cost $r(x)$ that depends on the joint action x and the cost of individual players. In this case, the externality for any player i depends only on the gradient of the public cost function $r(x)$ and not on the private cost of players.

$$e_i(x) = \frac{\partial \Phi(x)}{\partial x_i} - \frac{\partial \ell_i(x)}{\partial x_i} = \frac{\partial r(x)}{\partial x_i}.$$

12.2 General results

In Section 12.2, we characterize the set of fixed points of the updates (x-update)-(p-update) and ($\tilde{x}$-update)-($\tilde{p}$-update), and show that any fixed point corresponds to a socially optimal incentive mechanism such that the induced Nash equilibrium strategy profile minimizes the social cost. In Section 12.2, we provide a set of sufficient conditions that guarantee (local and global) convergence of incentive updates. Under these conditions, our adaptive incentive mechanism eventually induces a socially optimal outcome.

[3] For instance, the gradient based incentive update of atomic games studied in [253] takes the following form $p_{k+1} = p_k - \beta_k \nabla_p x^*(p_k)^\top \nabla_x \Phi(x^*(p_k))$, which is a gradient descent update on the function $\Phi(x^*(p))$. The authors estimate $\nabla_p x^*(p_k)^\top$ with $-\nabla_p J(x_k; p_k)^\top (\nabla_x J(x_k; p_k))^{-1}$, where $J(x; p) = (\partial c_i(x, p)/\partial x_i)_{i\in\mathcal{I}}$ is the *game Jacobian*. Therefore, these updates require second order information about the cost function of players. Meanwhile, our approach of externality based pricing only requires first-order information about the cost function of players.

Fixed point analysis

We first characterize the set of fixed points of the updates (x-update)-(p-update), and ($\tilde{x}$-update)-($\tilde{p}$-update) as follows:

$$\text{Atomic game } G,\ \{(x,p)|f(x,p)=x,\ e(x)=p\}\,, \tag{12.9a}$$

$$\text{Non-atomic game } \tilde{\mathcal{G}},\ \left\{(\tilde{x},\tilde{p})|\tilde{f}(\tilde{x},\tilde{p})=\tilde{x},\ \tilde{e}(\tilde{x})=\tilde{p}\right\}. \tag{12.9b}$$

Using (12.4), from (12.9a) – (12.9b), we can write the set of incentive mechanisms at the fixed point, $P^\dagger$ as follows:

$$\begin{aligned}&\text{Atomic game } G,\ P^\dagger=\{(p_i^\dagger)_{i\in I}|e(x^*(p^\dagger))=p^\dagger\},\\ &\text{Non-atomic game } \tilde{\mathcal{G}},\ \tilde{P}^\dagger=\{(\tilde{p}_i^\dagger)_{i\in\tilde{I}}|\tilde{e}(\tilde{x}^*(\tilde{p}^\dagger))=\tilde{p}^\dagger\}.\end{aligned} \tag{12.10}$$

That is, at any fixed point, the incentive of each player is set to be equal to the externality evaluated at their equilibrium strategy profile.

Our first result characterizes conditions under which the fixed point set $P^\dagger$ (resp. $\tilde{P}^\dagger$) is non-empty and singleton in G (resp. $\tilde{\mathcal{G}}$). Moreover, given any fixed point incentive mechanism $p^\dagger\in P^\dagger$ and $\tilde{p}^\dagger\in\tilde{P}^\dagger$, the corresponding Nash equilibrium is socially optimal.

Proposition 12.2.1. *Let Assumptions 12.1.1 hold and the strategy set X in an atomic game G be compact. The set $P^\dagger$ is a non-empty singleton set. The unique $p^\dagger\in P^\dagger$ is socially optimal, i.e. $x^*(p^\dagger)=x^\dagger$.*

Moreover, in a non-atomic game $\tilde{G}$ under Assumptions 12.1.1, $\tilde{P}^\dagger$ is a non-empty singleton set. The unique $\tilde{p}^\dagger\in\tilde{P}^\dagger$ is socially optimal, i.e., $\tilde{x}^*(\tilde{p}^\dagger)=\tilde{x}^\dagger$.

Advantage of externality-based incentive updates. Proposition 12.2.1 demonstrates that the externality-based incentive updates (p-update) and ($\tilde{p}$-update) ensure that any fixed point must achieve social optimality. In contrast, the gradient-based incentive update, commonly considered in the literature (e.g., [253, 243, 302, 12, 11]), does not guarantee that its fixed point corresponds to a socially optimal incentive mechanism. Typically, these works impose additional assumptions, such as the equilibrium social cost function $\mathbf{\Phi}(x^*(p))$ (resp. $\tilde{\mathbf{\Phi}}(\tilde{x}^*(\tilde{p}))$) being strongly convex in the incentive mechanism p (resp. $\tilde{p}$) [253, 243, 302], or that the gradient of the equilibrium strategy, $\nabla_p x^*(p)$ (resp. $\nabla_{\tilde{p}}\tilde{x}^*(\tilde{p})$), with respect to the incentive mechanism p (resp. $\tilde{p}$) is non-singular [12, 11], to ensure that the fixed points of the gradient-based update achieve the socially optimal outcome. In fact, in Chapter K.1, we provide an example of a two-link non-atomic routing game, where these assumptions are not satisfied and nearly all fixed points of the gradient-based incentive update fail to achieve the socially optimal outcome. Consequently, the gradient-based incentive update can lead to inefficient outcomes. In contrast, our externality-based incentive update has a unique fixed point that always induces the socially optimal outcome.

Proof of Proposition 12.2.1. First, we show that $P^\dagger$ is non-empty, i.e., there exists $p^\dagger$ such that $e(x^*(p^\dagger))=p^\dagger$. Define the function $\theta(p)=e(x^*(p))$. By Assumption 12.1.1, θ is well-defined. Thus, the problem reduces to proving the existence of a solution to $p=\theta(p)$.

We note that Assumption 12.1.1 ensures that $\theta(p)$ is a continuous function. Now, define $K \coloneqq \{\theta(p) : p \in \mathbb{R}^{|I|}\} \subseteq \mathbb{R}^{|I|}$. We claim that the set K is compact. Indeed, this follows from two observations. First, the externality function $e(\cdot)$ is continuous. Second, the range of the function $x^*(\cdot)$ is X, which is a compact set. These two observations ensure that $\theta(p) = e(x^*(p))$ is a bounded function. Let $\tilde{K} \coloneqq \mathsf{conv}(K)$ be the convex hull of K, which in turn is also a compact set. Let's denote the restriction of function θ on the set $\tilde{K}$ as $\theta_{|\tilde{K}} : \tilde{K} \to \tilde{K}$ where $\theta_{|\tilde{K}}(p) = \theta(p)$ for all $p \in \tilde{K}$. We note that $\theta_{|\tilde{K}}$ is a continuous function from a convex compact set to itself and therefore, the Schauder fixed point theorem ensures that there exists $p^\dagger \in \tilde{K}$ such that $p^\dagger = \theta_{|\tilde{K}}(p^\dagger) = \theta(p^\dagger)$ [388]. This concludes the proof of the existence of $p^\dagger$. Analogous argument applies for the non-atomic game $\tilde{G}$ to show that $\tilde{P}^\dagger$ is non-empty.

Next, we show that the incentive $p^\dagger$ aligns the Nash equilibrium with socially optimal strategy (i.e. for any $p^\dagger \in P^\dagger$, $x^*(p^\dagger) = x^\dagger$). For any $p^\dagger \in P^\dagger$ and any $i \in I$, it holds that $p_i^\dagger = e_i(x^*(p^\dagger))$. This implies that $\nabla_{x_i}\ell_i(x^*(p^\dagger)) + p_i^\dagger = \nabla_{x_i}\Phi(x^*(p^\dagger))$ for every $i \in I$, and thus

$$J(x^*(p^\dagger), p^\dagger) = \nabla\Phi(x^*(p^\dagger)), \tag{12.11}$$

where $J(x, p)$ is the *game Jacobian* defined as $J_i(x, p) = \nabla_{x_i}\ell_i(x) + p_i$ for every $i \in I$. From Assumption 12.1.1 and the first order necessary condition for Nash equilibrium [135], we know that the Nash equilibrium $x^*(p^\dagger)$ must satisfy

$$\langle J(x^*(p^\dagger), p^\dagger), x - x^*(p^\dagger)\rangle \geqslant 0, \quad \forall\ x \in X. \tag{12.12}$$

From (12.11) and (12.12), we observe that

$$\langle \nabla\Phi(x^*(p^\dagger)), x - x^*(p^\dagger)\rangle \geqslant 0, \quad \forall\ x \in X. \tag{12.13}$$

Further, from the first order conditions of optimality for social cost function we know that $x^\dagger$ is socially optimal if any only if it satisfies

$$\langle \nabla\Phi(x^\dagger), x - x^\dagger\rangle \geqslant 0, \quad \forall\ x \in X. \tag{12.14}$$

Comparing (12.13) with (12.14), we note that $x^*(p^\dagger)$ is the minimizer of social cost function Φ. This implies that $x^*(p^\dagger) = x^\dagger$, since $x^\dagger$ is the unique minimizer of the social cost function Φ under Assumption 12.1.1.

Similarly, for non-atomic game $\tilde{G}$, we show that the incentive $\tilde{p}^\dagger$ aligns the Nash equilibrium with social optimality. Fix $\tilde{p}^\dagger \in \tilde{P}^\dagger$. For every $j \in \tilde{S}_i$ and $i \in \tilde{I}$, it holds that $\tilde{p}_i^{j\dagger} = \tilde{e}_i^j(\tilde{x}^*(\tilde{p}^\dagger))$. Consequently,

$$\tilde{c}_i^j(\tilde{x}^*(\tilde{p}^\dagger), \tilde{p}^\dagger) = \nabla_{\tilde{x}_i^j}\tilde{\Phi}(\tilde{x}^*(\tilde{p}^\dagger)). \tag{12.15}$$

Under Assumption 12.1.1, $\tilde{x}^*(\tilde{p}^\dagger)$ is a Nash equilibrium only if

$$\langle \tilde{c}(\tilde{x}^*(\tilde{p}^\dagger), \tilde{p}^\dagger), \tilde{x} - \tilde{x}^*(\tilde{p}^\dagger)\rangle \geqslant 0, \quad \forall\ \tilde{x} \in \tilde{X}. \tag{12.16}$$

From (12.15) and (12.16), we observe that

$$\langle \nabla \tilde{\Phi}(\tilde{x}^*(\tilde{p}^\dagger)), \tilde{x} - \tilde{x}^*(\tilde{p}^\dagger) \rangle \geqslant 0, \quad \forall\ \tilde{x} \in \tilde{X}. \tag{12.17}$$

Comparing (12.17) with the first order necessary and sufficient conditions of optimality of social cost function, we note that $\tilde{x}^*(\tilde{p}^\dagger)$ is the minimizer of the social cost function $\tilde{\Phi}$. This implies that $\tilde{x}^*(\tilde{p}^\dagger) = \tilde{x}^\dagger$, since $\tilde{x}^\dagger$ is the unique minimizer of the social cost function $\tilde{\Phi}$ under Assumption 12.1.1.

Finally, we show that the set $P^\dagger$ is singleton. We prove this via contradiction. Suppose that $P^\dagger$ contains two element $p_1^\dagger, p_2^\dagger$, and both align the Nash equilibrium with social optimality. Then, $x^\dagger = x^*(p_1^\dagger) = x^*(p_2^\dagger)$. From (12.10), we know that $p_1^\dagger = e(x^*(p_1^\dagger))$ and $p_2^\dagger = e(x^*(p_2^\dagger))$. Thus, we must have $p_1^\dagger = e(x^\dagger) = p_2^\dagger$, which implies that $P^\dagger$ is a singleton. The proof of uniqueness of $\tilde{P}^\dagger$ follows analogously. □

Convergence to optimal incentive mechanism

In this subsection, we provide sufficient conditions for the convergence of strategy and incentive updates (x-update)-(p-update) and ($\tilde{x}$-update)-($\tilde{p}$-update). Before presenting the convergence result, we first introduce two assumptions.

Assumption 12.2.1. *The step sizes in* (x-update)-(p-update) *and* ($\tilde{x}$-update)-($\tilde{p}$-update) *satisfy the following conditions:*

(i) $\sum_{k=1}^{\infty} \gamma_k = \sum_{k=1}^{\infty} \beta_k = +\infty$, $\sum_{k=1}^{\infty} \gamma_k^2 + \beta_k^2 < +\infty$.

(ii) $\lim_{k\to\infty} \beta_k / \gamma_k = 0$.

Assumption 12.2.1-(i) is a standard assumption on step sizes that allows us to analyze the convergence properties of the discrete-time learning updates through that of a continuous-time dynamical system [60]. Assumption 12.2.1-(ii) ensures that the incentive update evolves on a slower timescale than the players' strategy updates [60, 225]. Any step sizes of the form $\gamma_k = k^{-a}$ and $\beta_k = k^{-b}$ with $0.5 < a < b \leqslant 1$, satisfy Assumption 12.2.1.

Assumption 12.2.1 has been adopted in several previous works on adaptive incentive design (e.g., [245, 89]). Under Assumption 12.2.1, the strategy update (x-update) represents a *fast transient*, whereas the incentive update (p-update) is a *slow component*. To analyze such discrete-time updates, we employ techniques from two-timescale approximation theory [60, 61, 89], which allows us to analyze the convergence of the strategy and incentive updates separately. An intermediate step in this process is to ensure that, for every $p, \tilde{p}$, the trajectories of the following continuous-time strategy dynamics globally converge (cf. [60, 61, 89]):

$$\dot{x}(t) = f(x(t), p) - x(t), \tag{x-dynamics}$$

$$\dot{\tilde{x}}(t) = \tilde{f}(\tilde{x}(t), \tilde{p}) - \tilde{x}(t). \tag{$\tilde{x}$-dynamics}$$

In this work, we do not focus on analyzing the convergence of (x-dynamics)-($\tilde{x}$-dynamics). Instead, we assume any off-the-shelf convergent strategy update that satisfies the following assumption:

Assumption 12.2.2. *For any incentive mechanism p (resp. $\tilde{p}$), the Nash equilibrium $x^*(p)$ (resp. $\tilde{x}^*(\tilde{p})$) is the globally asymptotically stable fixed point of the continuous-time dynamical system* (x-dynamics) *(resp.* ($\tilde{x}$-dynamics)*).*

Assumption 12.2.2 is satisfied for a variety of strategy updates in various games. This includes the best-response and fictitious play strategy update in zero-sum and potential games [185, 43, 401, 239], and gradient-based strategy update in continuous games [290, 296].

Our goal here is to characterize conditions under which the coupled strategy and incentive updates (x-update)-(p-update) and ($\tilde{x}$-update)-($\tilde{p}$-update) converge. Before stating the convergence results, we define two notions of convergence.

Definition 12.2.2. *We say that the coupled strategy and incentive updates* (x-update)-(p-update)

(i) globally converges to the fixed point $(x^\dagger, p^\dagger)$ if, for any initial condition $p_0 \in \mathbb{R}^{|I|}$ and $x_0 \in X$, and any selection of step sizes that satisfy Assumption 12.2.1, the discrete-time updates (x-update)-(p-update) *asymptotically converge to $(x^\dagger, p^\dagger)$.*

(ii) locally converges to the fixed point $(x^\dagger, p^\dagger)$ if there exist positive scalars $\bar{r}, \bar{\alpha}, \bar{\beta}, \bar{\gamma}$ such that, when $p_0 \in \mathcal{B}_{\bar{r}}(p^\dagger)$ and $x_0 \in \mathcal{B}_{\bar{r}}(x^*(p_0))$, the step sizes satisfy Assumption 12.2.1 and the following condition:

$$\sup_{k\in\mathbb{N}} \frac{\beta_k}{\eta_k} \leqslant \bar{\alpha}, \quad \sup_{k\in\mathbb{N}} \beta_k \leqslant \bar{\beta}, \quad \sup_{k\in\mathbb{N}} \eta_k \leqslant \bar{\gamma}, \tag{12.18}$$

then the discrete-time updates (x-update)–(p-update) *asymptotically converge to $(x^\dagger, p^\dagger)$.*

Local and global convergence are analogously defined for the updates ($\tilde{x}$-update)-($\tilde{p}$-update) *in non-atomic games.*

Proposition 12.2.3. *Consider the atomic game G with discrete-time update* (x-update)-(p-update) *that satisfy Assumptions 12.2.1 and 12.2.2. The following requirements provide sufficient conditions for* (x-update)-(p-update) *to locally converge to the fixed point*[4] *$(x^\dagger, p^\dagger)$ in the sense of Definition 12.2.2:*

(R1) $p^\dagger$ is a locally asymptotically stable equilibrium of the following continuous-time dynamical system:

$$\dot{\boldsymbol{p}}(t) = e(x^*(\boldsymbol{p}(t))) - \boldsymbol{p}(t). \tag{12.19}$$

[4]This result holds even if the updates (x-update)-(p-update) and ($\tilde{x}$-update)-($\tilde{p}$-update) are perturbed with square-integrable martingale difference noise [61].

(R2) The trajectories of the discrete-time updates satisfy the boundedness condition:

$$\sup_{k\in\mathbb{N}} \big(\|x_k\| + \|p_k\|\big) < +\infty.$$

Furthermore, the sufficient conditions for (x-update)-($\tilde{p}$-update) *to globally converge to the fixed point* $(x^\dagger, p^\dagger)$ *in the sense of Definition 12.2.2 are (R1') and (R2), where*

(R1') $p^\dagger$ is a globally asymptotically stable equilibrium of the continuous-time dynamical system:

$$\dot{\boldsymbol{p}}(t) = e(x^*(\boldsymbol{p}(t))) - \boldsymbol{p}(t). \tag{12.20}$$

Analogous result holds for the non-atomic game $\tilde{G}$.

Proposition 12.2.3 states two sets of generic conditions that can be verified when studying convergence in any specific game. In particular, by leveraging results from nonlinear dynamical systems theory, (R1) (or (R1')) can be verified by showing the existence of a Lyapunov function [366] or by establishing that the dynamical system is cooperative [181]; see Lemma 1 in Chapter K.4. Additionally, (R2) holds in any game with a compact strategy set. In games with an unbounded strategy set, (R2) can be verified by analyzing the global convergence of continuous-time ('scaled') strategy and incentive dynamics [225, Theorem 10]. In Section 12.3, we verify that the conditions in Proposition 12.2.3 are satisfied in atomic aggregative games and non-atomic routing games. **Proof of Proposition 12.2.3.** Assumption 12.2.1-(ii), allow us to study the convergence of (x-update)-(p-update) in two stages [59, 61]. First, we study the convergence of fast strategy updates, for every fixed value of incentive. Second, we study the convergence of slow incentive updates, assuming that the fast strategy updates have converged to the equilibrium.

Formally, to study the convergence of fast strategy updates, we re-write (x-update)-(p-update) as follows

$$\begin{aligned} x_{k+1} &= x_k + \gamma_k \left(f(x_k, p_k) - x_k\right), \\ p_{k+1} &= p_k + \gamma_k \frac{\beta_k}{\gamma_k} \left(e(x_k) - p_k\right). \end{aligned} \tag{12.21}$$

Since $\sup_{k\in\mathbb{N}}(\|x_k\| + \|p_k\|) < +\infty$ (cf. requirement (R2)) and $\lim_{k\to\infty} \beta_k/\gamma_k = 0$ (cf. Assumption 12.2.1), the term $\frac{\beta_k}{\gamma_k}\left(e(x_k) - p_k\right)$ in (12.21) goes to zero as $k \to \infty$. Consequently, leveraging the standard approximation arguments [60, Lemma 1, Section 2.2], we conclude that the asymptotic behavior of the updates in (12.21) is same as that of the following dynamical system

$$\dot{\mathbf{x}}(t) = f(\mathbf{x}(t), \mathbf{p}(t)) - \mathbf{x}(t), \quad \dot{\mathbf{p}}(t) = 0.$$

Using Assumption 12.2.2, we conclude that

$$\lim_{k\to\infty} (x_k, p_k) \to \{(x^*(p), p) : p \in \mathbb{R}^{|I|}\}. \tag{12.22}$$

Next, to study the convergence of the slow incentive updates, we re-write (p-update) as follows

$$p_{k+1} = p_k + \beta_k \left(e(x^*(p_k)) - p_k\right) + \beta_k \left(e(x_k) - e(x^*(p_k))\right). \tag{12.23}$$

We will show that $(p_k)_{k\in\mathbb{N}}$ will asymptotically follow the trajectories of the following continuous-time dynamics:

$$\dot{\mathbf{p}}(t) = e(x^*(\mathbf{p}(t))) - \mathbf{p}(t). \tag{12.24}$$

Note that $p^\dagger$ is the fixed point of the trajectories of the dynamical system (12.19) (cf. Proposition 12.2.1). Requirement (R1) in Proposition 12.2.3 ensures convergence of (12.19).

Let $D^\dagger$ denote the domain of attraction of $p^\dagger$ for the dynamical system (12.19). From the converse Lyapunov theorem [370], we know that there exists a continuously differentiable function $\bar{V} : D^\dagger \to \mathbb{R}_+$ such that $\bar{V}(p^\dagger) = 0$, $\bar{V}(p) > 0$ for all $p \in D^\dagger \backslash \{p^\dagger\}$ and $\bar{V}(p) \to \infty$ as $p \to \mathsf{boundary}(D^\dagger)$. For any $r > 0$, define $\bar{V}_r = \{p \in \mathsf{dom}(\bar{V}) : \bar{V}(p) \leqslant r\}$ to be a sub-level set of $\bar{V}$. There exists $0 < \bar{r}' < \bar{r}$ such that $\bar{V}_{\bar{r}'} \subsetneq \mathcal{B}_{\bar{r}'}(p^\dagger) \subsetneq \mathcal{B}_{\bar{r}}(p^\dagger) \subsetneq \bar{V}_{\bar{r}}$. Additionally, define $t_0 = 0, t_k = \sum_{i=1}^k \beta_i$ and $L_k = t_{n(k)}$ where $n(0) = 0$, and

$$n(k) = \min \left\{ m \geqslant n(k-1) : \sum_{j=n(k-1)+1}^{m} \beta_j \geqslant T \right\} \ \forall k \in \mathbb{N}. \tag{12.25}$$

Here, T is a positive integer to be described shortly. Furthermore, define $\bar{\mathbf{p}}^{(k)} : \mathbb{R}_+ \to \mathbb{R}^{|I|}$ to be a solution of (12.19) on $[L_k, \infty)$ such that $\bar{\mathbf{p}}^{(k)}(L_k) = p_{L_k}$.

To ensure that $\bar{\mathbf{p}}^{(k)}(L_k) \in \mathsf{dom}(\bar{V})$ for $k > 0$, we show that for an appropriate choice of T in (12.25), $p_{L_k} \in \mathsf{int}(D^\dagger)$ for every $k \in \mathbb{N}$. From [61, Theorem IV.1], we know that there exists $K > 0$ such that for all $k \in \mathbb{N}$,

$$\begin{aligned}
&\|p_k - \bar{\mathbf{p}}^{(0)}(t_k)\| \\
&\leqslant K \left(\sup_k \beta_k + \sup_k \gamma_k + \sup_k \frac{\beta_k}{\gamma_k} + \sup_k \frac{\beta_k}{\gamma_k} \|x_0 - x^*(p_0)\| \right) \\
&= K \left(\bar{\alpha} + \bar{\beta} + \bar{\gamma} + \bar{\alpha}\bar{r} \right) =: \kappa.
\end{aligned}$$

Consequently, using the triangle inequality, it holds that

$$\|p_k - p^\dagger\| \leqslant \kappa + \|\bar{\mathbf{p}}^{(0)}(t_k) - p^\dagger\|. \tag{12.26}$$

Since $\bar{V}$ is a Lyapunov function of (12.19) and $\bar{\mathbf{p}}^{(0)}(0) = p_0 \in \mathcal{B}_{\bar{r}}(p^\dagger) \subsetneq \bar{V}_{\bar{r}}$, there exists $\bar{k} \in \mathbb{N}$ such that for all $k \geqslant \bar{k}$, $\bar{\mathbf{p}}^{(0)}(t_k) \in \bar{V}_{\bar{r}'} \subsetneq \mathcal{B}_{\bar{r}'}(p^\dagger)$. If we choose $\kappa < \bar{r} - \bar{r}'$ then, from (12.26), it holds that for all $k \geqslant \bar{k}$, $p_k \in \mathcal{B}_{\bar{r}}(p^\dagger)$. Therefore, if we choose $T \geqslant \bar{k}$ in (12.25), it holds that

$$p_{L_k} \in \mathsf{dom}(\bar{V}), \quad \forall\ k \in \mathbb{N}. \tag{12.27}$$

Define $\hat{p} : \mathbb{R}_+ \to \mathbb{R}$ such that, for every $k \in \mathbb{N}$, $\hat{p}(t_k) = p_k$ with linear interpolation on $[t_k, t_{k+1}]$. Using the standard approximation arguments from [60, Chapter 6], it holds that[5]

$$\begin{aligned} &\sup_{t\in[L_k,L_{k+1}]} \|\hat{p}(t) - \bar{\mathbf{p}}^{(k)}(t)\| \\ &\leqslant \mathcal{O}\left(\sum_{m\geqslant L_k} \beta_m^2 + \sup_{m\geqslant L_k} \|x_m - x^*(p_m)\|\right). \end{aligned} \tag{12.28}$$

Using (12.22) and Assumption 12.2.1, we conclude that RHS in the above equation goes to zero as $k \to \infty$. Finally, using (12.27), (12.28) and [59, Lemma 2.1], we conclude that $p_k \to p^\dagger$ as $k \to \infty$. □

12.3 Applications

In this section, we study the applicability of the general results from Section 8.4 to study convergence of our externality-based incentive updates in two practically relevant classes of games: atomic aggregative games, and non-atomic routing games.

Atomic Aggregative Games

Here, we study *quadratic* networked aggregative games [65, 64, 378, 1]. Consider a game G comprised of a finite set of players I. The strategy set of every player is the entire real line $\mathbb{R}$. Given the joint strategy profile $x = (x_i)_{i\in I}$, the cost of each player $i \in I$ is given by

$$\ell_i(x) = \frac{1}{2} q_i x_i^2 + \alpha x_i (Ax)_i, \tag{12.29}$$

where $A \in \mathbb{R}^{|I|\times|I|}$ is the *network matrix*, with A_{ij} representing the impact of player j's strategy on the cost of player i. The parameter $\alpha > 0$ characterizes the impact of the aggregate strategy on the individual cost of players. Moreover, $q_i > 0$ determines the influence of each player's own strategy on their cost function. Without loss of generality, we consider $A_{ii} = 0$ for all $i \in I$. For notational brevity, we define $Q = \mathsf{diag}((q_i)_{i\in I}) \in \mathbb{R}^{|I|\times|I|}$.

A system operator designs incentives through a payment $p_i x_i$ for player i when choosing strategy x_i. Thus, the total cost of player i is given by $c_i(x, p) = \ell_i(x) + p_i x_i$. The system operator's cost is

$$\Phi(x) = \sum_{i=1}^{n} \frac{1}{2}(x_i - \zeta_i)^2, \tag{12.30}$$

where $\zeta = (\zeta_i)_{i\in I} \in \mathbb{R}^{|I|}$ denotes the socially optimal strategy. Similar cost function has been considered for systemic risk analysis in financial networks [2]. In Chapter K.1, we generalize our results for a broader class of social cost functions.

[5]For any $T \geqslant \bar{k}$ and $\delta > 0$, there exists $k(\delta)$ such that $\hat{p}(t_{k(\delta)} + \cdot)$ form a "(T, δ)" perturbation (cf. [59]) of (12.19).

Proposition 12.3.1. *Suppose $M := Q + \alpha A$ is invertible. Then, the Nash equilibrium is given by $x^*(p) = -M^{-1}p$. Furthermore, the set $P^\dagger$ is a singleton set.*

The proof follows by noting that the game is strongly convex and equilibrium is computed by first order conditions. Proof of Proposition 12.3.1 is provided in Chapter K.2.

Next, we provide sufficient conditions to ensure global convergence of (x-update)-(p-update) to the fixed points.

Proposition 12.3.2. *Consider the updates* (x-update)-(p-update) *associated to the aggregative game G. Suppose that Assumptions 12.2.1 and 12.2.2 are satisfied. Additionally, if*

(i) $M := Q + \alpha A$ is symmetric positive definite, and

(ii) The function $f_c(x,p) := \frac{1}{c}(f(cx,cp) - cx)$, satisfy $f_c \to f_\infty$ as $c \to \infty$, uniformly on the compacts, and for every incentive vector $p \in \mathbb{R}^{|I|}$, $x^*(p)$ is the globally asymptotically stable fixed point of

$$\dot{x}(t) = f_\infty(x(t), p), \tag{12.31}$$

where, for any $x \in X$ and $p \in \mathbb{R}^{|I|}$, $f_\infty(x,p) = \lim_{c\to\infty} f_c(x,p)$.

Then, the discrete-time updates (x-update) *and* (p-update) *globally converges to the fixed point $(x^\dagger, p^\dagger)$ in the sense of Definition 12.2.2.*

We establish Proposition 12.3.2 by verifying requirements (R1') and (R2) of Proposition 12.2.3. To verify (R1'), we use Proposition 12.3.2-(i) to show that $V(p) = (p-p^\dagger)^\top M^{-\top}(p-p^\dagger)$ serves as a Lyapunov function candidate for the dynamical system (12.20), guaranteeing global convergence. Next, we leverage Proposition 12.3.2-(ii) along with [225, Theorem 10] to show that (R2) of Proposition 12.2.3 holds. Proof of Proposition 12.3.2 is in Chapter K.2.

Condition (ii) in Proposition 12.3.2 and Assumption 12.2.2 both impose global convergence of a suitably defined continuous-time strategy dynamics. In general, one need not imply the other. However, the two conditions become equivalent if the strategy update rule $f(x,p)$ (cf. (x-update)) is linear in both x and p, which is the case if the strategy updates are best-response-based (12.6) or gradient-based (12.7) in aggregative game.

Non-atomic Traffic Routing on General Networks

Consider a routing game $\tilde{G}$ that models the interactions of strategic travelers over a directed graph $G = (\tilde{\mathcal{E}}, \tilde{\mathcal{N}})$, where $\tilde{\mathcal{N}}$ is the set of nodes and $\tilde{\mathcal{E}}$ is the set of edges. Let $\tilde{I}$ be the set of origin-destination (o-d) pairs. Each o-d pair $i \in \tilde{I}$ is connected by a set of routes,[6] denoted by $\mathbf{R}_i$. Let $\mathbf{R} = \bigcup_{i\in\tilde{I}} \mathbf{R}_i$ represent the set of all routes in the network.

An infinitesimal traveler on the network is associated with an o-d pair and chooses a route to commute between the o-d pair. Let the total population of travelers associated with

[6] A route is a sequence of contiguous edges.

any o-d pair $i \in \tilde{I}$ be denote by $\tilde{M}_i$. Let $\tilde{x}_i^j$ be the amount of travelers taking route $j \in \mathbf{R}_i$ to commute between o-d pair $i \in \tilde{I}$ and $\tilde{x} = (\tilde{x}_i^j)_{j \in \mathbf{R}_i, i \in \tilde{I}}$ is a vector which contains, as its entries, the route flow of all population on different routes. Naturally, for every $i \in \tilde{I}$, it holds that $\sum_{j \in \mathbf{R}_i} \tilde{x}_i^j = \tilde{M}_i$. Any route flow $\tilde{x}$ induces a flow on the edges of the network, denoted by $\tilde{w}$, such that $\tilde{w}_a = \sum_{i \in \tilde{I}} \sum_{j \in \mathbf{R}_i} \tilde{x}_i^j \mathbb{1}(a \in j)$, for every $a \in \tilde{\mathcal{E}}$. We denote the set of feasible route flows by $\tilde{X}$ and the set of feasible edge flows by $\tilde{W} = \{(\tilde{w}_a)_{a \in \tilde{\mathcal{E}}} : \exists \tilde{x} \in \tilde{X}, \tilde{w}_a = \sum_{i \in \tilde{I}} \sum_{j \in \mathbf{R}_i} \tilde{x}_i^j\}$. For any o-d pair $i \in \tilde{I}$ and route flow $\tilde{x} \in \mathbb{R}^{|\mathbf{R}|}$, the cost experienced by travelers using route $j \in \mathbf{R}_i$ is $\tilde{\ell}_i^j(\tilde{x}) = \sum_{a \in \tilde{\mathcal{E}}} l_a(\tilde{w}_a) \mathbb{1}(a \in j)$, where $l_a(\cdot)$ is the *edge latency function* that depends on the edge flows. For every edge $a \in \tilde{\mathcal{E}}$, we assume that the edge latency function $l_a(\cdot)$ is convex and strictly increasing. This property of edge latency function captures the congestion effect on the transportation network [39, 354]. A system operator designs incentives by setting tolls on the edges of the network in the form of edge tolls[7], denoted by $\tilde{p} = (\tilde{p}_a)_{a \in \tilde{\mathcal{E}}}$. Every edge toll vector induces a unique route toll vector $\tilde{P}$. That is, for any o-d pair $i \in \tilde{I}$, the toll on route $j \in \mathbf{R}_i$ is

$$\tilde{P}_i^j = \sum_{a \in \tilde{\mathcal{E}} : a \in j} \tilde{p}_a. \tag{12.32}$$

Consequently, the total cost experienced by travelers on o-d pair $i \in \tilde{I}$ who choose route $j \in \mathbf{R}_i$ is $\tilde{c}_i^j(\tilde{x}, \tilde{P}) = \tilde{\ell}_i^j(\tilde{x}) + \tilde{P}_i^j$. Let $\tilde{x}^*(\tilde{P})$ denote a Nash equilibrium (also known as Wardrop equilibrium in non-atomic routing games literature) corresponding to route tolls $\tilde{P}$. Owing to (12.32), with slight abuse of notation, we shall frequently use $\tilde{x}^*(\tilde{P})$ and $\tilde{x}^*(\tilde{p})$ interchangeably. Typically, the equilibrium route flows can be non-unique but the corresponding edge flows $\tilde{w}^*(\tilde{p})$ are unique. Furthermore, the function $\tilde{p} \mapsto \tilde{w}^*(\tilde{p})$ is a continuous function [435].

The system operator's objective is to design tolls that ensure that the resulting equilibrium minimizes the overall travel time incurred by travelers on the network, characterized as the minimizer of

$$\tilde{\mathbf{\Phi}}(\tilde{x}) = \sum_{i \in \tilde{I}} \sum_{j \in \mathbf{R}_i} \tilde{x}_i^j \tilde{\ell}_i^j(\tilde{x}). \tag{12.33}$$

Note that the optimal route flow can be non-unique but the optimal edge flow, denoted by $w^\dagger$, is unique [435].

Using the description of travelers' costs, the externality caused by travelers from o-d pair $i \in \tilde{I}$ using route $j \in \mathbf{R}_i$, based on (12.8b), is given by

$$\tilde{e}_i^j(\tilde{x}) = \sum_{i' \in \tilde{I}} \sum_{j' \in \mathbf{R}_i} \tilde{x}_{i'}^{j'} \frac{\partial \tilde{\ell}_{i'}^{j'}(\tilde{x})}{\partial \tilde{x}_i^j} \overset{(a)}{=} \sum_{a \in \tilde{\mathcal{E}} : a \in j} \nabla l_a(\tilde{w}_a) \tilde{w}_a, \tag{12.34}$$

[7] If we directly use the setup of non-atomic games presented in Section 12.1, we would require the system operator to use *route-based* tolls rather than *edge-based* tolls. Our approach of using edge-based tolls is rooted in practical consideration with implementation of tolls.

where (a) is due to Lemma 2 in Chapter K.4.

From (12.34), we note that the externality on any route j is the sum externality on every edge on that route. Therefore, we study the following incentive update, which updates the edge-tolls as follows:

$$\tilde{p}_{a,k+1} = (1-\beta_k)\tilde{p}_{a,k} + \beta_k \tilde{e}_a(\tilde{x}_k), \quad \forall\ a \in \tilde{\mathcal{E}}, \tag{12.35}$$

where $\tilde{e}_a(\tilde{x}_k) = \nabla l_a(\tilde{w}_{a,k})\tilde{w}_{a,k}$, and $w_{a,k} = \sum_{i\in\tilde{I}}\sum_{j\in\tilde{R}_i}\tilde{x}^j_{i,k}$. Define

$$\tilde{\mathbf{P}}^\dagger = \{(\tilde{p}^\dagger_a)_{a\in\tilde{\mathcal{E}}} : \tilde{p}^\dagger_a = \tilde{w}^*_a(\tilde{p}^\dagger)\nabla l_a(\tilde{w}^*_a(\tilde{p}^\dagger)), \forall\ a \in \tilde{\mathcal{E}}\Big\},$$

to be the fixed point of the joint update ($\tilde{x}$-update)-(12.35).

Proposition 12.3.3. *The set $\tilde{\boldsymbol{P}}^\dagger$ is non-empty singleton set. The unique $p^\dagger \in \tilde{\boldsymbol{P}}^\dagger$ is socially optimal, i.e. $\tilde{w}(p^\dagger) = w^\dagger$.*

Proof of Proposition 12.3.3 follows in two steps. First, we show that any $p^\dagger \in \tilde{\mathbf{P}}^\dagger$ aligns the Nash equilibrium with social optimality, i.e. $\tilde{w}(p^\dagger) = w^\dagger$. Next, using contradiction argument similar to the proof of Proposition 12.2.1, we show that $\tilde{\mathbf{P}}^\dagger$ is singleton. Detailed proof of Proposition 12.3.3 is provided in Chapter K.3.

Next, we provide sufficient conditions for local convergence of the updates ($\tilde{x}$-update)-($\tilde{p}$-update).

Proposition 12.3.4. *Consider the updates* ($\tilde{x}$-update)-($\tilde{p}$-update) *associated with the routing game $\tilde{G}$. Suppose that Assumptions 12.2.1 and 12.2.2 are satisfied, and there exists an equilibrium route flow $\tilde{x}^*(\tilde{p}^\dagger)$ such that for every $i \in \tilde{I}, j, j' \in \tilde{R}_i$,*

$$\tilde{c}^j_i(\tilde{x}^*(\tilde{p}^\dagger)) \leqslant \tilde{c}^{j'}_i(\tilde{x}^*(\tilde{p}^\dagger)) \implies \tilde{x}^{j,*}_i(\tilde{p}^\dagger) > 0. \tag{12.36}$$

The discrete-time updates ($\tilde{x}$-update)-($\tilde{p}$-update) *locally converges to fixed point $(\tilde{x}^\dagger, \tilde{p}^\dagger)$ in the sense of Definition 12.2.2.*

Remark 12.3.1. *There is a subtle distinction between the definition of Nash equilibrium (cf.* (12.3)*) and* (12.36)*. The former states that at equilibrium, any route with a positive flow must have the minimum cost. In contrast,* (12.36) *further requires that all minimum-cost routes have strictly positive equilibrium flow. This regularity condition, commonly used in transportation literature ([435, Chapter 4]), ensures the differentiability of link flows $\tilde{w}^*(p)$ in the neighborhood of $\tilde{p}^\dagger$.*

We show Proposition 12.3.4 by verifying the requirements (R1)-(R2) in Proposition 12.2.3. (R2) holds due to the fact that $\tilde{X}$ is compact. Thus, it only remains to verify (R1). Towards this goal, we define $\Delta \in \mathbb{R}^{|\tilde{\mathcal{E}}|\times|\tilde{\mathcal{E}}|}$ to be a diagonal matrix such that, for every $a \in \tilde{\mathcal{E}}$,

$$\Delta_{a,a} = (\nabla l_a(\tilde{w}^*_a(\tilde{p}^\dagger)) + \tilde{w}^*_a(\tilde{p}^\dagger)\nabla^2 l_a(\tilde{w}^*_a(\tilde{p}^\dagger)))^{-1}. \tag{12.37}$$

Using condition (12.36), we show that $V(\tilde{p}) = (\tilde{p} - \tilde{p}^\dagger)^\top \Delta(\tilde{p} - \tilde{p}^\dagger)$, acts as a Lyapunov function candidate for the following dynamical system

$$\dot{\tilde{\mathbf{p}}}_a(t) = \tilde{w}_a^*(\tilde{\mathbf{p}}) \nabla l_a(\tilde{w}_a^*(\tilde{\mathbf{p}})) - \tilde{\mathbf{p}}_a, \quad \forall\ a \in \tilde{\mathcal{E}}. \tag{12.38}$$

Detailed proof is provided in Chapter K.3.

12.4 Concluding Remarks

We propose an adaptive incentive mechanism that updates based on agents' externalities, operates independently of their learning rules, and evolves on a slower timescale, forming a two-timescale coupled strategy and incentive dynamics. We show that its fixed point corresponds to an optimal incentive ensuring he Nash equilibrium of the corresponding game achieves social optimality. Additionally, we provide sufficient conditions for convergence of the coupled dynamics and validate our approach in atomic quadratic aggregative games and non-atomic routing games.

Part IV

Market Mechanisms for Emerging Advanced Air Mobility: A Case Study

Chapter 13

Incentive-Compatible Market Mechanisms for Advanced Air Mobility

Advanced air mobility (AAM) encompasses the utilization of unmanned aerial vehicles (UAVs), air taxis, and various cargo and passenger transport solutions. This innovative approach taps into previously unexplored airspace, poised to revolutionize urban airspace. A recent report forecasts the air mobility market alone to exceed US$50 billion by 2035, underlining this area's immense growth potential [101].

Despite the widespread optimism surrounding AAM, the design of regulatory policies remains an open problem. While ideas from conventional air traffic management (e.g. [49, 50, 48, 355, 323]) could be leveraged, they often fall short in accommodating the dynamic and adaptable nature of AAM operations [53], resulting from on-demand requests from operators with heterogeneous private valuations [385, 375]. Indeed, the administrative management methods prevalent in traditional air traffic management, such as grand-fathering rights, flow management, and first-come-first-serve, prove ineffective for AAM operations [166, 132] as these approaches fail to elicit the heterogeneous private valuations (arising from different aircraft specifications, demand realization, etc.) different operators have on using AAM resources. Furthermore, they risk fostering inefficient and anti-competitive outcomes, as evidenced in traditional airspace operations [122]. Recognizing the need for tailored regulation, the Federal Aviation Administration (FAA) is actively developing a *clean-slate* congestion management framework for AAM operations to ensure efficiency, fairness, and safety [5].

Market-based congestion management mechanisms have been proposed as potential solutions for AAM operations [96, 346, 415, 132, 385, 375]. Even in conventional airspace management, market-based mechanisms are extensively studied such as [29, 34, 82, 294], where both theoretical and empirical evidence show their precedence over administrative approaches [122]. However, the design of market-based mechanisms that guarantee safety, efficiency, and fairness under the heterogeneous and on-demand nature of AAM operations has remained elusive as the existing approaches concentrate heavily on tactical deconflic-

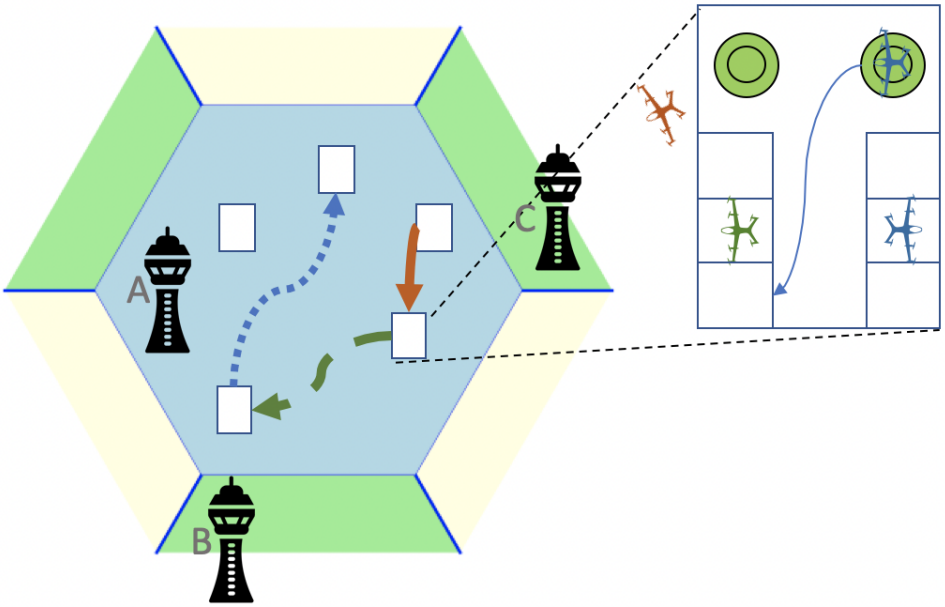


Figure 13.1: Schematic representation of the air traffic network with a service provider tasked with coordinating the movement of aircraft of various fleet operators between vertiports in its domain. Each vertiport has a constraint on the number of arriving aircraft, departing aircraft, and parked aircraft.

tion [211, 45], while not accounting for efficiency, fairness and the economic incentives of operators [95, 399, 94, 96, 415, 346, 166, 132].

In this chapter, we introduce an auction-based mechanism for a prominent AAM scenario of vertiport reservation, where electric vertical take-off and landing (eVTOL) operators with heterogeneous private valuations need to be coordinated to use vertiports based on their realized demands. This problem is challenging for three main reasons. First, the resulting reservation must ensure efficient, fair, and safe allocation of resources. Second, the operators may misreport their private valuations and demands to gain access to more valuable airspace resources (i.e. ensuring *incentive compatibility*). Third, the computation of these auction mechanisms is combinatorial, as evidenced by existing air traffic flow management frameworks [49, 50, 48] (i.e. ensuring fast computability). Thus, the main question we set out for this work is:

> *How to design an efficient, fair, and safe vertiport reservation mechanism for heterogeneous and on-demand nature of eVTOL operators, while ensuring incentive compatibility and faster computation?*

We consider an air transportation network (ATN) managed by a service provider (SP). The SP is responsible for ensuring the efficient, safe, and fair movement of aircraft operated

by various fleet operators (FOs) between vertiports (as depicted in Figure 13.1). The goal of the SP is to maximize a metric of *social welfare* that is comprised of two objectives: (i) maximize the overall (weighted) valuations[1] of all FOs, and (ii) minimize excessive congestion at vertiports[2]. Additionally, the SP must (iii) enforce arrival, departure, and parking capacity constraints at vertiports, and (iv) elicit truthful valuations from heterogeneous FOs in the form of bids.

We propose an auction mechanism, to be used by the SP, that satisfies $(i) - (iv)$. In this mechanism, using the bids submitted by FOs, the SP allocates the resources by maximizing social welfare, subject to capacity constraints. Next, the SP charges each FO a payment based on the *externality* imposed by them, which is assessed by the difference in the optimal social welfare of remaining FOs when this FO is included versus when it is excluded from the auction environment. Note that this payment mechanism is inspired by the generalized Vickrey–Clarke–Groves (VCG) mechanism [318]. We theoretically study the properties of the proposed mechanism in terms of incentive compatibility, individual rationality, and social welfare maximization (cf. Theorem 13.2.1).

There are two computational challenges associated with designing this mechanism. First, naively optimizing social welfare over the set of feasible allocations can be computationally challenging. To address this, we frame the problem as a mixed binary linear program by constructing a network-flow graph, which reduces the number of binary variables. Second, computing the externality in the payment mechanism— which requires maximizing social welfare over the set of feasible allocations— entails characterizing the set of feasible allocations when an FO is excluded from the auction environment. This is non-trivial, as the underlying resource allocation problem is an exchange problem. To overcome this, we introduce the idea of *pseudo-bids*, where we simply set a bid of 0 for an FO while computing the optimal allocation when this FO is excluded from the auction environment.

We note two important features of the problem we study in this chapter. First, we focus only on strategic deconfliction where the safety is encoded in the form of minimizing congestion and ensuring capacity constraints, and not on tactical deconfliction. However, our approach can be integrated into the airborne automation workflow proposed in [426] to also account for tactical deconfliction. Second, this problem is an "exchange problem", where some of the resources desired by any FO could be occupied by aircraft of other FOs, and a feasible allocation in this setting needs to exchange the resources between FOs while respecting capacity constraints. In contrast, the standard slot allocation problems studied in conventional air traffic literature (cf. [122, 294, 29, 347, 340, 30, 53]) are "assignment problems" where the slots need to be assigned to airlines and not exchanged between airlines.

Notation: We denote the set of real numbers by $\mathbb{R}$, non-negative real numbers by $\mathbb{R}_+$, integers by $\mathbb{Z}$, non-negative integers by $\mathbb{Z}_+$, and natural numbers by $\mathbb{N}$. For $N \in \mathbb{N}$, we define $[N] := \{1, 2, ..., N\}$. The indicator function is denoted as $\mathbf{1}(\cdot)$, which is 1 when $(\cdot)$

[1] We allow the SP to weigh FOs differently in order to encourage new-comers in this emerging market.

[2] Note that we *only* consider congestion at the vertiports in this work. An extension to airborne congestion is discussed in Subsection 13.4.

is true and 0 otherwise. When indexing a set $b = \{b_1, b_2, ..., b_N\}$, we follow the standard game-theoretic notation: $b_{-i} := \{b_1, ..., b_{i-1}, b_{i+1}, ..., b_N\}$.

13.1 Problem Setup

System Model

We consider an air transportation network (ATN), comprised of multiple vertiports, which are used by electric vertical take-off and landing (eVTOL) aircraft. We focus on a strategic deconfliction mechanism that complements the tactical deconfliction algorithms proposed in [211, 432, 45, 381]. The scheduling mechanism proceeds over non-overlapping time slots with a receding time horizon. At the beginning of each time slot, all fleet operators (FOs) submit a *menu* of desired origin-destination pairs and the corresponding bids specifying how much they are willing to pay for getting scheduled. Then, the service provider (SP) will compute a feasible allocation and payment and execute them in the next time slot. The granted aircraft can now go to their desired locations. In most congested vertiports, when the parking capacity is fully utilized, any additional arrival would necessitate a simultaneous departure of an aicraft from that vertiport. Thus, this is an "exchange problem" as opposed to the "assignment problem" studied in other air traffic allocation problems [122, 294, 29, 347, 340, 30, 53].

We denote the set of vertiports by R, the set of FOs by F, and the set of eVTOL aircraft by A. We consider the problem for H time slots.

Vertiports At any time $t \in [T]$, each vertiport $r \in R$ has three kinds of capacity constraints [3]: *(i) arrival capacity constraints*, denoted by $\mathsf{arr}(r,t) \in \mathbb{Z}_+$, that restrict the number of eVTOLs that can land at vertiport r at time t; *(ii) departure capacity constraints*, denoted by $\mathsf{dep}(r,t) \in \mathbb{Z}_+$, that restrict the number of eVTOLs that can depart from vertiport r at time t; *(iii) parking capacity constraints*, denoted by $\mathsf{park}(r,t) \in \mathbb{Z}_+$, that restrict the number of eVTOLs that can park at vertiport r at time t.

Fleet Operators Let A_i be the fleet of aircraft operated by FO $i \in F$, and $A := \{a_{i,j} | i \in F, j \in A_i\}$ be the set of all aircraft using the ATN. Each aircraft $a_{i,j}$ is identified by a tuple $\left(r_{i,j}^{\mathsf{orig}}, m_{i,j}, \{t_{i,j,k}^{\mathsf{dep}}, t_{i,j,k}^{\mathsf{arr}}, v_{i,j,k}, b_{i,j,k}, r_{i,j,k}^{\mathsf{dest}}\}_{k \in m_{i,j}}\right)$ where (i) $r_{i,j}^{\mathsf{orig}} \in R$ is the origin vertiport of aircraft $a_{i,j}$, (ii) $m_{i,j}$ is the menu of available routes to aircraft $a_{i,j}$; (iii) any route $k \in m_{i,j}$ implies that aircraft $a_{i,j}$ departs from $r_{i,j}^{\mathsf{orig}} \in R$ at time $t_{i,j,k}^{\mathsf{dep}}$ to arrive at $r_{i,j,k}^{\mathsf{dest}} \in R$ at time $t_{i,j,k}^{\mathsf{arr}}$; (iv) $v_{i,j,k}$ denotes the private valuation of aircraft $a_{i,j}$ to choose the route $k \in m_{i,j}$; and (v) $b_{i,j,k} \in \mathbb{R}_+$ is the bid submitted by FO i to schedule aircraft $a_{i,j}$ on route $k \in m_{i,j}$.

[3] The arrival, departure and parking capacity constraints in our model are exogeneously determined at every time step and are un-correlated between two consecutive time steps. Extending our model to account for correlations is an interesting direction of future research.

Note that we include the option to stay parked at the same vertiport in $m_{i,j}$, denoted by $\varnothing$, and set its departure time to 0.

Additionally, we denote the joint bid profile of all aircraft operated by FO $i \in F$ by $B_i := (b_{i,j,k})_{j\in A_i, k\in m_{i,j}}$ and joint valuation profile of its fleet by $V_i := (v_{i,j,k})_{j\in A_i,k\in m_{i,j}}$. For succinct notation, we denote the joint bid and valuation profile of all FOs as $B := (B_i)_{i\in F}$ and $V := (V_i)_{i\in F}$, respectively.

Problem Formulation

We consider an SP tasked with coordinating[4] the movement of aircraft by allocating them to their desired vertiports while ensuring that the capacity constraints are met. Formally, the SP needs to decide on a feasible allocation $x = (x_{i,j,k} \in \{0,1\} | i \in F, j \in A_i, k \in m_{i,j})$, where

$$x_{i,j,k} = \begin{cases} 1, & \text{if aircraft } a_{i,j} \text{ is allocated route } k \in m_{i,j}, \\ 0, & \text{otherwise.} \end{cases}$$

Given an allocation x, let $S(r,t,x) \in \mathbb{Z}_+$ denote the number of aircraft occupying the parking spots at vertiport $r \in R$ at time $t \in [T]$. For every $r \in R$, the initial occupation $S(r,1,x)$ is

$$S(r,1,x) = \sum_{i\in F} \sum_{j\in A_i} \mathbf{1}(r^{\mathsf{orig}}_{i,j} = r).$$

For concise notation, we shall denote $S(r,1,x)$ by $\bar{S}(r)$ for every $r \in R$ since it does not depend on x. Naturally, it must hold that, for every $r \in R, t \in \{2,\dots,T\}$,

$$\begin{aligned} S(r,t,x) = S(r,t-1,x) &+ \sum_{i\in F} \sum_{j\in A_i} \sum_{k\in m_{i,j}} x_{i,j,k} \mathbf{1}(r^{\mathsf{dest}}_{i,j,k} = r, t^{\mathsf{arr}}_{i,j,k} = t) \\ &- \sum_{i\in F} \sum_{j\in A_i} \sum_{k\in m_{i,j}} x_{i,j,k} \mathbf{1}(r^{\mathsf{orig}}_{i,j} = r, t^{\mathsf{dep}}_{i,j,k} = t), \end{aligned} \tag{13.1}$$

where the second (resp. third) term on the RHS in the above equation denotes the set of incoming (resp. departing) aircraft in vertiport r at time t. The *residual capacity* at vertiport $r \in R$ at time $t \in [H]$ is $Z(r,t,x) := \mathsf{park}(r,t) - S(r,t,x)$. To ensure the existence of a feasible allocation as defined later in (13.2), we assume that $\mathsf{park}(r,t) - \bar{S}(r) \geqslant 0, \forall r \in R, t \in [H]$.

An allocation x is called feasible if it satisfies the following constraints:

(C1) Each aircraft is allocated at most one route. That is, for every $i \in F, j \in A_i$, $\sum_{k\in m_{i,j}} x_{i,j,k} \leqslant 1$.

[4]We do not impose the information sharing constraints in [346, 94], where different sectors have different operators, and an SP only provides the identities, but not the positions, of aircraft to neighboring sectors. We follow the architecture in the current ATFM framework [323, 49, 50, 48, 355], where a central SP can aggregate information from all the sectors and make decisions.

(C2) Arrival and departure capacity constraints must be satisfied at every vertiport r at all times. That is, for every $r \in R, t \in [T]$,

$$\sum_{i\in F}\sum_{j\in A_i}\sum_{k\in m_{i,j}} x_{i,j,k}\mathbf{1}(r^{\mathsf{dest}}_{i,j,k} = r, t^{\mathsf{arr}}_{i,j,k} = t) \leqslant \mathsf{arr}(r,t),$$
$$\sum_{i\in F}\sum_{j\in A_i}\sum_{k\in m_{i,j}} x_{i,j,k}\mathbf{1}(r^{\mathsf{orig}}_{i,j} = r, t^{\mathsf{dep}}_{i,j,k} = t) \leqslant \mathsf{dep}(r,t).$$

(C3) Parking capacity constraints must be satisfied. That is, for every vertiport $r \in R$ at any time $t \in [T]$, $Z(r,t,x) \geqslant 0$.

Consequently, we define

$$X := \left\{ x \in \{0,1\}^{\sum_{i\in F}\sum_{j\in A_i}|m_{i,j}|} \,\middle|\, x \text{ satisfies (C1)-(C3)} \right\} \tag{13.2}$$

to be the set of feasible allocations.

Definition 13.1.1 (Social Welfare)**.** *Given $x \in X$,* social welfare *is defined as follows.*

$$\mathsf{SW}(x;V) := \sum_{i\in F}\rho_i \sum_{j\in A_i}\sum_{k\in m_{i,j}} v_{i,j,k}\cdot x_{i,j,k} - \lambda \sum_{r\in R}\sum_{t\in[T]} C_{r,t}(S(r,t,x)), \tag{13.3}$$

where (i) $\rho_i \in \mathbb{R}_+$ is the weight factor specifying the relative importance of different FOs[5]*, (ii) $C_{r,t} : \mathbb{Z}_+ \to \mathbb{R}_+$ with $C_{r,t}(0) = 0$ is discrete convex*[6] *to capture increasing marginal cost of congestion*[7]*, and (iii) $\lambda \in \mathbb{R}_+$ is the ratio between the congestion cost and the cumulative weighted valuations of FOs. Furthermore, we define an optimal allocation as*

$$x^*(V) \in \arg\max_{x\in X} \mathsf{SW}(x;V), \tag{13.4}$$

where ties are resolved arbitrarily.

Remark 13.1.1. *The social welfare objective* (13.3) *captures three main desiderata: efficiency, fairness, and safety. The objective* (13.3) *incorporates efficiency through additive valuations of FOs. Additionally, it incorporates the* proportional fairness *criterion*[8] *by assigning different weights to the valuations of different FOs, denoted by $(\rho_i)_{i\in F}$. Well-constructed weights can prevent larger FOs from monopolizing the resources; for example, using the logarithm of the number of aircraft as an FO's weight. Finally, it encompasses safety considerations in two ways: first, through capacity constraints; and second, by introducing a congestion-dependent term in* (13.3) *that penalizes vertiports when the number of aircraft increases. With these three considerations, the definition of social welfare aligns closely with that presented in [122].*

[5]Similar weight factors, termed as *remote city opportunity factor*, are used in [122].

[6]Based on [308], a function $f : \mathbb{Z} \to \mathbb{R}$ is discrete convex if $f(x+1) - f(x) \geqslant f(x) - f(x-1)$, $\forall x \in \mathbb{Z}$.

[7]While we only consider the congestion resulting from parked aircraft, it is straightforward to extend our formulation to arriving and departing aircraft; see Subsection 13.4.

[8]We emphasize that the fairness is at the FO-level.

We assume the SP does not have access to the true valuations V, as it is private information. Instead, the SP must use bids B reported by the FOs to allocate the aircraft to vertiports through an auction mechanism. More formally, given a bid profile B, the SP uses a mechanism $\bar{M} = (\bar{x}, (\bar{p}_i)_{i \in F})$, where for a given bid profile B, (i) $\bar{x}(B) \in X$ is the allocation proposed by the mechanism; and (ii) $\bar{p}_i(B) \in \mathbb{R}$ denotes the payment charged to FO $i \in F$. Under the mechanism $\bar{M}$, the utility derived by any FO $i \in F$ is

$$U_i(B; \bar{M}) = \sum_{j \in A_i} \sum_{k \in m_{i,j}} v_{i,j,k} \mathbf{1}(\bar{x}_{i,j,k}(B)) - \bar{p}_i(B). \tag{13.5}$$

Given any arbitrary valuation profile V, the goal is to design a vertiport reservation mechanism $\bar{M} = (\bar{x}, \bar{p})$ with the following desiderata.

(D1) **Incentive Compatibility (IC):** Bidding truthfully is each FO's (weakly) dominant strategy, i.e., for every $i \in F$, $B_{-i} \in \mathbb{R}_+^{\sum_{\ell \in F \setminus \{i\}} \sum_{j \in A_\ell} |m_{\ell,j}|}$,

$$V_i \in \underset{B_i \in \mathbb{R}_+^{\sum_{j \in A_i} |m_{i,j}|}}{\arg\max} U_i(B_i, B_{-i}; \bar{M}).$$

(D2) **Individual Rationality (IR):** Bidding truthfully results in non-negative utility, i.e., for every $i \in F$,

$$U_i(V_i, B_{-i}; \bar{M}) \geqslant 0, \quad \forall\ B_{-i} \in \mathbb{R}_+^{\sum_{\ell \in F \setminus \{i\}} \sum_{j \in A_\ell} |m_{\ell,j}|}.$$

(D3) **Social Welfare Maximization (SWM):** The resulting allocation maximizes social welfare, i.e.,

$$\bar{x}(B) \in \underset{x \in X}{\arg\max}\ \mathsf{SW}(x; V).$$

13.2 Mechanism Design

In this section, we present an auction mechanism that satisfies (D1)-(D3) in Subsection 13.2 and prove its theoretical properties in Subsection 13.2. We defer the optimization algorithm to Section 13.3.

Mechanism

Inspired by Myerson's lemma [310], our approach is to separate the allocation and payment functions so that the latter can ensure IC and IR as long as the former ensures maximization of total welfare in terms of bids submitted.

Allocation Function: Given a bid profile $B \in \mathbb{R}_+^{\sum_{i\in F}\sum_{j\in A_i}|m_{i,j}|}$, the allocation is obtained by

$$\bar{x}(B) \in \arg\max_{x\in X} \mathsf{SW}(x;B). \tag{13.6}$$

Payment Function: We first define a function

$$\theta : F \times \mathbb{R}_+^{\sum_{i\in F}\sum_{j\in A_i}|m_{i,j}|} \to \mathbb{R}_+^{\sum_{i\in F}\sum_{j\in A_i}|m_{i,j}|}$$

such that for any $\ell \in F$ and bid $B \in \mathbb{R}_+^{\sum_{i\in F}\sum_{j\in A_i}|m_{i,j}|}$,

$$\theta_{i,j,k}(\ell,B) = \begin{cases} b_{i,j,k}, & \text{if } i \neq \ell, \\ 0, & \text{if } i = \ell, \end{cases} \quad \forall\, i \in F, j \in A_i, k \in m_{i,j}. \tag{13.7}$$

The payment function, given a bid profile B, is

$$\bar{p}_i(B) = \frac{1}{\rho_i}\left(\max_{x'\in X} \mathsf{SW}_{-i}(x';\theta(i,B)) - \mathsf{SW}_{-i}(\bar{x};B)\right), \tag{13.8}$$

where for every $i \in F$, $x \in X$, and $B \in \mathbb{R}_+^{\sum_{i\in F}\sum_{j\in A_i}|m_{i,j}|}$,

$$\mathsf{SW}_{-i}(x;B) := \sum_{\ell\in F_{-i}} \rho_\ell \sum_{j\in A_\ell} \sum_{k\in m_{\ell,j}} b_{\ell,j,k}\cdot x_{\ell,j,k} - \lambda \sum_{r\in R}\sum_{t\in[T]} C_{r,t}(S(r,t,x)). \tag{13.9}$$

Remark 13.2.1. *The payment rule is inspired by the VCG mechanism, where each FO is charged a payment based on the* externality *created by them. Particularly, the typical VCG payment for any player is determined by assessing the difference in the optimal social welfare of players when they are present, versus when they are excluded from the auction environment.*

Remark 13.2.2. *There are some notable differences between the VCG payment and* (13.8)*. First, since our problem is an "exchange problem" and not the typical "assignment problem", we need to be cognizant of the physical resources occupied by the aircraft of that operator. However, this would require us to enumerate all the feasible combinations if we were to directly implement VCG mechanisms. To overcome the problem of enumerating all feasible solutions while computing payments, we adopt a novel approach of "pseudo-bids", where while computing the payments, each non-participating aircraft is considered to be using a bid of* 0*, as formally described in* (13.7)*.*

Second, since the objective function (13.3) *is not the summation of the participants' valuations, the typical VCG auction is not directly applicable. Instead, we follow [122, 318] to devise the payment rule for any $i \in F$ and $b \in \mathbb{R}_+^{\sum_{j\in A_i}|m_{i,j}|}$.*

Theoretical Analysis

Theorem 13.2.1. *The proposed mechanism $\bar{M} := (\bar{x}, \bar{p})$, defined by (13.6) and (13.8) is IC, IR, and SWM.*

Proof. Observe from (13.3) that $\mathsf{SW}(x; V)$ is a weighted summation of FOs' valuations and the congestion cost. Since the congestion cost is independent of valuations,

$$\bar{x}(V) \in \arg\max_{x \in X} \mathsf{SW}(x; V)$$

is an affine maximizer with respect to FOs' valuations, as defined in [318, Definition 9.30]. Thus, the allocation function (13.6) and the payment function (13.8) form a generalized VCG mechanism, and IC directly follows from [318, Proposition 9.31]. Finally, IR follows from [318, Lemma 9.20] since the bids are non-negative, and the allocation is an affine maximizer, as formally proved below.

For any $B_{-i} \in \mathbb{R}_+^{\sum_{\ell \in F \setminus \{i\}, j \in A_\ell} |m_{\ell,j}|}$,

$$\begin{aligned} U_i(V_i, B_{-i}; \bar{M}) &= \sum_{j \in [A_i]} \sum_{k \in [m_{i,j}]} v_{i,j,k} \mathbf{1}(\bar{x}_{i,j,k}(V_i, B_{-i})) - \bar{p}_i(V_i, B_{-i}) \\ &= \frac{1}{\rho_i} \Bigg(\rho_i \sum_{j \in [A_i]} \sum_{k \in [m_{i,j}]} v_{i,j,k} \mathbf{1}(\bar{x}_{i,j,k}(V_i, B_{-i})) + \mathsf{SW}_{-i}(\bar{x}; V_i, B_{-i}) \\ &\qquad - \max_{x' \in X} \mathsf{SW}_{-i}(x'; \theta(i, V_i, B_{-i})) \Bigg) \\ &= \frac{1}{\rho_i} \Bigg(\mathsf{SW}(\bar{x}; V_i, B_{-i}) - \max_{x' \in X} \mathsf{SW}_{-i}(x'; \theta(i, V_i, B_{-i})) \Bigg). \end{aligned} \tag{13.10}$$

Since $\bar{x} \in \arg\max_{x \in X} \mathsf{SW}(\bar{x}; V_i, B_{-i})$, it holds that $\mathsf{SW}(\bar{x}; V_i, B_{-i}) \geqslant \mathsf{SW}(x^\dagger; V_i, B_{-i})$, where $x^\dagger \in \arg\max_{x' \in X} \mathsf{SW}_{-i}(x'; \theta(i, V_i, B_{-i}))$. Thus, we obtain

$$\begin{aligned} U_i(V_i, B_{-i}; \bar{M}) &\geqslant \frac{1}{\rho_i} \Big(\mathsf{SW}(x^\dagger; V_i, B_{-i}) - \mathsf{SW}_{-i}(x^\dagger; \theta(i, V_i, B_{-i})) \Big) \\ &= \frac{1}{\rho_i} \sum_{j \in [A_i]} \sum_{k \in [m_{i,j}]} v_{i,j,k} \mathbf{1}(x^\dagger_{i,j,k}(V_i, B_{-i})) \geqslant 0. \end{aligned}$$

□

13.3 Optimization Algorithm

In this section, we formulate (13.6) as a mixed binary linear program (MBLP), as shown in (13.16). We derive this in three steps. First, in Subsection 13.3, we construct a time-extended flow network, where vertices are vertiport-time and aircraft-time pairs with edges

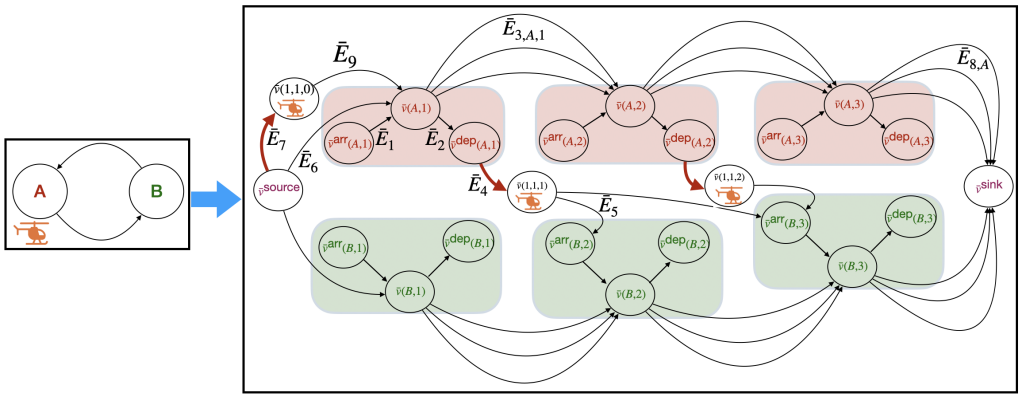


Figure 13.2: Auxiliary graph $\bar{G}$ constructed from an ATN with two vertiports and one aircraft over three time slots.

capturing capacity constraints and route allocation. Then, using binary variables ($\delta_{i,j,\tau}$ as formally defined later in (13.12d) and (13.12e)) to ensure that each aircraft is allocated one route, we formulate a mixed integer linear program (MILP (13.12)) in Subsection 13.3. This MILP has fewer binary variables than (13.6) when the number of unique departure times for any aircraft is less than the size of its menu. Finally, in Subsection 13.3, we show that the total unimodularity of the constraint matrix ($\bar{I}_\star$ in (13.12b)) guarantees that all flows are integral for each binary variable assignment, so we can drop the integrality constraint (13.12f) and get the final MBLP formulation (13.16).

Auxiliary Graph

We construct an auxiliary graph $\bar{G} = (\bar{V}, \bar{E})$ as detailed below. Figure 13.2 shows a pictorial depiction.

(i) *Set of vertices* $\bar{V} = \cup_{\ell=1}^{3} \bar{V}_\ell$. We define these sets below:

- $\bar{V}_1 := \{(\bar{\nu}(r,t), \bar{\nu}^{\mathsf{arr}}(r,t), \bar{\nu}^{\mathsf{dep}}(r,t)) | r \in R, t \in [T]\}$:
  We consider three replica for each vertiport $r \in R$ at time $t \in [T]$, denoted as $\bar{\nu}(r,t), \bar{\nu}^{\mathsf{arr}}(r,t)$, and $\bar{\nu}^{\mathsf{dep}}(r,t)$. These vertices, along with $\bar{E}_1$, $\bar{E}_2$, $\bar{E}_3$, and $\bar{E}_8$ defined later, embed capacity constraints and congestion costs into the graph structure.
- $\bar{V}_2 := \{\bar{\nu}(i,j,\tau) | i \in F, j \in A_i, \tau \in T^{\mathsf{dep}}_{i,j}\}$:
  For each $i \in F, j \in A_i$, we consider one vertex corresponding to all routes that have the same departure time. More formally, for every $i \in F, j \in A_i$, define $T^{\mathsf{dep}}_{i,j} := \cup_{k \in m_{i,j}} \left\{t^{\mathsf{dep}}_{i,j,k}\right\}$, to be the set of unique departure times amongst all routes. We consider one vertex corresponding to each $i \in F, j \in A_i$, and $\tau \in T^{\mathsf{dep}}_{i,j}$,

denoted as $\bar{\nu}(i,j,\tau)$, which, along with $\bar{E}_4$, $\bar{E}_5$, $\bar{E}_7$, and $\bar{E}_9$ defined later, embeds the route choice of the aircraft.

- $\bar{V}_3 := \{\bar{\nu}^{\mathsf{source}}, \bar{\nu}^{\mathsf{sink}}\}$:
  $\bar{\nu}^{\mathsf{source}}$ and $\bar{\nu}^{\mathsf{sink}}$ denote the source and sink in the flow network (to be described shortly). These vertices, along with $\bar{E}_6$, $\bar{E}_7$, and $\bar{E}_8$, ensure flow conservation of the parking aircraft.

(ii) *Set of edges* $\bar{E} = \cup_{\ell=1}^{9} \bar{E}_\ell \subseteq \bar{V} \times \bar{V} \times \mathbb{Z}_+ \times \mathbb{Z}_+ \times \mathbb{R}$, where each edge is identified with a tuple $(r, r', \overline{\mathbf{c}}, \underline{\mathbf{c}}, \bar{\mathbf{w}})$ such that (i) $r, r' \in R$ are the upstream and downstream vertiport on an edge, respectively, (ii) $\overline{\mathbf{c}}, \underline{\mathbf{c}} \in \mathbb{Z}_+$ are the upper and lower bound on the capacity of the edge, respectively, and (iii) $\bar{\mathbf{w}} \in \mathbb{R}$ is the edge weight.

- $\bar{E}_1 := \{(\bar{\nu}^{\mathsf{arr}}(r,t), \bar{\nu}(r,t), \overline{\mathbf{c}} = \mathsf{arr}(r,t), \underline{\mathbf{c}} = 0, \bar{\mathbf{w}} = 0) | r \in R, t \in [T]\}$.
- $\bar{E}_2 := \{(\bar{\nu}(r,t), \bar{\nu}^{\mathsf{dep}}(r,t), \overline{\mathbf{c}} = \mathsf{dep}(r,t), \underline{\mathbf{c}} = 0, \bar{\mathbf{w}} = 0) | r \in R, t \in [T]\}$.
- $\bar{E}_3 := \cup_{r \in R, t \in [T-1]} \bar{E}_{3,r,t}$:
  For every $r \in R, t \in [H-1]$, we consider $\mathsf{park}(r,t)$ edges connecting $\bar{\nu}(r,t)$ and $\bar{\nu}(r,t+1)$. We denote this set by $\bar{E}_{3,r,t}$. For any $q \in [\mathsf{park}(r,t)]$, we denote the weight of the $q-$th edge in $\bar{E}_{3,r,t}$ by $\bar{\mathbf{w}}_{q,r,t}$, and upper and lower capacity by $\overline{\mathbf{c}}_{q,r,t}$ and $\underline{\mathbf{c}}_{q,r,t}$, respectively. For any $r \in R, q \in [\mathsf{park}(r,t)]$, $\bar{\mathbf{w}}_{q,r,t} = -\lambda(C_{r,t}(q) - C_{r,t}(q-1))$, $\overline{\mathbf{c}}_{q,r,t} = 1$, and $\underline{\mathbf{c}}_{q,r,t} = 0$.
- $\bar{E}_4 := \{(\bar{\nu}^{\mathsf{dep}}(r^{\mathsf{orig}}_{i,j}, \tau), \bar{\nu}(i,j,\tau), \overline{\mathbf{c}} = \underline{\mathbf{c}} = \delta_{i,j,\tau}, \bar{\mathbf{w}} = 0) | i \in F, j \in A_i, \tau \in T^{\mathsf{dep}}_{i,j} \backslash \{0\}\}$:
  $\delta_{i,j,\tau} \in \{0,1\}$ is a variable defined later.
- $\bar{E}_5 := \{(\bar{\nu}(i,j,t^{\mathsf{dep}}_{i,j,k}), \bar{\nu}^{\mathsf{arr}}(r^{\mathsf{dest}}_{i,j,k}, t^{\mathsf{arr}}_{i,j,k}), \overline{\mathbf{c}} = 1, \underline{\mathbf{c}} = 0, \bar{\mathbf{w}} = \rho_i b_{i,j,k}) | i \in F, j \in A_i, k \in m_{i,j} \backslash \{\varnothing\}\}$.
- $\bar{E}_6 := \{(\bar{\nu}^{\mathsf{source}}, \bar{\nu}(r,1), \overline{\mathbf{c}} = \underline{\mathbf{c}} = \bar{S}(r) - \sum_{i \in F} \sum_{j \in A_i} \delta_{i,j,0} \mathbf{1}(r^{\mathsf{orig}}_{i,j} = r), \bar{\mathbf{w}} = 0) | r \in R\}$[9].
- $\bar{E}_7 := \{(\bar{\nu}^{\mathsf{source}}, \bar{\nu}(i,j,0), \overline{\mathbf{c}} = \underline{\mathbf{c}} = \delta_{i,j,0}, \bar{\mathbf{w}} = \rho_i b_{i,j,\varnothing}) | i \in F, j \in A_i\}$:
  $\delta_{i,j,0} \in \{0,1\}$ is a variable which would be defined shortly, and $b_{i,j,\varnothing}$ is the bid placed by aircraft $a_{i,j}$ on staying parked at the same location.
- $\bar{E}_8 := \cup_{r \in R} \bar{E}_{8,r}$:
  For every $r \in R$, we consider $\mathsf{park}(r,H)$ edges connecting $\bar{\nu}(r,H)$ and $\bar{\nu}^{\mathsf{sink}}$. We denote these edges by $\bar{E}_{8,r}$. For any $q \in [\mathsf{park}(r,H)]$, we denote the weight of the $q-$th edge in $\bar{E}_{8,r}$ by $\bar{\mathbf{w}}_{q,r,H}$, and upper and lower capacity by $\overline{\mathbf{c}}_{q,r,H}$ and $\underline{\mathbf{c}}_{q,r,H}$ respectively. For any $r \in R, q \in [\mathsf{park}(r,H)]$, $\bar{\mathbf{w}}_{q,r,H} = -\lambda(C_{r,H}(q) - C_{r,H}(q-1))$, $\overline{\mathbf{c}}_{q,r,H} = 1$, and $\underline{\mathbf{c}}_{q,r,H} = 0$.

[9] Recall that $\bar{S}(r)$ is the state of occupancy of vertiport r at $t = 1$.

– $\bar{E}_9 := \{(\bar{\nu}(i,j,0), \bar{\nu}(r^{\mathsf{orig}}_{i,j}, 1), \overline{\mathbf{c}} = \underline{\mathbf{c}} = \delta_{i,j,0}, \bar{\mathbf{w}} = 0) | i \in F, j \in A_i\}$.

Remark 13.3.1. *In the preceding construction, the capacity of any outgoing edge (resp. incoming edge) from a node which does not have an incoming edge (resp. outgoing edge), other than $\bar{\nu}^{\mathsf{source}}$ and $\bar{\nu}^{\mathsf{sink}}$, is set to 0.*

Mixed Binary Linear Program Formulation

We concatenate the weight, upper capacity bound, and lower capacity bound of each edge as $\overline{\mathbf{W}} \in \mathbb{R}^{|\bar{E}|}$, $\overline{\mathbf{C}} \in \mathbb{Z}_+^{|\bar{E}|}$, and $\underline{\mathbf{C}} \in \mathbb{Z}_+^{|\bar{E}|}$, respectively. Define an incidence matrix of the graph $\bar{G}$ as $\bar{I} \in \{-1, 0, 1\}^{|\bar{V}| \times |\bar{E}|}$, where

$$\bar{I}_{ij} = \begin{cases} 1, & \text{if edge } j \text{ ends at vertex } i, \\ -1, & \text{if edge } j \text{ starts from vertex } i, \\ 0, & \text{otherwise.} \end{cases} \tag{13.11}$$

Defining a truncated incidence matrix $\bar{I}_\star$ obtained from $\bar{I}$ by removing rows corresponding to $\bar{\nu}^{\mathsf{source}}$ and $\bar{\nu}^{\mathsf{sink}}$, we have the following optimization problem.

$$\max_{\mathbf{A}, \delta} \ \overline{\mathbf{W}}^\top \mathbf{A} \tag{13.12a}$$

$$\text{s.t.} \quad \bar{I}_\star \mathbf{A} = \mathbf{0} \tag{13.12b}$$

$$\underline{\mathbf{C}}(\delta) \leqslant \mathbf{A} \leqslant \overline{\mathbf{C}}(\delta) \tag{13.12c}$$

$$\sum_{\tau \in T^{\mathsf{dep}}_{i,j}} \delta_{i,j,\tau} = 1, \forall\ i \in F, j \in A_i \tag{13.12d}$$

$$\delta_{i,j,\tau} \in \{0,1\}, \forall\ i \in F, j \in A_i, \tau \in T^{\mathsf{dep}}_{i,j} \tag{13.12e}$$

$$\mathbf{A} \in \mathbb{Z}_+^{|\bar{E}|} \tag{13.12f}$$

$$\mathbf{A}_{q+1,r,t} \leqslant \mathbf{A}_{q,r,t}, \forall r \in R, t \in [T], q \in [\mathsf{park}(r,t) - 1]. \tag{13.12g}$$

Here, (13.12b) denotes the "flow balance" constraint at every node in $\bar{V} \backslash \bar{V}_3$; (13.12c) denotes the capacity constraints where we have explicitly denoted the dependence of constraints on δ (cf. definitions of $\bar{E}_4$ and $\bar{E}_7$); (13.12d) and (13.12e) denote the constraint that each aircraft must be allocated exactly one route; (13.12f) denotes the integrality constriants; (13.12g) denotes additional constraints which require that edges in $\bar{E}_{3,r,t}$ and $\bar{E}_{8,r}$ are allocated in an increasing order.

Next, we highlight the connection between the optimization problems (13.6) and (13.12).

Lemma 13.3.1. *Given the values of $\boldsymbol{A}_e$ for $e \in \bar{E}_3 \cup \bar{E}_5 \cup \bar{E}_8$ that satisfy the capacity constraints* (13.12c)*, there exists a unique feasible solution $(\boldsymbol{A}, \delta)$ that satisfies* (13.12b)-(13.12g)*.*

Proof. First, note that any feasible solution to (13.12b)-(13.12g) has the same value of $\mathbf{A}_e(x)$ for $e \in \bar{E}_6 \cup \bar{E}_7 \cup \bar{E}_9$ since the lower and upper bound on capacity are the same on these edges by construction. Thus, it is sufficient to show that the values of $\mathbf{A}_e$ for $e \in \bar{E}_3 \cup \bar{E}_5 \cup \bar{E}_6 \cup \bar{E}_7 \cup \bar{E}_8 \cup \bar{E}_9$ uniquely determine a feasible solution $(\mathbf{A}, \delta)$ that satisfies (13.12b)-(13.12g). Particularly, we will show that we can uniquely recover the values of $\mathbf{A}_e$ for $e \in \bar{E}_1 \cup \bar{E}_2 \cup \bar{E}_4$.

To show this claim, we leverage the flow balance constraint (13.12b) at every node. Below, we state the incoming and outgoing edges from every type of node in the network.

Vertex	Incoming Edges	Outgoing Edges
$\bar{\nu}^{\mathsf{arr}}(r,t)$	$\bar{E}_5$	$\bar{E}_1$
$\bar{\nu}(i,j,\tau)$	$\bar{E}_4$	$\bar{E}_5$
$\bar{\nu}^{\mathsf{dep}}(r,t)$	$\bar{E}_2$	$\bar{E}_4$
$\bar{\nu}(i,j,0)$	$\bar{E}_7$	$\bar{E}_9$
$\bar{\nu}(r,t)$	$\bar{E}_1, \bar{E}_3, \bar{E}_6, \bar{E}_9$	$\bar{E}_2, \bar{E}_3, \bar{E}_8$

Note that flow balance at nodes of the form $\bar{\nu}^{\mathsf{arr}}(r,t)$ will determine the values $\mathbf{A}_e$ on edge $\bar{E}_1$, as we know these values for edges $\bar{E}_5$. Next, flow balance at nodes of the form $\bar{\nu}(i,j,\tau)$ will determine the values $\mathbf{A}_e$ on edge $\bar{E}_4$, as we know these values for edges $\bar{E}_5$. This and the capacity constraints on $\bar{E}_4$, ensure that we know the value of δ. Next, flow balance at nodes of the form $\bar{\nu}^{\mathsf{dep}}(r,t)$ will determine the values $\mathbf{A}_e$ on edge $\bar{E}_2$, as we can uniquely determine these values on $\bar{E}_4$. Finally, flow balance at nodes of the form $\bar{\nu}(r,t)$ will determine the values $\mathbf{A}_e$ on edge $\bar{E}_2$, as we can uniquely determine these values on $\bar{E}_1 \cup \bar{E}_3 \cup \bar{E}_6 \cup \bar{E}_8 \cup \bar{E}_9$. □

Proposition 13.3.2. *Suppose* $(A^\dagger, \delta^\dagger)$ *is an optimal solution to* (13.12). *Then* $\overline{\boldsymbol{W}}^\top \boldsymbol{A}^\dagger = \max_{x \in X} \mathit{SW}(x; B)$. *Additionally, using* $A^\dagger$ *we can uniquely determine* $x^\dagger \in X$ *such that* $x^\dagger \in \arg\max_{x \in X} \mathit{SW}(x; B)$.

Proof. First, we show that, for every $x \in X$, there exists a unique $(\mathbf{A}(x), \delta(x))$ satisfying (13.12b)-(13.12g) and $\overline{\mathbf{W}}^\top \mathbf{A}(x) = \mathsf{SW}(x; B)$. Indeed, we construct $(\mathbf{A}(x), \delta(x))$ such that

(i) for every $e \in \bar{E}_5$, where $(\bar{\nu}(i,j,t^{\mathsf{dep}}_{i,j,k}), \bar{\nu}^{\mathsf{arr}}(r^{\mathsf{dest}}_{i,j,k}, t^{\mathsf{arr}}_{i,j,k})) \in e$ for some $i \in F, j \in A_i, k \in m_{i,j}$, it holds that $\mathbf{A}_e(x) = x_{i,j,k}$;

(ii) for every $r \in R, t \in [T]$, and $q \in [\mathsf{park}(r,t)]$, it holds that $\mathbf{A}_e(x) = \mathbf{1}(q \leqslant S(r,t,x))$, where e is the $q-$th edge in $\bar{E}_{3,r,t} \cup \bar{E}_{8,r}$.

The above construction specifies the values of $\mathbf{A}_e(x)$ for $e \in \bar{E}_3 \cup \bar{E}_5 \cup \bar{E}_8$. Additionally, by Lemma 13.3.1, there exists a unique feasible solution $(\mathbf{A}(x), \delta(x))$, and we get

$$\overline{\mathbf{W}}^\top \mathbf{A}(x) = \sum_{e \in \bar{E}} \bar{\mathbf{w}}_e \mathbf{A}_e(x) = \sum_{e \in \bar{E}_3 \cup \bar{E}_5 \cup \bar{E}_7 \cup \bar{E}_8} \bar{\mathbf{w}}_e \mathbf{A}_e(x),$$

where the last equality holds because $\bar{\mathbf{w}}_e = 0$ for $e \in \bar{E}_1 \cup \bar{E}_2 \cup \bar{E}_4 \cup \bar{E}_6 \cup \bar{E}_9$.

Then, we examine each term. First, observe the following.

$$\sum_{e\in\bar{E}_5} \bar{\mathbf{w}}_e \mathbf{A}_e(x) = \sum_{i\in F} \sum_{j\in A_i} \sum_{k\in m_{i,j}} \rho_i b_{i,j,k} x_{i,j,k}.$$
$$\sum_{e\in\bar{E}_7} \bar{\mathbf{w}}_e \mathbf{A}_e(x) = \sum_{i\in F} \sum_{j\in A_i} \rho_i b_{i,j,\phi} x_{i,j,0}.$$

Next, we use the definition of weights in $\bar{E}_{3,r,t}$.

$$\begin{aligned}
\sum_{e\in\bar{E}_3} \bar{\mathbf{w}}_e \mathbf{A}_e(x) &= \sum_{r\in R} \sum_{t=1}^{T-1} \sum_{e\in\bar{E}_{3,r,t}} \bar{\mathbf{w}}_e \mathbf{A}_e(x) \\
&= \sum_{r\in R} \sum_{t=1}^{T-1} \sum_{q=1}^{\mathsf{park}(r,t)} \bar{\mathbf{w}}_{q,r,t} \mathbf{A}_{q,r,t}(x) \\
&= -\lambda \sum_{r\in R} \sum_{t=1}^{T-1} \sum_{q=1}^{S(r,t,x)} (C_r(q) - C_r(q-1)) \\
&= -\lambda \sum_{r\in R} \sum_{t=1}^{T-1} C_r(S(r,t,x)).
\end{aligned}$$

Similarly, we get $\sum_{e\in\bar{E}_8} \bar{\mathbf{w}}_e \mathbf{A}_e(x) = -\lambda \sum_{r\in R} C_r(S(r,H,x))$.

To summarize, we obtain

$$\overline{\mathbf{W}}^\top \mathbf{A}(x) = \mathsf{SW}(x;B). \tag{13.13}$$

Using this, we conclude that

$$\begin{aligned}
&\max_{x\in X} \mathsf{SW}(x;B) = \max_{x\in X} \overline{\mathbf{W}}^\top \mathbf{A}(x) \\
&\leqslant \max_{(\mathbf{A},\delta) \text{ s.t. } (13.12b)-(13.12g)} \overline{\mathbf{W}}^\top \mathbf{A} = \overline{\mathbf{W}}^\top \mathbf{A}^\dagger.
\end{aligned} \tag{13.14}$$

Next, we show that for every $(\mathbf{A},\delta)$ satisfying (13.12b)-(13.12g), there exists $x(\mathbf{A},\delta) \in X$ such that $\mathsf{SW}(x(\mathbf{A},\delta)) = \overline{\mathbf{W}}^\top \mathbf{A}$. Indeed, we construct $x(\mathbf{A},\delta)$ such that for every $i \in F, j \in A_i, k \in m_{i,j}$ it holds that $x_{i,j,k} = \mathbf{A}_e$ for $e \in \bar{E}_5$ such that $(\bar{\nu}(i,j,t^{\mathsf{dep}}_{i,j,k}), \bar{\nu}^{\mathsf{arr}}(r^{\mathsf{dest}}_{i,j,k}, t^{\mathsf{arr}}_{i,j,k})) \in e$ or $e \in \bar{E}_9$ such that $(\bar{\nu}(i,j,0), \bar{\nu}(r^{\mathsf{dest}}_{i,j,k}, 1)) \in e$. Note that due to capacity constraints on these edges, $x_{i,j,k} \in \{0,1\}$. Additionally, the flow balance at the nodes of the form $\bar{\nu}(i,j,\tau)$, for some $i \in F, j \in A_i, \tau \in T^{\mathsf{dep}}_{i,j}$, ensures that

$$\begin{aligned}
\delta_{i,j,\tau} =& \sum_{k\in m_{i,j}} \sum_{e\in\bar{E}_9} \mathbf{A}_e \mathbf{1}((\bar{\nu}(i,j,0), \bar{\nu}(r^{\mathsf{dest}}_{i,j,k}, 1)) \in e, \tau = 0) \\
&+ \sum_{k\in m_{i,j}} \sum_{e\in\bar{E}_5} \mathbf{A}_e \mathbf{1}((\bar{\nu}(i,j,t^{\mathsf{dep}}_{i,j,k}), \bar{\nu}^{\mathsf{arr}}(r^{\mathsf{dest}}_{i,j,k}, t^{\mathsf{arr}}_{i,j,k})) \in e, \tau = t^{\mathsf{dep}}_{i,j,k}).
\end{aligned}$$

Summing over τ, we get

$$\begin{aligned}\sum_{\tau \in T^{\mathsf{dep}}_{i,j}} \delta_{i,j,\tau} &= \sum_{k \in m_{i,j}} \sum_{e \in \bar{E}_9} \mathbf{A}_e \mathbf{1}((\bar{\nu}(i,j,0), \bar{\nu}(r^{\mathsf{dest}}_{i,j,k}, 1)) \in e) \\ &\quad + \sum_{k \in m_{i,j}} \sum_{e \in \bar{E}_5} \mathbf{A}_e \mathbf{1}((\bar{\nu}(i,j,t^{\mathsf{dep}}_{i,j,k}), \bar{\nu}^{\mathsf{arr}}(r^{\mathsf{dest}}_{i,j,k}, t^{\mathsf{arr}}_{i,j,k})) \in e) \\ &= \sum_{k \in m_{i,j}} x_{i,j,k}(\mathbf{A}, \delta).\end{aligned}$$

Using (13.12d), we conclude that $\sum_{k \in m_{i,j}} x_{i,j,k}(\mathbf{A}, \delta) = 1$.

Next, we use the flow balance at nodes of the form $\bar{\nu}^{\mathsf{arr}}(r,t)$, for every $r \in R, t \in [T]$, to ensure that

$$\sum_{\substack{i \in F, j \in A_i, \\ k \in m_{i,j}, e \in \bar{E}_5}} \mathbf{A}_e \mathbf{1}((\bar{\nu}(i,j,t^{\mathsf{dep}}_{i,j,k}), \bar{\nu}^{\mathsf{arr}}(r^{\mathsf{dest}}_{i,j,k}, t^{\mathsf{arr}}_{i,j,k})) \in e, r^{\mathsf{dest}}_{i,j,k} = r, t^{\mathsf{arr}}_{i,j,k} = t) \leqslant \mathsf{arr}(r,t).$$

By using

$$\sum_{e \in \bar{E}_5} \mathbf{A}_e \mathbf{1}((\bar{\nu}(i,j,t^{\mathsf{dep}}_{i,j,k}), \bar{\nu}^{\mathsf{arr}}(r^{\mathsf{dest}}_{i,j,k}, t^{\mathsf{arr}}_{i,j,k})) \in e) = x_{i,j,k}(\mathbf{A}, \delta),$$

we get[10]

$$\sum_{i \in F} \sum_{j \in A_i} \sum_{k \in m_{i,j}} x_{i,j,k} \mathbf{1}(r^{\mathsf{dest}}_{i,j,k} = r, t^{\mathsf{arr}}_{i,j,k} = t) \leqslant \mathsf{arr}(r,t).$$

Analogously, the flow balance equations at the nodes of the form $\bar{\nu}^{\mathsf{dep}}(r,t)$, for some $r \in R, t \in [T]$, ensure that $\sum_{i \in F} \sum_{j \in A_i} \sum_{k \in m_{i,j}} x_{i,j,k} \mathbf{1}(r^{\mathsf{orig}}_{i,j} = r, t^{\mathsf{dep}}_{i,j,k} = t) \leqslant \mathsf{dep}(r,t)$. Finally, we can establish $S(r,t,x(\mathbf{A},\delta)) = \sum_{q=1}^{\mathsf{park}(r,t)} \mathbf{A}_{q,r,t}$ through the flow balance equation at $\bar{\nu}(r,t)$ and (13.1). Since $\sum_{q=1}^{\mathsf{park}(r,t)} \mathbf{A}_{q,r,t} \leqslant \mathsf{park}(r,t)$, due to the capacity constraints on the edge $\bar{E}_{3,r,t}$, it holds that $S(r,t,x(\mathbf{A},\delta)) \leqslant \mathsf{park}(r,t)$. Thus, we conclude that $x(\mathbf{A},\delta) \in X$.

Additionally, using the analysis to show (13.13) in the backward direction and the construction of $x(\mathbf{A},\delta)$, we can establish that $\mathsf{SW}(x(\mathbf{A},\delta)) = \overline{W}^\top \mathbf{A}$. Thus, we conclude that

$$\begin{aligned}\overline{\mathbf{W}}^\top \mathbf{A}^\dagger &= \max_{(\mathbf{A},\delta) \text{ s.t. } (13.12b)-(13.12g)} \overline{\mathbf{W}}^\top \mathbf{A} \\ &= \max_{(\mathbf{A},\delta) \text{ s.t. } (13.12b)-(13.12g)} \mathsf{SW}(x(\mathbf{A},\delta)) \leqslant \max_{x \in X} \mathsf{SW}(x).\end{aligned} \tag{13.15}$$

By (13.14) and (13.15), we get $\overline{\mathbf{W}}^\top \mathbf{A}^\dagger = \max_{x \in X} \mathsf{SW}(x)$. □

[10]When $t = 1$, the arrival capacity constraints are trivially satisfied since there is no incoming aircraft.

Reduction to Mixed Binary Linear Program

Instead of solving (13.12), we can obtain $(\mathbf{A}^\dagger, \delta^\dagger)$ by solving the following MBLP. We establish this fact in Proposition 13.3.3.

$$\max_{\mathbf{A},\delta} \ \overline{\mathbf{W}}^\top \mathbf{A} \tag{13.16a}$$

$$\text{s.t. } (13.12\text{b}) - (13.12\text{e}) \tag{13.16b}$$

$$\mathbf{A} \in \mathbb{R}_+^{|\bar{E}|}. \tag{13.16c}$$

Proposition 13.3.3. *The optimal values of* (13.12) *and* (13.16) *are equal.*

Proof. First, we prove that we can drop (13.12g) when solving (13.12). Suppose there exists $r \in R, t \in [T]$, $q \in [\mathsf{park}(r,t) - 1]$ such that $\mathbf{A}^\dagger_{q,r,t} < \mathbf{A}^\dagger_{q+1,r,t}$. By swapping the value of $\mathbf{A}^\dagger_{q+1,r,t}$ with that of $\mathbf{A}^\dagger_{q,r,t}$, we get a new feasible allocation with a weakly higher objective value. This is because $\bar{\mathbf{w}}_{q+1,r,t} \leqslant \bar{\mathbf{w}}_{q,r,t}$ as $\lambda \geqslant 0$ and $C_{r,t}(\cdot)$ is discrete convex. Then, for any feasible value of δ, the optimization problem (13.12) is an integer linear program where the constraint matrix $\bar{I}_\star$ satisfies total unimodularity, so it is guaranteed to have an integral solution [371, Chapter 19]. □

For any fixed values of binary variables $(\delta_{i,j,\tau})_{i\in F, j\in A_i, \tau\in T^{\mathsf{dep}}_{i,j}}$, the optimization problem (13.12) is a maximum-weight flow problem. Thus, one can enumerate all the departure time combinations, and solve each maximum-weight flow problem with the number of scenarios being $\prod_{i\in F, j\in A_i} |\{t^{\mathsf{dep}}_{i,j,k} | k \in m_{i,j}\}|$. The complete problem can be solved efficiency using the above MBLP approach, which will provide speed-up due to some techniques implemented in commercial solvers such as branch and bound, cutting-plane methods, etc.

13.4 Discussions

We show how the proposed mechanism generalizes existing works in Subsection 13.4 and present some extensions in Subsection 13.4.

Connections to Existing Mechanisms

We consider $H = 1$, $\mathsf{arr}(r,1) = \infty$, $\mathsf{dep}(r,1) = \infty$, $\forall r \in R$, and $|A_i| = 1$, $\forall i \in F$.

(i) **Air Traffic Protocol:** When we treat each vertiport $r \in R$ as a sector with $\mathsf{park}(r,1)$ being the sector capacity, our model generalizes the problem studied in [346], where the authors did not consider arrival and departure capacities and assumed single-aircraft FOs.

(ii) **Airport Time Slot Auction:** When we treat each vertiport $r \in R$ as a time slot with $\mathsf{park}(r, 1)$ being the slot capacity, our model subsumes the framework in [122]. Therefore, our formulation becomes a two-sided matching problem as detailed in [122] and is subject to a faster strongly polynomial-time algorithm.

Extensions of the Proposed Mechanism

(i) **Arrival, Departure, and Airborne Congestion:** To consider congestion due to arriving and departing aircraft, we can apply the same technique in $\bar{E}_3$ to $\bar{E}_1$ and $\bar{E}_2$ by constructing corresponding edge weights. To consider airborne congestion, we treat waypoints in the airspace as vertiports and setting corresponding capacities and congestion costs.

(ii) **External Demand:** Aircraft that are not available in the service area of the SP at $t = 0$ can be incorporated in our framework by setting $r^{\mathsf{orig}}_{i,j} = \mathcal{O}$ and $t^{\mathsf{dep}}_{i,j,k} = 0, \forall k \in m_{i,j}$[11].

(iii) **Entire Trajectory:** We can extend each route to an entire trajectory with multiple vertiport-time pairs. By setting a binary variable for each route and combining those variables when two routes only differ in one time slot, we can apply the same MBLP approach.

(iv) **Cancellation Policy:** It is possible to cancel or re-allocate some of the previously scheduled flights due to changing vertiport capacities or newly emerging aircraft. While there is no single re-allocation policy, it is typical to consider three aspects: congestion, efficiency, and fairness, where we cancel flights from congested vertiports, with low valuations, or at random, respectively.

13.5 Concluding Remarks

In this chapter, we propose an auction mechanism to incentivize fleet operators to report their valuations truthfully and consequently perform a socially optimal allocation of vertiport access. This approach adapts the popular Vickrey–Clarke–Groves mechanism while considering the egalitarian, congestion-aware, and computational issues. The proposed framework could be of interest beyond air traffic management, such as multi-robot coordination.

[11] In this case, $r^{\mathsf{orig}}_{i,j}$ and $t^{\mathsf{dep}}_{i,j,k}$ do not affect our analysis, so we can set them arbitrarily.

Chapter 14

Privacy Preserving Market Mechanisms

Integrating advanced air mobility (AAM) into existing air traffic management (ATM) systems presents complex and unresolved challenges. Projections of AAM traffic density and operational complexity have raised concerns about the scalability of traditional ATM infrastructure. The Federal Aviation Administration (FAA) has acknowledged these challenges [4, 136], stating:

> "Given the number, type, and duration of UAS operations envisioned, the existing ATM system infrastructure and associated resources cannot cost-effectively scale to deliver services for UAS." — *FAA UAS/UTM Con-Ops* ([4])

The limitations of conventional ATM approaches become evident when examining their underlying design principles. Existing systems [49, 50, 48, 355, 323, 29, 30] are designed to manage fixed-wing aircraft operating between established airports, with flight schedules planned weeks or months in advance to minimize overall delays. In contrast, AAM introduces a fundamentally different paradigm: a high volume of electric vertical take-off and landing (eVTOL) aircraft and UAVs operating on demand and pursuing diverse objectives. These vehicles will not only travel between fixed vertiports but also serve ad-hoc destinations—such as residential areas for package deliveries—further straining legacy ATM systems.

To address this challenge, the FAA [4, 136] and other Air Navigation Service Providers (ANSPs) worldwide [131] have stated that the day-to-day traffic management of AAM operations will be delegated to third-party service providers (SPs). These providers will coordinate directly with AAM vehicles to allocate airspace efficiently and safely, thus reducing the dependence on the FAA.

Beyond operational challenges, the introduction of third-party SPs in AAM systems raises additional concerns, particularly related to privacy. A key issue stems from the heterogeneous private valuations of AAM vehicles, which can vary significantly depending on their specific use cases. For instance, a passenger air taxi may have strict scheduling constraints, whereas a regional cargo flight might be more flexible with delays [385, 375]. This variability in

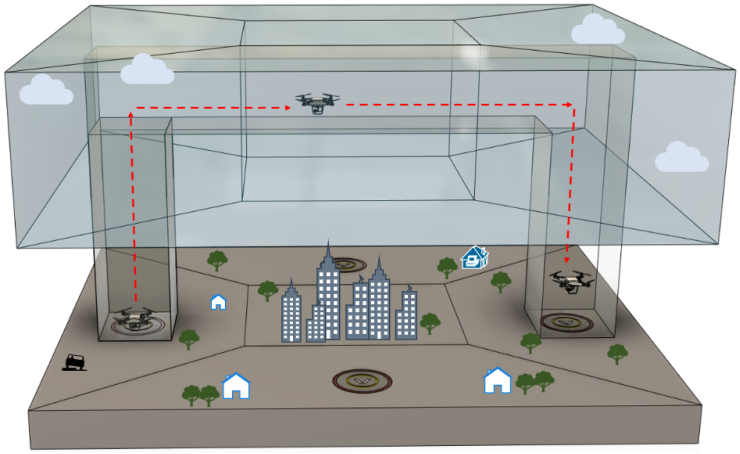


Figure 14.1: A schematic of a city with 3D airspace segmented into regions. Drones depart from vertiports, ascend to cruising altitude, traverse horizontal paths across regions, and descend to land at distant pads.

preferences, coupled with the sensitivity of business and personal data, makes AAM operators reluctant to disclose their private valuation information to SPs.

Against this backdrop, this work aims to answer the following question:

> How can SPs allocate capacity-constrained airspace resources to dynamically arriving AAM vehicles with heterogeneous private valuations, in a way that enables these vehicles to achieve an (approximately) optimal allocation based on their valuations, without requiring them to disclose this private information to the SPs?

We address this question by focusing on a single SP managing access to a specific region of airspace, as this problem remains an open challenge even under a single SP.

We model the airspace as a set of contiguous regions, each with specific capacity constraints on the number of AAM vehicles (modeled as eVTOLs) that can arrive, depart, or remain in a given region at any time (see Fig. 14.1). Vehicles submit requests for airspace access through a menu of feasible (discrete-time) time-trajectories (or air corridors), where each trajectory corresponds to a sequence of tuples specifying the time step and the sector the vehicle wishes to occupy at that time. We formulate the allocation of these time-trajectories within capacity-constrained airspace as a path allocation problem on a time-extended graph (Definition 14.1.1), where all capacity constraints are represented as constraints on the graph's edges.

To prevent monopolization of airspace by major players, we introduce an artificial currency-based auction mechanism [385]. Furthermore, to accommodate the dynamically arriving requests of AAM vehicles, we propose implementing this auction mechanism in a receding-

horizon manner (see Section 14.2). This approach periodically collects AAM vehicle requests and determines allocations based on the proposed auction mechanism.

In each of these auction mechanisms, the SP allocates "air-credits" (the artificial currency) to each AAM vehicle requesting airspace access and charges an anonymous price (in air-credits) for using different airspace regions (i.e., resources on the time-extended graph). Based on these prices, each vehicle selects its most preferred time-trajectory on the time-extended graph, maximizing its valuation while adhering to its budget constraint. The SP's objective is to design prices and allocate airspace efficiently and safely, guided by the following desiderata: (i) Given the price vector, the SP's allocation should be optimal for every AAM vehicle, ensuring that each vehicle maximizes its private valuation subject to budget constraints; (ii) the capacity constraints of the airspace must be respected; (iii) prices must be nonnegative, and if strictly positive, the capacity of each airspace region must match its demand. An allocation-price tuple satisfying these conditions is known as a *competitive equilibrium* in economics, which may not always exist [74]. However, drawing inspiration from Fisher markets under linear constraints [193], we establish the existence of a *fractional competitive equilibrium*—a relaxation of the competitive equilibrium that allows fractional allocations (Proposition 14.3.1). Moreover, we demonstrate that the prices at a fractional competitive equilibrium can be computed as the optimal dual multipliers of a *budget-adjusted welfare problem* (see (14.3)), which is a convex optimization problem (Lemma 14.3.2). Notably, the budget adjustment for each vehicle is determined by the optimal dual multiplier associated with a linear constraint of this optimization problem. Consequently, computing a fractional competitive equilibrium reduces to solving a fixed-point problem (Proposition 14.3.3).

Building on these theoretical insights, we propose a *two-step algorithmic procedure* to allocate AAM vehicles to airspace. In the first step, we develop a *two-loop* algorithm to compute the fractional competitive equilibrium without requiring information about the vehicles' private valuations. Specifically, this step involves solving the fixed-point problem stated in Proposition 14.3.3 using a two-loop algorithm (see Algorithm 11) that mimics fixed-point iteration. The inner loop solves a reformulated budget-adjusted welfare problem (cf. (14.4)) in a distributed manner using the Alternating Direction Method of Multipliers (ADMM) (see Appendix L.3 for a review of ADMM). This ensures that AAM vehicles do not need to share their valuations with the SP or other AAM vehicles. The outer loop then updates the budget adjustment parameter using the latest value of the dual multiplier associated with each AAM vehicle's individual constraints.

Algorithm 11 can be interpreted as an online learning process, where AAM vehicles iteratively refine their trajectory choices based on anonymized market signals, such as expected demand and prices, while the SP dynamically adjusts these signals based on observed demand from AAM vehicles. This enables the SP to estimate equilibrium prices without access to private valuations of AAM vehicles while vehicles deconflict through indirect, price-mediated coordination, eliminating the need for direct trajectory or valuation sharing. This mechanism aligns with the literature on online learning of market mechanisms (cf. [212, 246, 54]), as both the agents and the SP update their strategies sequentially based on observed feedback.

In the second step (i.e., Algorithm 12), we derive an integral allocation from the fractional competitive equilibrium obtained in the first step while keeping the prices unchanged. We rank the vehicles according to the fractional allocation they received for their most desired resource in the first step. The SP then allocates resources to the vehicles sequentially according to this ranking, updating the remaining capacity after each allocation (Algorithm 12). Importantly, in both Algorithms 11–12, the SP requires only information on resource demands, feasible time trajectories, and the most desired path of each AAM vehicle, without accessing any private valuation data. Likewise, individual AAM vehicles do not obtain information about the time trajectories or private valuations of other vehicles, thereby ensuring privacy is preserved throughout the process.

To validate the effectiveness of our approach, we analyze drone-based package delivery using a dataset of drone trajectories generated by Airbus over the city of Toulouse, France, using realistic physical models. A further study of the scheduling problem for electric air taxis on a hypothetical air traffic network in Northern California, United States, is provided in Appendix L.4.

Related Works

The literature on market mechanisms for airspace management in Advanced Air Mobility (AAM) remains relatively limited [375]. Recent works [346, 96] have explored airspace management using second-price auctions combined with congestion management algorithms, as well as combinatorial auctions [237]. However, these approaches are restricted to unit-capacity regions, limiting their applicability to real-world AAM systems that require flexible allocation across multiple capacity-constrained airspace sectors.

Su et al. [397] addressed these limitations by employing a generalized Vickrey–Clarke–Groves (VCG) auction for AAM resource management, incorporating considerations of social welfare, safety, and congestion. Their approach ensures proportional fairness by optimizing a social cost function based on the weighted utilities of all fleet operators. A key design goal in their work is to elicit truthful preferences from AAM vehicles to enable efficient airspace allocation. However, this reliance on truthful bidding raises privacy concerns, as vehicles must disclose their exact valuations through bids—potentially exposing sensitive operational information to competitors or regulators. Balakrishnan and Chandran [28] proposed a column generation algorithm that iteratively updates prices to determine an allocation satisfying capacity constraints. However, their method relies on vehicles reporting their private valuations, which may raise privacy concerns. In contrast, our approach eliminates the need for AAM vehicles to disclose private valuations, thereby enhancing privacy while still achieving efficient allocation.

A distinguishing feature of our approach, compared to existing market mechanisms for AAM, is the use of artificial currency. While monetary transactions are effective in eliciting preferences, as noted by [375], they can disproportionately advantage operators with greater financial resources. Our approach mitigates this issue by introducing a system of artificial currency designed to promote fairness. The idea of using artificial currency for fair and

efficient resource allocation has been extensively studied in economics, beginning with [411]. These works typically consider environments in which agents are endowed with artificial currency that holds no value outside the market. However, most of these studies focus on the allocation of substitutable goods, whereas our setting requires agents to select multiple goods over a time-extended network, resulting in complementarities rather than substitutability. Artificial currency mechanisms that address complementarities have been explored in [74, 75]. The key distinction of our work is that agents are allowed to save currency for future use, and the saved budget directly influences their utility in a quasi-linear manner. This feature aligns our model more closely with combinatorial auction environments.

Combinatorial auctions enable participants to bid on bundles of items rather than on individual items [108]. Due to the exponential complexity inherent in these auctions, no single format universally applies across all settings. In environments with budget constraints, iterative auction formats—such as the Simultaneous Ascending Auction, the Ascending Proxy Auction, and the Clock-Proxy Auction [23, 109]—are particularly relevant. These formats allow agents to observe current prices and iteratively adjust their bids within budget constraints prior to final submission. Additionally, they offer a privacy advantage by enabling incremental bid submission, thereby reducing the exposure of complete preference information. However, these mechanisms may suffer in efficiency when applied to combinatorial auction settings with strong complementarities. In our approach, we propose a novel method for modeling complementarities using linear equality constraints and leverage recent advances in Fisher markets with linear constraints to determine allocations. Specifically, we compute a fractional competitive equilibrium, which yields a relaxed solution that is easier to compute. From this fractional equilibrium, we then derive an integral allocation that satisfies the problem's capacity constraints.

Organization The chapter is organized as follows. In Section 14.1, we introduce the model of airspace management studied in this chapter. Section 14.2 presents a high-level overview of our approach, which implements an artificial-currency-based auction mechanism in a receding-horizon manner. In Section 14.3, we formally describe this auction mechanism and provide theoretical results on fractional competitive equilibrium. Section 14.4 outlines the algorithmic procedure used to compute an approximate market mechanism. In Section 14.5, we validate the performance of our mechanism using a drone delivery dataset generated from a realistic drone dynamics model developed by Airbus. We discuss the limitations of our approach in Section 14.6 and conclude in Section 14.7.

14.1 Model of Advanced Air Mobility

Consider the problem of allocating airspace to Advanced Air Mobility (AAM) vehicles, such as drones and eVTOLs, which enable novel AAM services in urban environments. In our model, we segment the urban airspace into contiguous regions or sectors, denoted by $\mathcal{R}$. The spatial configuration of the airspace is represented as a graph $\mathcal{G} = (\mathcal{R}, \mathcal{E})$, where $\mathcal{R}$

corresponds to the set of vertices, and $\mathcal{E} \subseteq \mathcal{R} \times \mathcal{R}$ denotes the set of edges connecting adjacent regions, indicating feasible movements for AAM vehicles. To account for the temporal dimension, we divide the day into T time steps, with each time step comprising τ seconds. AAM vehicles arrive dynamically, each requesting access to the airspace.

Each AAM vehicle has a feasible set of time trajectories (also referred to as "air corridors") that it can utilize within the airspace. Each vehicle can independently determine its set of feasible trajectories by accounting for energy consumption, travel time, and other operational factors. Furthermore, the feasible set includes only time trajectories that start before the vehicle's takeoff; mid-flight trajectory changes are not permitted. Fig. 14.2 illustrates a schematic representation of a time trajectory for a package delivery scenario. For each feasible trajectory, the AAM vehicle generates a *private valuation* that may differ across trajectories, reflecting its preferences.

In our model, we assume that each region has a limit on the number of vehicles that can simultaneously arrive, depart, or remain in that region at any given time[1]. Due to capacity constraints, it may not be feasible to allocate each AAM vehicle its *most preferred* time trajectory, as this could violate airspace constraints. In such a scenario, a service provider (SP) is typically responsible for managing the airspace and assigning each AAM vehicle a feasible time trajectory while ensuring compliance with airspace constraints. To facilitate this process, we introduce the framework of a *time-extended graph*, which is essential for integrating arrival, departure, and transit constraints when allocating time trajectories to AAM vehicles.

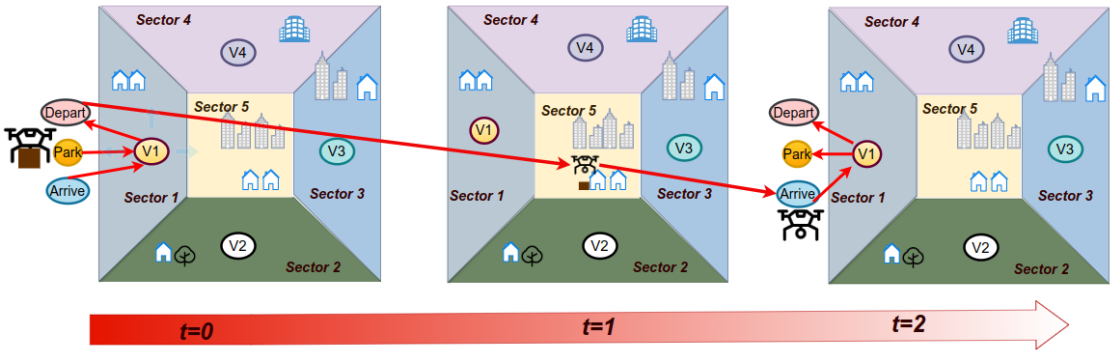


Figure 14.2: A time trajectory diagram of a drone delivering a package in an urban setting. The drone starts from the launch pad V1 (Sector 1) and needs to drop a package in Sector 5 before returning. Here, we show a simple trajectory that moves between regions in one time step, but in general, such trajectories can remain in any region for multiple time steps.

[1]This constraint arises from safety concerns that limit the number of vehicles that can be autonomously de-conflicted in a confined space [37].

Definition 14.1.1 (Time-extended Graph)**.** *We define* $\tilde{\mathcal{G}} = (\tilde{\mathcal{R}}, \tilde{\mathcal{E}})$ *as the* time-extended graph *with horizon* T*, for some positive integer* T*, such that*

(i) $\tilde{\mathcal{R}} = \cup_{t=1}^{T} \cup_{r \in \mathcal{R}} \{\nu(r,t), \nu^{\mathsf{arr}}(r,t), \nu^{\mathsf{dep}}(r,t)\}$*, where* $\nu(r,t)$*,* $\nu^{\mathsf{arr}}(r,t)$ *, and* $\nu^{\mathsf{dep}}(r,t)$ *are three replicas of region* r *at time* t*.*

(ii) $\tilde{\mathcal{E}} = \cup_{j=1}^{4} \tilde{\mathcal{E}}^{(j)} \subseteq \tilde{\mathcal{R}} \times \tilde{\mathcal{R}}$*, where*

- $\tilde{\mathcal{E}}^{(1)} := \cup_{t=1}^{T} \{(\nu^{\mathsf{arr}}(r,t), \nu(r,t))\}$.
  Any edge of the type $(\nu^{\mathsf{arr}}(r,t), \nu(r,t))$ *has capacity*[2] $C^{\mathsf{arr}}(r,t)$.
- $\tilde{\mathcal{E}}^{(2)} := \cup_{t=1}^{T} \{(\nu(r,t), \nu^{\mathsf{dep}}(r,t))\}$.
  Any edge of the type $(\nu(r,t), \nu^{\mathsf{dep}}(r,t))$ *has capacity* $C^{\mathsf{dep}}(r,t)$.
- $\tilde{\mathcal{E}}^{(3)} := \cup_{t=1}^{T-1} \{(\nu(r,t), \nu(r,t+1))\}$.
  Any edge of the type $(\nu(r,t), \nu(r,t+1))$ *has capacity* $C^{\mathsf{stay}}(r,t)$.
- $\tilde{\mathcal{E}}^{(4)} := \cup_{t=1}^{T-1} \cup_{(r,r') \in \mathcal{E}} \{(\nu^{\mathsf{dep}}(r,t), \nu^{\mathsf{arr}}(r',t+1))\}$.
  Any edge of the type $(\nu^{\mathsf{dep}}(r,t), \nu^{\mathsf{arr}}(r',t+1))$ *is unconstrained.*

Any time trajectory of an AAM vehicle is a path on this time-extended graph. A simple example describing the time-extended graph and time trajectories is provided in Appendix L.1.

In this work, we propose a market-based mechanism for the service provider (SP) to allocate capacity-constrained airspace infrastructure to dynamically arriving AAM vehicles. Our mechanism ensures that: (i) the capacity constraints of the airspace are strictly satisfied, (ii) each AAM vehicle receives an (approximately) optimal allocation according to its preferences, and (iii) AAM vehicles are not required to disclose their private valuations to the SP. The detailed design and implementation of this mechanism are discussed in the following section.

14.2 High-level Overview of Market-based Mechanism

We propose an auction-based approach that allocates airspace to dynamically arriving AAM vehicles in a *receding-horizon* fashion by *periodically* allocating the available airspace capacity through an auction mechanism. In particular, for a total duration of T time steps (with each time step comprising τ seconds), the service provider (SP) conducts I auctions. In each auction, the SP allocates airspace to AAM vehicles participating in that auction.

[2]The arrival and departure constraints are primarily required for airspace regions comprising of vertiports/launch pads.

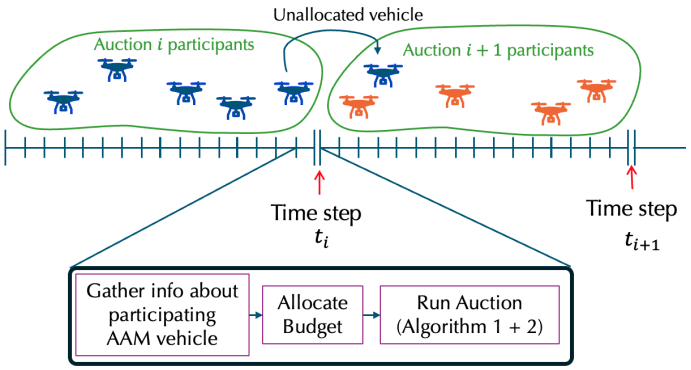


Figure 14.3: A schematic depiction of the receding horizon approach.

The allocation of the i-th auction is determined at time-step $t_i = (i-1)\lfloor T/I \rfloor + 1$. The set of vehicles participating in the i-th auction comprises new vehicles that have requested airspace access from the SP after the $(i-1)$-th auction (i.e., after time-step t_{i-1}), as well as those that were not allocated in previous auctions. Each AAM vehicle has a feasible set of time trajectories that it can follow[3]. In each auction, the SP assigns a time trajectory to each participating AAM vehicle from its feasible set[4]. It is important to emphasize that the flight trajectory of each AAM vehicle is finalized before takeoff, and they are not allowed to participate in subsequent auctions to modify their trajectory mid-flight. Finally, the SP updates the remaining airspace capacity before initiating the next auction. A schematic of the receding-horizon approach is provided in Fig. 14.3.

At a high level, in each auction, the SP allocates a certain amount of "air credits" to each participating AAM vehicle. These air credits, along with other credits they have from past auctions, act as an artificial currency for purchasing airspace access during that auction. Additionally, the SP imposes a payment (in air credits) for the use of each edge in the time-extended graph. AAM vehicles can utilize any leftover budget in future auctions to purchase access to airspace.

The main design component of this auction mechanism is to determine these prices in a way that allows each AAM vehicle to afford an (approximately) optimal time-trajectory—one that maximizes its utility within its allocated budget—while ensuring that the overall airspace allocation adheres to capacity constraints. Moreover, these prices must be computed in a manner that preserves the privacy of AAM vehicles, ensuring that they do not need to disclose their private valuations.

In Section 14.3, we formally present one such auction mechanism used in the receding

[3]Note that the time trajectory allocated in auction i may start at time-step t_i and can end at time-step T.

[4]Note that the feasible set of trajectories for each vehicle includes an option $\varnothing$, which indicates that the vehicle will not take off in that auction (i.e., remain unallocated), ensuring that our problem always has a feasible allocation.

horizon approach discussed above, assuming that the SP has access to private valuations. Additionally, we study the theoretical properties of the proposed auction mechanism. In Section 14.4, we relax the assumption that the SP knows the private valuations of AAM vehicles and develop an algorithmic approach (Algorithm 11-12) to implement the proposed auction mechanism without requiring this knowledge.

14.3 Auction Mechanism: Design and Analysis

In this section, we formally describe the elements of our proposed auction mechanism, along with theoretical guarantees, assuming that the service provider (SP) has access to the private valuations of each AAM vehicle.

Let U be the set of AAM vehicles requesting access to airspace in the current auction. Each AAM vehicle $u \in U$ is allocated a budget of air credits, denoted by $w_u \geqslant 0$[5]. The set of feasible time-trajectories for any AAM vehicle $u \in U$ is given by $M_u = R_u \cup \{\varnothing\}$, where R_u represents a subset of paths on the time-extended graph (cf. Definition 14.1.1), and $\varnothing$ denotes the option to drop out of the system if no feasible path is available due to high congestion.

We define $x_{u,e} \in \{0,1\}$ to indicate whether an AAM vehicle u is using edge $e \in \tilde{\mathcal{E}}$, and $x_{u,\varnothing} \in \{0,1\}$ to indicate whether the AAM vehicle u has dropped out of the system. Furthermore, each AAM vehicle has the option to convert its unused budget into an "outside option" for future auctions[6]. We use $x_{u,\mathsf{o}}$ to denote the amount of outside options that the AAM vehicle consumes in any auction. For concise notation, we define $\mathbf{x}_u = \left(x_{u,e}\right)_{e\in\tilde{\mathcal{E}}}$ and $\bar{\mathbf{x}}_u = \left[\mathbf{x}_u^\top, x_{u,\mathsf{o}}, x_{u,\varnothing}\right]^\top$.

The utility derived by any vehicle $u \in U$ from selecting a route $s \in R_u$ is denoted by $v_{u,s} \in \mathbb{R}_+$, the utility from selecting the option $\varnothing$ is denoted by $v_{u,\varnothing} \in \mathbb{R}_+$, and for the per-unit consumption of the outside option is given by $v_{u,\mathsf{o}} \in \mathbb{R}_+$. Therefore, the overall utility derived by AAM vehicle u is given by

$$f_u(\bar{\mathbf{x}}_u) = \sum_{s\in R_u} v_{u,s} x_{u,e^*(s)} + v_{u,\mathsf{o}} x_{u,\mathsf{o}} + v_{u,\varnothing} x_{u,\varnothing}, \tag{14.1}$$

where $e^*(s)$ denotes the *departing edge*[7] on the route s.

[5]The variation in budgets among AAM vehicles can arise due to two factors: (i) savings accumulated from previous auctions, and (ii) the priority given by the SP, for instance, in the case of disaster relief and emergency service vehicles.

[6]The budget converted into the outside option is carried forward and added to the AAM vehicle's budget in future auctions.

[7]In our framework, we use the convention that the AAM vehicles place all their valuation of a route on the first edge on the time-extended graph that goes out from the origin node. This is an edge of type $\tilde{\mathcal{E}}^{(4)}$ in Definition 14.1.1. Additionally, we add a constraint (cf. (14.3b)) that ensures that if a vehicle selects a departing edge corresponding to a route, then all edges on that route will be selected.

The SP charges a price p_e for any AAM vehicle using the edge $e \in \tilde{\mathcal{E}}$, and a payment $p_{\mathsf{o}} \geqslant 0$ for the consumption per unit of the outside option. Upon observing the prices, each AAM vehicle $u \in U$ solves the following optimization problem:

$$\begin{aligned} \max_{\bar{\mathbf{x}}_u} \quad & f_u(\bar{\mathbf{x}}_u) && \text{(IOP)} \\ \text{s.t.} \quad & \mathbf{p}^\top \mathbf{x}_u + p_{\mathsf{o}} x_{u,\mathsf{o}} \leqslant w_u && \text{(14.2a)} \\ & \tilde{\mathbf{a}}_u^\top \mathbf{x}_u + x_{u,\varnothing} = 1 && \text{(14.2b)} \\ & \tilde{\mathbf{A}}_u \mathbf{x}_u = \mathbf{0} && \text{(14.2c)} \\ & \mathbf{x}_u \in \{0,1\}^{|\tilde{\mathcal{E}}|}, x_{u,\varnothing} \in \{0,1\}, && \text{(14.2d)} \end{aligned}$$

where $\tilde{\mathbf{a}}_u \in \mathbb{R}^{|\tilde{\mathcal{E}}|}$ is such that the constraint (14.2b) enforces $\sum_{s \in R_u} x_{u,e^*(s)} + x_{u,\varnothing} = 1$, indicating that the AAM vehicle u will either select a path in R_u or will drop out. The matrix $\tilde{\mathbf{A}}_u \in \mathbb{R}^{K \times (|\tilde{\mathcal{E}}|)}$ in (14.2c) represents two types of constraints: (i) $\mathbf{A}_u \mathbf{x}_u = \mathbf{0}$, where $\mathbf{A}_u \in \mathbb{R}^{|\tilde{\mathcal{R}}| \times |\tilde{\mathcal{E}}|}$ is an incidence matrix of the time-extended graph encoding flow-balance constraints at each node in $\tilde{\mathcal{G}}$; and (ii) $\mathbf{B}_{u,s} \mathbf{x}_u = 0$ for each $s \in R_u$, where $\mathbf{B}_{u,s} \in \mathbb{R}^{(K - |\tilde{\mathcal{R}}|) \times |\tilde{\mathcal{E}}|}$ encodes the constraint that the flow on any edge connecting two different regions along path s matches the flow on the departing edge $e^*(s)$. Intuitively, (ii) ensures that any feasible solution that satisfies (14.2b) and (i) results in a unique edge flow. We present a simple example in Appendix L.1 to describe these constraints.

In (14.2), (IOP) represents the utility derived by the AAM vehicle u; (14.2a) denotes the budget constraint of AAM vehicle u; (14.2b) represents the requirement that AAM vehicle u must select at least one of the paths in R_u or consider dropping out; (14.2c) ensures a unique edge flow for every feasible solution from (14.2b); and (14.2d) ensures that the selections made by AAM vehicles are integral. Note that the feasible set in (14.2) is non-empty. This is because $\mathbf{x}_u = \mathbf{0}, x_{u,\varnothing} = 1$, and $x_{u,\mathsf{o}} = w_u / p_{\mathsf{o}}$ is always a feasible solution.

The goal of the SP is to set the prices such that the resulting allocation is a *competitive equilibrium*:

Definition 14.3.1. $(\bar{\mathbf{x}}^*, \mathbf{p}^*)$ *is said to be a* competitive equilibrium *if the following conditions are satisfied*

(i) For every $u \in U$, $\bar{\mathbf{x}}_u^*$ is an optimal solution of (14.2) with prices set to $\mathbf{p}^*$;

(ii) The capacity constraints are satisfied. That is, for every $e \in \tilde{\mathcal{E}}$, $\sum_{u \in U} x_{u,e}^* \leqslant \ell_e$.

(iii) $p_e^* \geqslant 0$ for all $e \in \tilde{\mathcal{E}}$; and if $p_e^* > 0$ then $\sum_{u \in U} x_{u,e}^*(p^*) = \ell_e$.

We call $(\bar{\mathbf{x}}^*, \mathbf{p}^*)$ the market clearing allocation and market clearing prices, respectively. In general, a competitive equilibrium may not always exist [74]. Therefore, we introduce a relaxed version of competitive equilibrium, where we relax the requirement that allocations are integral.

Definition 14.3.2. $(\bar{\mathbf{x}}^*, \mathbf{p}^*)$ *is called a* fractional-competitive equilibrium *if all conditions in Definition 14.3.1 are satisfied, except that in Definition 14.3.1-(i) the integral constraint* (14.2d) *is relaxed to a positivity constraint.*

This relaxation is inspired by the competitive equilibrium framework studied in the literature on Fisher markets with linear constraints [193].

In what follows, we demonstrate that a fractional competitive equilibrium always exists and can be computed as the solution to a fixed-point problem. Then, in Section 14.4, we leverage this property to develop a two-step algorithmic procedure that produces an integral allocation and prices approximating a competitive equilibrium, without requiring knowledge of the private valuations of the AAM vehicles.

Existence and Computation of Fractional Competitive Equilibrium

We state the following existence result regarding the fractional competitive equilibrium.

Proposition 14.3.1. *There exists a fractional-competitive equilibrium.*

The proof builds on the result establishing the existence of a competitive equilibrium in Fisher markets with auxiliary inequality constraints [193]. Specifically, our proof accounts for auxiliary *equality* constraints that arise from (14.2b)-(14.2c). A detailed proof of Proposition 14.3.1 is provided in Appendix L.2.

Next, we present a computational framework that can be used by the service provider for computing a fractional-competitive equilibrium, if they know the private valuations of all AAM vehicles. Consider the following optimization problem, parametrized by $\omega \in \mathbb{R}_{\geqslant 0}^{|U|}$:

$$\max_{\bar{\mathbf{x}}=(\bar{\mathbf{x}}_u)_{u\in U}} \quad \sum_{u\in U} (w_u + \omega_u) \log\left(f_u(\bar{\mathbf{x}}_u)\right) - \sum_{u\in U} p_{\mathsf{o}} x_{u,\mathsf{o}} \tag{14.3a}$$

$$\text{s.t.} \quad \sum_{u\in U} x_{u,e} \leqslant \ell_e, \quad \forall\, e \in \tilde{\mathcal{E}}, \tag{14.3b}$$

$$\tilde{\mathbf{a}}_u^\top \mathbf{x}_u + x_{u,\varnothing} = 1, \quad \forall\, u \in U, \tag{14.3c}$$

$$\tilde{\mathbf{A}}_u \mathbf{x}_u = \mathbf{0}, \quad \forall\, u \in U, \tag{14.3d}$$

$$x_{u,e} \geqslant 0, \quad \forall\, u \in U, e \in \tilde{\mathcal{E}} \cup \{\mathsf{o}, \varnothing\}, \tag{14.3e}$$

where the first term in (14.3a) represents the "budget-adjusted" weighted geometric mean of the utilities of all AAM vehicles, while the second term accounts for the total expenditure on the outside option. Constraint (14.3b) enforces the capacity limit on every edge, and constraints (14.3c)–(14.3d) are analogous to (14.2b)–(14.2c). Additionally, (14.3e) is a relaxation of the integrality constraint in (14.2d). The objective in (14.3) is related to the "budget-adjusted social optimization problem" studied in [193] for Fisher markets, with the key difference being the second term, which ensures that a smaller amount of credits is spent

on the outside option. On an intuitive level, the weighted geometric mean structure of the objective in (14.3a) can be seen as finding an allocation that balances the trade-offs between different AAM vehicles' utilities, weighted by their budget adjustments. It ensures that an improvement in a vehicle's utility contributes to the overall objective in proportion to its market power. At the optimal point, this results in a fair and efficient allocation of airspace among AAM vehicles. On a more technical level, the weighted geometric mean is fundamental in the proof of Proposition 14.3.3, which ensures that if we can adjust the weights of AAM vehicles in an appropriate manner (through a careful choice of ω), then the optimal dual multiplier of (3b) is the market-clearing price and the optimizer of (3) yields market-clearing allocations. Before stating Proposition 14.3.3, we present an important property of (14.3).

Lemma 14.3.2. *The constraints* (14.3b)-(14.3e) *always have a feasible solution. Furthermore, for any* $\omega \in \mathbb{R}_+^{|U|}$, (14.3) *is a convex optimization problem.*

The proof of this result follows from the fact that the constraint set in (14.3) is a polytope. Moreover, the objective $f_u(\cdot)$ is a linear function with positive coefficients. For any $\omega \in \mathbb{R}_+^{|U|}$, let $\mathbf{p}^\dagger(\omega)$ denote an optimal dual multiplier corresponding to the constraint (14.3b), $\lambda^\dagger(\omega)$ denote an optimal dual multiplier corresponding to the constraint (14.3c), and $\bar{\mathbf{x}}^\dagger(\omega)$ denote an optimal solution to (14.3).

Proposition 14.3.3. *Suppose there exists* $\omega^* \in \mathbb{R}_+^{|U|}$ *that is a fixed point of the mapping* $\omega \mapsto \lambda^\dagger(\omega)$. *Then* $(\bar{\mathbf{x}}^\dagger(\omega^*), \mathbf{p}^\dagger(\omega^*))$ *is a fractional-competitive equilibrium.*

The proof relies on the convexity of the optimization problems (14.3) (Lemma 14.3.2) and (14.2) (after relaxing the integrality constraint in (14.2d) to fractional in (14.3e)), along with matching the KKT conditions for optimality. The detailed proof of Proposition 14.3.3 is provided in Chapter L.2.

Remark 14.3.1. *The proof of Proposition 14.3.3 can be extended to settings where a fixed point may not exist. Suppose there exists* $\omega^* \in \mathbb{R}_+^{|U|}$ *such that, for each* $u \in U$, $\omega_u^* - \lambda_u^\dagger(\omega^*) = \epsilon_u$ *for some* $\epsilon_u \in \mathbb{R}$ *ensuring that* $w_u + \epsilon_u \geqslant 0$. *Then* $(\bar{\mathbf{x}}^\dagger(\omega^*), \mathbf{p}^\dagger(\omega^*))$ *is a fractional competitive equilibrium of a market where, for each* $u \in U$, *the budget is adjusted to* $w_u + \epsilon_u$.

14.4 Algorithmic Design of Auction Mechanism without Private Valuations

In this section, we outline our algorithmic procedure for the service provider (SP) to compute (approximate) competitive equilibria using Proposition 14.3.3, without requiring knowledge of the private valuations of AAM vehicles. Our approach unfolds in two stages. First, we introduce an algorithm that solves the fixed-point equation from Proposition 14.3.3 to compute the fractional competitive equilibrium in a distributed manner (cf. Algorithm 11). The SP then generates a ranked list of AAM vehicles using the prices and fractional

allocations derived from this step. This ranking allows the SP to obtain an integral allocation by successively assigning regions to AAM vehicles according to the ranking (cf. Algorithm 12). In the following subsections, we elaborate on each of these steps.

Step 1: Distributed Algorithm for Computing Fractional-Competitive Equilibrium

To compute a fractional-competitive equilibrium, we propose Algorithm 11 to solve the fixed-point problem described in Proposition 14.3.3 in a distributed manner. Algorithm 11 emulates the fixed-point iteration for the mapping

$$\omega \mapsto \lambda^{\dagger}(\omega). \tag{FP}$$

Since the SP lacks access to $\lambda^{\dagger}(\omega)$, as computing it requires solving (14.3), which in turn depends on the private valuations of AAM vehicles, we adopt a two-loop approach to circumvent this challenge. In the inner loop, we iteratively solve the convex optimization problem (14.3) in a distributed manner that does not require the SP to access the private valuations of AAM vehicles. This is achieved by repeatedly interacting with the AAM vehicles for a finite number of rounds, while holding ω constant, to approximate $\lambda^{\dagger}(\omega)$. This approximation is then used to update ω using a fixed-point iteration in the outer loop.

To solve the inner-loop problem in a distributed manner, we reformulate (14.3) as (14.4) by introducing two additional variables, $\mathbf{y}$ and $\mathbf{z}$. This reformulation enables the use of distributed optimization techniques, facilitating computation across multiple agents while preserving the structure of the original problem.

$$\begin{aligned}
\min_{(\bar{\mathbf{x}}_u, \mathbf{y}_u)_{u\in U}, (z_e)_{e\in\tilde{\mathcal{E}}}} \quad & \sum_{u\in U} (w_u + \omega_u) \log\left(f_u(\bar{\mathbf{x}}_u)\right) - \sum_{u\in U} p_{\mathsf{o}} x_{u,\mathsf{o}} && \text{(14.4a)}\\
\text{s.t.} \quad & \mathbf{y}_u = \mathbf{x}_u, \quad \forall\, u \in U, && \text{(14.4b)}\\
& \sum_{u\in U} y_{u,e} + z_e = \ell_e, \quad \forall\, e \in \tilde{\mathcal{E}}, && \text{(14.4c)}\\
& \tilde{\mathbf{a}}_u^{\top} \mathbf{x}_u + x_{u,\varnothing} = 1, \quad \forall\, u \in U && \text{(14.4d)}\\
& \tilde{\mathbf{A}}_u \mathbf{x}_u = \mathbf{0}, \quad \forall\, u \in U, && \text{(14.4e)}\\
& \bar{\mathbf{x}}_u \geqslant \mathbf{0}, \mathbf{z} \geqslant \mathbf{0}, \mathbf{y}_u \in \mathbb{R}^{|\tilde{\mathcal{E}}|}, \quad \forall\, u \in U. && \text{(14.4f)}
\end{aligned}$$

In this reformulation, Equation (14.4b) enforces equality between $\mathbf{x}$ and $\mathbf{y}$, while Equation (14.4c) ensures that the capacity constraints are satisfied. Constraints (14.4d)–(14.4f) are identical to (14.3c)–(14.3e), with additional positivity constraints on $\mathbf{y}$ and $\mathbf{z}$. This reformulation ensures that the Lagrangian of (14.4) becomes a separable function of $\bar{\mathbf{x}}_u$, allowing the problem to be solved in a distributed manner using the ADMM algorithm [176, 193]. The variable $\mathbf{y}$ can be interpreted as the "expected allocation" estimate of the service provider (SP), while $\mathbf{z}$ represents the "resource surplus" in each region. Next, we describe the inner

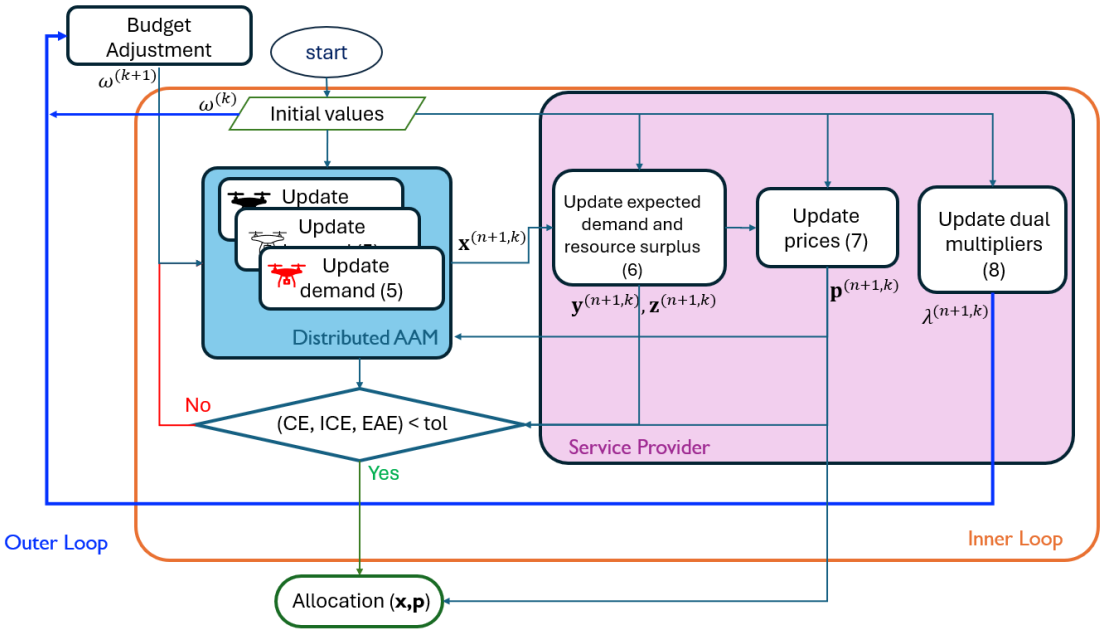


Figure 14.4: Flowchart of Algorithm 11, illustrating the processes executed independently by the SP and AAM vehicles, as well as the steps computed within the inner and outer loops.

and outer loops in detail. We index the inner-loop iterations by $n = 1, 2, \ldots, N$ and the outer-loop iterations by $k = 1, 2, \ldots, K$. A flowchart of our algorithm in Step 1 is shown in Figure 14.4.

Inner Loop: Inner loop iterations are obtained by performing ADMM iterations[8] for (14.4) with step-size parameter β. For any $\omega \in \mathbb{R}^{|U|}_{\geqslant 0}$, this implementation allows us to estimate the dual multiplier $\lambda^{\dagger}(\omega)$ in a distributed manner without requiring knowledge of the private valuations of AAM vehicles. Next, we describe these iterations in more detail.

Given that the outer loop is at iteration k, at any iteration n of the inner loop: (a) each AAM vehicle u keeps track of its individual demand $\bar{\mathbf{x}}_u^{(n,k)}$; and (b) the SP keeps track of four quantities: the estimate of expected allocations $\mathbf{y}^{(n,k)}$, the expected resource surplus $\mathbf{z}^{(n,k)}$, the prices of all regions $\mathbf{p}^{(n,k)}$, and a dual multiplier $\lambda^{(n,k)}$, which is used to adjust the budgets of AAM vehicles.

Local update for each AAM vehicle: Given that the outer loop is at iteration k, at every iteration n of the inner loop, each AAM vehicle u receives its expected allocation $\mathbf{y}_u^{(n,k)}$, the current prices on regions $\mathbf{p}^{(n,k)}$, and the dual multiplier corresponding to its local constraints $\lambda_u^{(n,k)}$. Using this information, the AAM vehicle updates its requested demand using the

[8]Derivation of ADMM updates for (14.4) is provided in Appendix L.3.

Algorithm 11 Distributed Algorithm for Fractional Competitive Equilibrium

Input: $\mathbf{p}^{(0,0)} = \mathbf{0}, \boldsymbol{\lambda}^{(0,0)} = \mathbf{0}, \mathbf{y}^{(0,0)} = \mathbf{0}, \omega^{(0)} = \mathbf{0}, \mathsf{tol}_{\mathsf{CE}}, \mathsf{tol}_{\mathsf{ICE}}, \mathsf{tol}_{\mathsf{EAE}}, \beta$
for $k = 0$ to $K-1$ **do**
  // Outer loop
  for $n = 0$ to $N-1$ **do**
    // Inner loop
    // Distributed Updates by AAM vehicles
    Each $u \in U$ updates $\bar{\mathbf{x}}_u^{(n+1,k)}$ using (14.5)
    // Service Provider (SP) Updates
    Update $\mathbf{y}^{(n+1,k)}$ and $\mathbf{z}^{(n+1,k)}$ using (14.6)
    Update prices $\mathbf{p}^{(n+1,k)}$ using (14.7)
    For each $u \in U$, update $\lambda_u^{(n+1,k)}$ using (14.8)
    if $\mathsf{CE} \leqslant \mathsf{tol}_{\mathsf{CE}}$ and $\mathsf{ICE} \leqslant \mathsf{tol}_{\mathsf{ICE}}$ and $\mathsf{EAE} \leqslant \mathsf{tol}_{\mathsf{EAE}}$ **then**
      Return: $\bar{\mathbf{x}}^\dagger = \bar{\mathbf{x}}^{(n+1,k)},\ \bar{\mathbf{p}}^\dagger = \mathbf{p}^{(n+1,k)}$
    end if
  end for
  Set $\omega^{(k+1)} = \lambda^{(N,k)}$
  Set $\mathbf{p}^{(0,k+1)} = \mathbf{p}^{(N,k)},\ \mathbf{y}^{(0,k+1)} = \mathbf{y}^{(N,k)}$
end for
Return: $\bar{\mathbf{x}}^\dagger = \bar{\mathbf{x}}^{(N,K)},\ \bar{\mathbf{y}}^\dagger = \mathbf{p}^{(N,K)}$

following update and shares this with the SP.

$$\begin{aligned}\bar{\mathbf{x}}_u^{(n+1,k)} = \underset{\bar{\mathbf{x}}_u,\ \text{s.t. (14.3d)}-\text{(14.3e) hold}}{\arg\max} \Bigg(& (w_u + \omega_u^{(k)}) \log\left(f_u(\bar{\mathbf{x}}_u)\right) - p_{\mathsf{o}} x_{u,\mathsf{o}} \\ & - \sum_{e \in \tilde{\mathcal{E}}} p_e^{(n,k)} x_{u,e} - \lambda_u^{(n,k)} \cdot (\tilde{\mathbf{a}}_u^\top \mathbf{x}_u + x_{u,\varnothing} - 1) \\ & - \frac{\beta}{2}(\tilde{\mathbf{a}}_u^\top \mathbf{x}_u + x_{u,\varnothing} - 1)^2 - \frac{\beta}{2}\|\mathbf{y}_u^{(n,k)} - \mathbf{x}_u\|^2 \Bigg),\end{aligned} \tag{14.5}$$

where β is a positive scalar that represents the step-size in the ADMM algorithm (cf. Appendix L.3).

Remark 14.4.1. *The update in Equation* (14.5) *can be implemented through a "proxy bidding agent" in place of an actual AAM vehicle. The AAM vehicle operator can have the proxy agent at their local end, where they feed the AAM vehicle's valuation, and the proxy agent then participates on behalf of the AAM vehicle. This ensures that no on-board energy resources are used to run Algorithm 11. The proxy agent attempts to maximize the vehicle's budget-adjusted utility while penalizing deviations from the expected allocation, overspending*

artificial currency, and violating constraints. These constraints ensure that the bundle of edges selected by the AAM vehicle results in a feasible path on the time-extended graph.

Updates by SP: Using the demand from AAM vehicles, the SP updates the expected allocation and the excess supply through (14.6). The objective in (14.6) requires the SP to minimize three terms: (i) the difference between the expected allocation by the SP and the demand sent by AAM vehicles; (ii) the violation of resource constraints; and (iii) the minimization of unused capacity on any resource with a positive price.

$$\begin{aligned}&(\mathbf{y}^{(n+1,k)}, \mathbf{z}^{(n+1,k)})\\ &= \underset{\mathbf{y}\in\mathbb{R}^{U\times|\tilde{\mathcal{E}}|},\ \mathbf{z}\in\mathbb{R}_+^{|\tilde{\mathcal{E}}|}}{\arg\max} \left(-\frac{\beta}{2}\sum_{u\in U}\|\mathbf{y}_u - \mathbf{x}_u^{(n+1,k)}\|^2 - \frac{\beta}{2}\|\sum_{u\in U}\mathbf{y}_u + \mathbf{z} - \ell\|^2 - \sum_{e\in\tilde{\mathcal{E}}} p_e^{(n,k)} z_e \right).\end{aligned} \tag{14.6}$$

Next, the SP updates the price estimates using the updated values of the expected allocation and the excess supply through (14.7). Equation (14.7) reflects the idea that the SP should increase the price if the capacity constraint is violated and reduce it if there is available capacity.

$$\mathbf{p}^{(n+1,k)} = \mathbf{p}^{(n,k)} + \beta\left(\sum_{u\in U}\mathbf{y}_u^{(n+1,k)} + \mathbf{z}^{(n+1,k)} - \ell\right). \tag{14.7}$$

Finally, the SP updates the dual multiplier estimate λ_u for each AAM vehicle u using (14.8).

$$\lambda_u^{(n+1,k)} = \lambda_u^{(n,k)} + \beta\left(\tilde{\mathbf{a}}_u^\top \mathbf{x}_u^{(n+1,k)} + x_{u,\varnothing}^{(n+1,k)} - 1\right). \tag{14.8}$$

Outer loop. In the outer loop, the SP updates the budget adjustment after every N iteration of the inner loop, using the value of $\lambda^{(N,k)}$ to approximate the fixed-point iteration (line 12 of Algorithm 11). This step ensures that budget adjustments progressively converge toward equilibrium by iteratively refining the dual variables based on the current solution from the inner loop.

Termination criterion[9]: We terminate the algorithm once **all** of the following errors fall below their predefined threshold:

- Complementarity error (CE): Smaller values of CE ensure that resources with a positive price maintain a balance between demand and supply, while resources priced at zero satisfy capacity constraints. We define

$$\mathsf{CE}(\bar{\mathbf{x}}, \mathbf{p}) = \sqrt{\sum_{e\in\tilde{\mathcal{E}}} p_e^2 z_e^2}, \tag{14.9}$$

[9]Reaching a fixed point may be used as an alternative stopping criteria.

where $z_e = \sum_{u \in U} x_{u,e} - \ell_e$ is the *excess demand*. The definition of CE is motivated from the "complementarity condition" in general equilibrium theory in economics [288]. We define $\mathsf{tol}_{\mathsf{CE}}$ to be the threshold for this error.

- Individual constraint error (ICE): Smaller values of ICE ensure that (14.4d) constraint is satisfied. We define

$$\mathsf{ICE}(\bar{\mathbf{x}}) = \max_{u \in U} \|\tilde{\mathbf{a}}_u^\top \mathbf{x}_u + x_{u,\varnothing} - 1\|_\infty. \tag{14.10}$$

  We define $\mathsf{tol}_{\mathsf{ICE}}$ to be the threshold for this error.

- Expected allocation error (EAE): Smaller value of EAE ensure that (14.4b) is satisfied. We define

$$\mathsf{EAE}(\mathbf{x}, \mathbf{y}) = \max_{u \in U} \|\mathbf{y}_u - \mathbf{x}_u\|_\infty. \tag{14.11}$$

  We define $\mathsf{tol}_{\mathsf{EAE}}$ to be the threshold for this error.

We represent the output of Algorithm 11 by $(\bar{\mathbf{x}}^\dagger, \mathbf{p}^\dagger)$.

Remark 14.4.2. *Algorithm 11 is similar to online learning algorithms for computing market equilibrium [212, 246, 54]. In this process, AAM vehicles iteratively adjust their demand based on anonymized signals—such as prices, expected allocation, resource surplus, and dual multipliers—broadcast by the SP. Meanwhile, the SP dynamically updates these signals in response to the aggregate demand from AAM vehicles, without requiring knowledge of their private valuations.*

Step 2: Computing Integral Allocation

Using the output from Algorithm 11, the SP computes an integral allocation in a distributed manner using Algorithm 12. The SP sets the airspace price to $\mathbf{p}^\dagger$ (the output of Algorithm 11) and generates a *ranked list* of AAM vehicles based on $\bar{\mathbf{x}}^\dagger$ (cf. Sec. 14.4). This list is created by ranking the AAM vehicles in descending order according to the numerical value of the (fractional) allocation of their preferred routes (cf. Sec. 14.4).

Ranked List.

To generate a ranked list of AAM vehicles, we define $s^*(u)$ to be the most desired route of AAM vehicle u in R_u. Using $\bar{\mathbf{x}}^\dagger$ from Algorithm 11, the SP creates a ranking over agents based on decreasing values of $x^\dagger_{u,e^*(s^*(u))}$, where $e^*(s^*(u))$ denotes the departing edge on the route $s^*(u)$. We denote the ranked list[10] of AAM vehicles by $\mathbf{rank}$.

[10]Ties are broken arbitrarily.

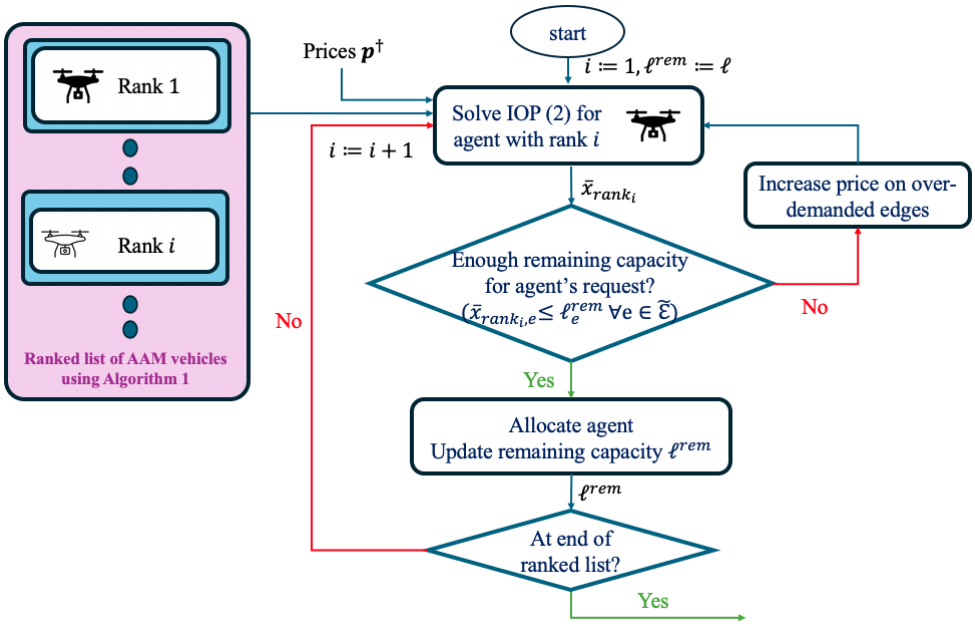


Figure 14.5: Flowchart of Algorithm 12, illustrating the ranking system and the removal of over-demanded edges.

Integral Allocation.

After generating the ranked list, the SP fixes the prices for all resources based on the output of Algorithm 11 (i.e., $\bar{\mathbf{p}}^\dagger$) and iterates over AAM vehicles according to **rank** (cf. Line 3 in Algorithm 12). Each AAM vehicle is allocated its most desired feasible route (cf. Lines 5–10 in Algorithm 12), subject to the current resource capacity [11]. If a capacity constraint is violated on any resource, the SP either removes that resource from the available pool for all remaining agents or increases its price to infinity for the remaining AAM vehicles (as described in Line 12 of Algorithm 12). See Fig. 14.5 for a schematic illustration.

Remark 14.4.3. *For Algorithms 11-12, the SP only requires information on the resource demands of AAM vehicles, their feasible time-trajectories, and their most preferred time-trajectory. Importantly, the SP does* ***not*** *need any details regarding the private valuations of individual AAM vehicles. Moreover, each AAM vehicle does not have access to the demands or private valuations of other vehicles, ensuring that privacy is maintained throughout the process.*

[11] The IOP in Algorithm 12 is constrained by the adjusted budget $w_u^\dagger = w_u + \omega_u$.

Algorithm 12 Integral Allocation Based on Ranked List

Input: $\mathbf{p}^\dagger, \mathbf{w}^\dagger$, **rank**
Initialize remaining capacity: $\ell^{\text{rem}} \leftarrow \ell$
for $i = 1$ **to** $|U|$ **do**
  $u \leftarrow \textbf{rank}_i$
  // Select AAM vehicle from the ranked list
  `AAM_vehicle_allocated` $\leftarrow$ False
  while `AAM_vehicle_allocated` = False **do**
    AAM vehicle u reports its integral allocation $\bar{\mathbf{x}}_u$ by solving (14.2)
    if $x_{u,e} \leqslant \ell_e^{\text{rem}}$ for every $e \in \tilde{\mathcal{E}}$ **then**
      Update remaining capacity: $\ell_e^{\text{rem}} \leftarrow \ell_e^{\text{rem}} - \bar{x}_{u,e},\ \forall e \in \tilde{\mathcal{E}}$
      `AAM_vehicle_allocated` $\leftarrow$ True
    else
      Define contested set $C \leftarrow \{e \in \tilde{\mathcal{E}} \mid x_{u,e} > \ell_e^{\text{rem}}\}$
      // Identify contested goods
      for each $e \in C$ **do**
        Update price: $p_e \leftarrow \infty$
        // Remove e from further consideration
      end for
    end if
  end while
end for
Return: $\bar{\mathbf{x}}$

14.5 Drone Delivery in Toulouse: A Case Study

In this section, we validate our proposed mechanism using a dataset of simulated package deliveries by Airbus, as shown in Fig. 14.7. Specifically, we use a synthetic dataset generated by Airbus to simulate a drone-based package delivery scenario in Toulouse, France [127, 94].

Dataset Specification: The data involves four warehouses located on the periphery of the city, which serve as hubs for UAV take-off and landing. Delivery requests are generated with spatial locations drawn uniformly across Toulouse and their temporal occurrences follow a Poisson process. Each request triggers a UAV to depart from the launchpad of a warehouse, deliver a package to the specified destination, and return to its origin. The UAV flight trajectory, including take-off, cruise, and landing phases, is generated by Airbus's high-fidelity trajectory simulator, ensuring that the simulated operations closely mimic real-world conditions. The data set includes data corresponding to 177 UAV flights, which spans roughly 6000 seconds (100 minutes). The average length of a UAV flight from a warehouse to the delivery location is approximately 300 seconds (5 minutes).

Airspace Specifications: To implement our auction approach, we partitioned Toulouse's airspace into 12 "cruising-altitude" regions, along with 4 warehouses, as shown in Fig. 14.7.

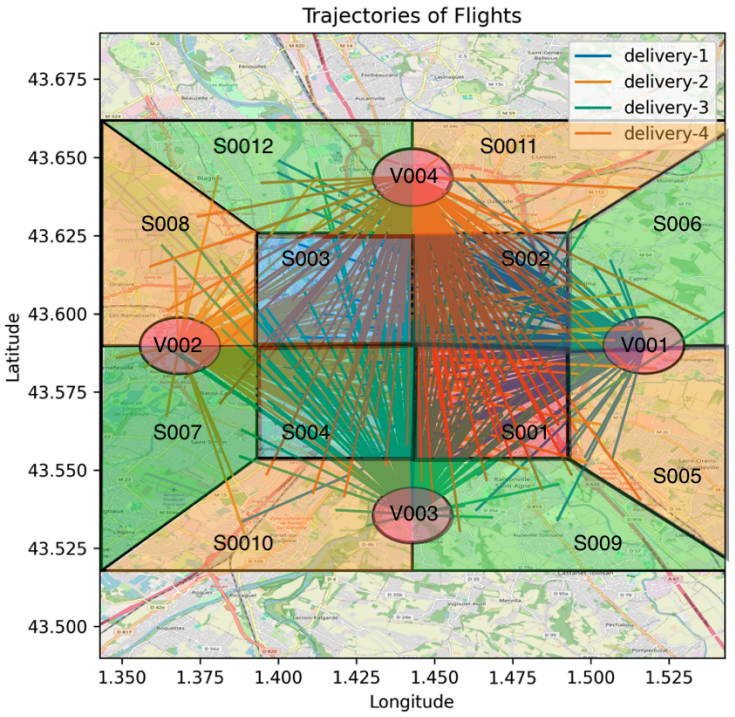


Figure 14.6: Division of Toulouse airspace into 12 cruising sectors (polygons) and 4 launch-pad sectors (circles). The lines show the trajectory of UAVs in the dataset. Labels indicate the sector (S#) or vertiport (V#).

Figure 14.7: A map is shown overlaid with 12 regions labeled S001 to S0012. There are also four circles labeled V001 to V004, which are in between regions S005 and S006, S007 and S008, S009 and S0010, and S0011 and S0012, respectively. Lines for the flight paths are shown emanating from each circle.

We construct a time-extended graph (cf. Definition 14.1.1) with a total of $T = 400$ time-steps, where each step of corresponds to $\tau = 15$ seconds. Based on the data we find that the minimum capacity needed to accommodate all requests from UAVs is 14 units in all 12 cruising regions of the airspace and 4 units for vertiport departure and arrival at the warehouse locations. Therefore, to make the problem more interesting, we set the capacity of each airspace region to 50% of the maximum number of vehicles it can accommodate at each time step unless otherwise specified.

UAV Specifications: Each UAV is either allocated a path on the time-extended graph or is rebased. For every UAV, the feasible set of paths on the time-extended graph includes its most desirable path, along with four alternative paths, each incurring a one-time-step delay. If a UAV is rebased, it converts its budget into outside options to be used again in the

next auction window. Every rebased UAV requests access to the airspace from the start of the next auction window. Any UAV may be rebased at most twice, after which it is dropped and no longer considered in future auctions.

The private value generated by each UAV on its most desirable path is a uniformly random number between 150 and 250 units. The utility decreases by a factor of 0.95 for each time-step delay in departure. If a UAV is rebased, its utility decreases by a factor of 0.5. The utility derived by any UAV from dropping out is 40 units.

Implementation Specifications: In each auction window, the SP allocates an additional budget of artificial currency to each participating UAV, randomly sampled between 150 and 250 units. This amount is then added to the UAV's existing budget accumulated from past auctions. For our numerical study, we set the nominal tolerance values in Algorithm 11 as follows: $\mathsf{tol}_{\mathsf{CE}} = 0.1\%$, $\mathsf{tol}_{\mathsf{ICE}} = 0.01\%$, and $\mathsf{tol}_{\mathsf{EAE}} = 0.1\%$ of their respective maximum attainable values. Additionally, in Algorithm 11, we set $\beta = 50$ and $N = 10$ for the qualitative analysis and $\beta = 50$ and $N = 30$ for the sensitivity analysis. We implemented our approach in Python and ran the simulations on a laptop equipped with a 12th-gen Intel Core i7-1200H CPU (14 cores, 20 threads) and 32GB DDR4 RAM. The operating system used was Ubuntu 22.04. Our code is publicly available at https://github.com/sastry-group/Mechanism-Design-for-AAM.

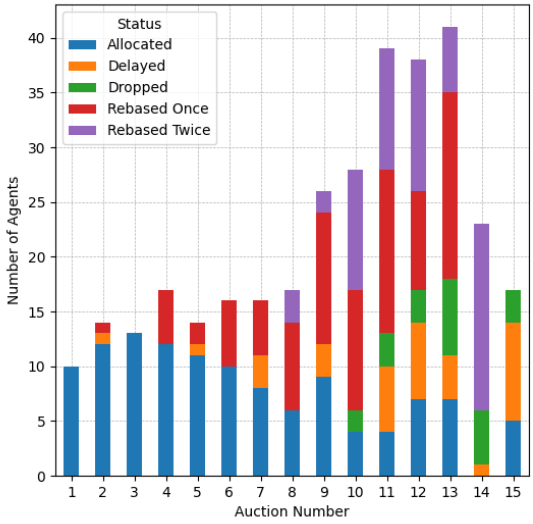
Qualitative Analysis

In this subsection, we present the outcome of our receding-horizon auction approach that we conducted for a total of 15 (i.e, $I = 15$) auctions. Before the start of every auction, the SP gathers the demand of AAM vehicles requesting access to the airspace. The overall budget of any UAV includes the new air credits they received and any unused air credits from a previous round. Additionally, we also compare our performance with two baselines.

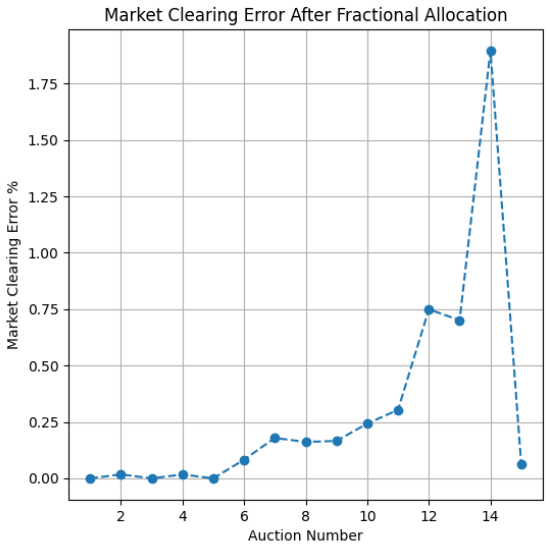
In Fig. 14.8a, we present the number of agents that were allocated, delayed, dropped, rebased-once, and rebased-twice across different auctions. We observe that the number of rebased and dropped agents increases in later rounds. This is because congestion builds up over time, as we operate in highly contested settings where the maximum capacity of every region is set to 50%.

In Fig. 14.8b, we present the market clearing error (MCE) after Algorithm 12, which captures the fraction of edges on the time-extended graph that have a positive price despite congestion being below capacity. This metric aligns with the third component of the competitive equilibrium definition in Definition 14.3.1, reflecting the economic principle that a good should not carry a positive price if it is underutilized. We find that this error is small—under 2%—highlighting that our approach (Algorithm 11 + 12) does not impose prices on uncontested goods.

In Table 14.5.1, we compare our approach to two baselines based on the ascending clock auction [109]. Both baselines implement simultaneous ascending clock auctions using β as the price increment. Agents are allowed to perform price discovery by bidding only on their most beneficial request, rather than on all goods across their preferred and delayed options.



(a) A bar plot shows the number of agents with status Allocated, Delayed, Dropped, Rebased Once, and Rebased Twice for each auction. The earlier auctions show mostly allocated agents with more Delayed, Dropped, and Rebased agents in the later auctions.



(b) A line graph shows the value of the market-clearing error for each auction. The error for all auctions except for the 14th is below 1 percent.

Figure 14.8: Properties of allocation finalized by our receding-horizon auction approach.

Table 14.5.1: Comparison to baseline auction approaches under different capacities.

Approach	Capacity	Num. Times Rebased	Num. Delayed UAVs	Avg. Delay (time steps)	Num. Rebased UAVs	Avg. Times Rebased	Num. Never Allocated
Budget-based	60%	20	12	1.83	17	1.18	**0**
	50%	**148**	**22**	**1.22**	**84**	1.76	43
Profit-based	60%	59	17	2.41	32	1.84	24
	50%	164	**22**	2.86	87	1.89	70
Ours	60%	**14**	**10**	**1.1**	**14**	**1**	**0**
	50%	153	35	1.51	91	**1.68**	**25**

Due to constraint (14.2b), we can assume that agents will always bid either on a request or on the outside option, and are therefore always active. The overall procedure is outlined in Algorithm 13. In the *Budget-based* comparison, agents solve their individual optimization problem (IOP) to determine their bids. In the *Profit-based* comparison, agents determine their bids based on their profit (value minus price). In each round, all agents submit bids,

and prices are increased on contested goods until no further goods are contested.

Algorithm 13 Ascending Clock Auction Comparisons

Initialize prices: $\mathbf{p} \leftarrow \mathbf{0}$
repeat
  $\hat{\mathbf{x}}_u \leftarrow$ AAM vehicle u's integral output from (14.2) *// Budget-based Bid*
  $\hat{\mathbf{x}}_u \leftarrow \arg\max_{\bar{\mathbf{x}}_u} \sum_{s \in R_u} v_{u,s} x_{u,e^*(s)} + v_{u,o} x_{u,o} - \mathbf{p}^T \mathbf{x}_u - p_o x_{u,o}$
    subject to: constraints (14.2b)–(14.2d), $\forall u \in U$ *// Profit-based Bid*
  $\mathcal{C} \leftarrow \left\{ e \in \tilde{\mathcal{E}} \mid \ell_e < \sum_{u \in U} \hat{x}_{u,e} \right\}$
  $p_e \leftarrow p_e + \beta \quad \forall e \in \mathcal{C}$ *// Increase price for overcapacitated goods*
until $\mathcal{C} = \emptyset$
Return: $\hat{\mathbf{x}}_{u \in U}$

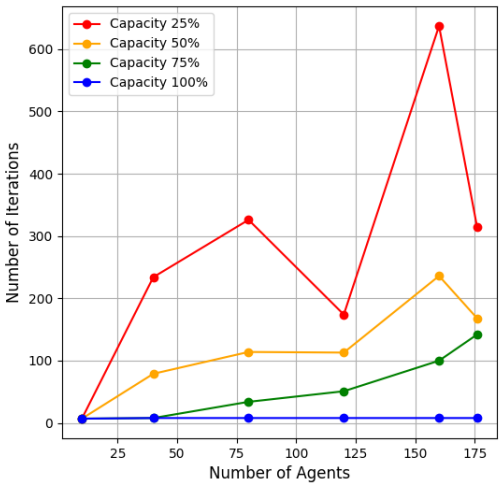
We compare these approaches using the Toulouse example at a 50% capacity level and also consider a slightly less constrained case (60% capacity). The set of goods remains consistent across all approaches. Our evaluation includes the following metrics: the number of agents that are never allocated (dropped agents), the number of delayed agents, the average delay duration for delayed agents, the total number of times agents are rebased (including both once- and twice-rebased cases), the number of agents rebased at least once, and the average number of times agents are rebased. Lower values are preferable across all metrics, and the lowest value for each metric is shown in **green and bolded**.

For the 60% capacity case, both our approach and the Budget-based comparison allocate all agents, with our approach additionally resulting in fewer delayed and rebased aircraft. In the more constrained 50% case, our approach significantly reduces the number of unallocated agents. While the Budget-based approach results in fewer delayed aircraft and a lower number of rebased agents, minimizing the number of unallocated agents is the more desirable outcome. The Profit-based approach performs worse in both cases across nearly every metric, further highlighting the benefits of using artificial currency. We attribute this to the high cumulative costs incurred when agents must bid on multiple goods in this setting. These results demonstrate that, for comparable values of β, our approach is better equipped to manage increasing congestion, as evidenced by the lower numbers of unallocated and rebased agents.

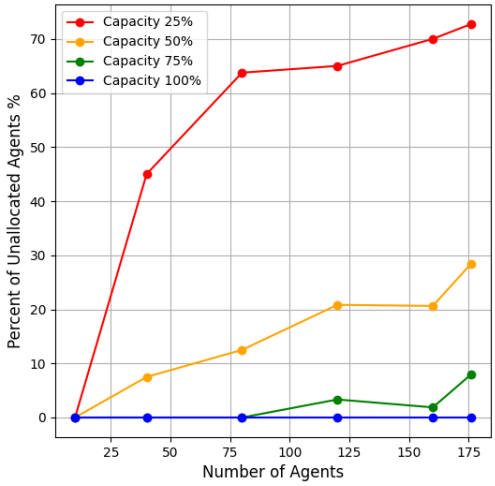
Sensitivity Analysis

In this subsection, we conduct numerical studies to evaluate the impact of key design parameters on the performance of our algorithmic approach.

In Fig. 14.9a, we present the variation in the number of iterations of Algorithm 11 with respect to different numbers of agents under various capacity constraints. This scenario can be interpreted as a setting with a single auction window in which all agents simultaneously request their desired goods on the time-extended graph. We observe that an increase in the



(a) Iterations vs. Number of UAVs

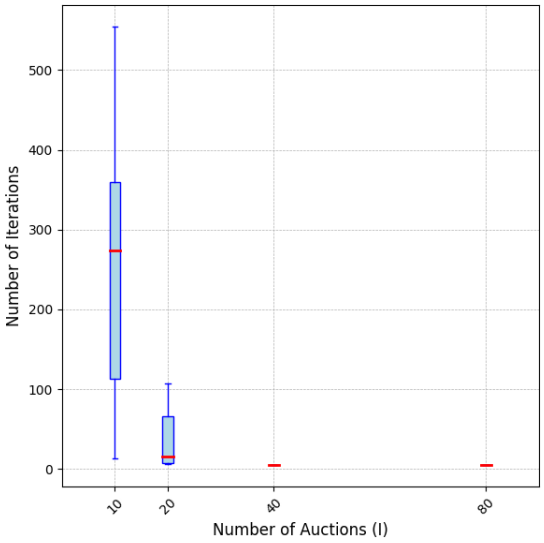


(b) Percent of Unallocated UAVs vs. Number of UAVs

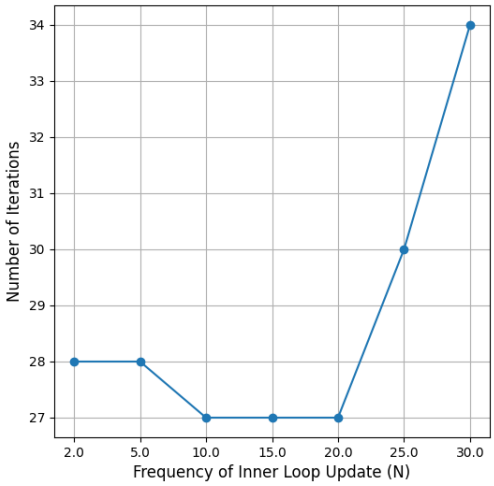
Figure 14.9: Variation of the number of iterations and percent of unallocated agents with respect to the number of agents as we vary capacity constraints on the resources.

number of agents leads to a higher number of contested goods, which requires more iterations of Algorithm 11 to compute a fractional competitive equilibrium. At 100% capacity, there are no contested resources, and the algorithm requires only 8 iterations regardless of the number of agents. As the capacity decreases, the number of iterations increases due to the higher number of contested goods. In Fig. 14.9b, we analyze the variation in the percentage of unallocated agents with respect to different numbers of agents under various capacity constraints. We observe that when the resource capacity is 100%, all UAVs receive their desired paths. However, as the capacity decreases, a greater number of agents remain unallocated. Furthermore, for any given capacity level, the percentage of unallocated agents trends upwards as more agents participate in the auction.

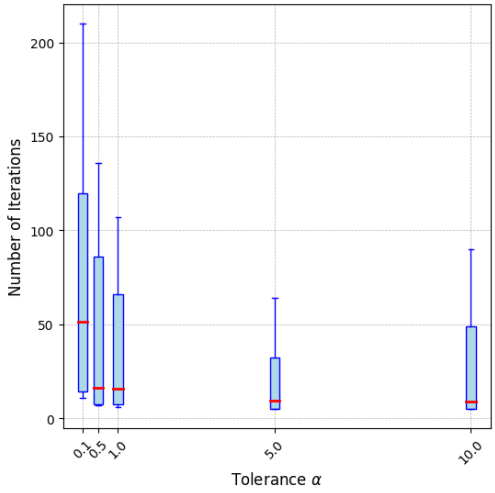
In Fig. 14.10, we study the variation in the number of iterations of Algorithm 11 with respect to different market parameters. First, in Fig. 14.10a, we examine how the number of iterations of Algorithm 11 changes as we vary the number of auctions in our receding horizon approach. As the number of auctions increases, the number of UAVs participating in each auction decreases, resulting in fewer contested goods. Consequently, Algorithm 11 converges more quickly. Next, in Fig. 14.10b, we evaluate the impact of the number of inner loop updates (parameter N in Algorithm 11). Recall that the goal of Algorithm 11 is to emulate fixed-point iteration (FP). Toward this goal, the role of inner loop updates in Algorithm 11 is to estimate $\lambda^{\dagger}(\omega^{(k)})$ for every update of $\omega^{(k)}$ in the outer loop. The results show that when N is lower, the number of iterations of Algorithm 11 is higher, as the



(a) Iterations vs. Number of Auctions (I)



(b) Iterations vs. Inner Loop Update Parameter in Algorithm 11 (N)



(c) Iterations vs. Tolerance in Algorithm 11

Figure 14.10: Variation in number of iterations of Algorithm 11 with different algorithmic parameters.

inner loop cannot accurately estimate $\lambda^{\dagger}(\omega^{(k)})$. Consequently, increasing N decreases the number of iterations up to a point. However, if we continue to increase N past this point, the additional steps in the inner loop do not help the convergence of fixed-point iteration (FP) and instead cause the iterations of Algorithm 11 to start increasing again. Finally, in Fig. 14.10c, we analyze how the number of iterations varies with different convergence tolerances. For clarity, we introduce the variable α, which is a multiplier for the nominal value of tolerances (i.e., $\mathsf{tol}_{\mathsf{CE}}, \mathsf{tol}_{\mathsf{ICE}}, \mathsf{tol}_{\mathsf{EAE}}$) described in our setup above, and we study the variation in the number of iterations with respect to α. As expected, stricter tolerances (i.e., lower α) generally increase the number of iterations, since the algorithm must satisfy more stringent stopping conditions.

14.6 Limitations

We highlight several limitations of the current modeling framework, each of which presents a promising direction for future research:

(i) We assume that the AAM vehicles are not malicious and follow Algorithms 11 and 12 as intended. In practice, this can be implemented before takeoff via a 'proxy agent', which is an autonomous entity that bids on behalf of each vehicle. The use of proxy agents is a well-established approach for implementing iterative auctions [109]. Designing effective auditing mechanisms to ensure agent compliance remains an important open problem.

(ii) We assume that communication between the service provider and each UAV is delay-free and noiseless. Once a vehicle has taken off, its route is fixed, and no further in-flight communication is required.

(iii) The current framework assumes that a single service provider manages the entire airspace. Extending the model to support multiple, potentially competing, service providers is an important direction for future exploration.

(iv) Each UAV is assigned a trajectory before takeoff and does not deviate from it during flight.

(v) We assume that sufficient infrastructure and operational measures are in place to manage emergencies.

(vi) Budgets are randomly assigned to AAM vehicles by the SP. Developing improved budget allocation mechanisms that promote social welfare or fairness is a compelling direction for future research.

14.7 Concluding Remarks

In this work, we introduce a novel mechanism that enables service providers to allocate on-demand requests from AAM vehicles—each with heterogeneous private valuations—to a capacity-constrained airspace. This is the first work in the AAM literature to allocate constrained airspace resources to dynamically arriving AAM vehicles without requiring knowledge of their private valuations. Central to our approach is an artificial currency-based auction mechanism implemented in a receding-horizon manner. In every auction, we use a distributed iterative algorithm that accounts for individual agent preferences while ensuring system efficiency and safety. We evaluate the effectiveness of our approach using an urban air delivery dataset.

Part V

Final Remarks

Chapter 15

Summary & Future Directions

This dissertation advances the rapidly evolving field of autonomous AI technologies, which are increasingly being integrated into societal systems where multiple autonomous agents and humans interact strategically under uncertainty, dynamic environments, resource constraints, and limited information. Addressing these challenges requires not only the development of novel paradigms for efficient, real-time learning in multi-agent environments but also developing mechanisms that ensure socially efficient, equitable, and safe outcomes.

To this end, this dissertation advocates for a new systems-theoretic paradigm that synthesizes concepts from machine learning and AI for algorithmic decision-making, optimization and statistics for performance guarantees, and game theory and mechanism design for modeling and shaping strategic interactions.

Structured in four parts, the dissertation develops new theory and algorithms for the design, analysis, and regulation of dynamic multi-agent autonomous systems for societal transformation. It discusses several applications spanning from mobility to autonomous high-performance robotics. Looking forward, several promising research directions emerge by leveraging ideas from different parts. These directions present not only technical challenges but also opportunities to influence how autonomous systems are engineered, deployed, and governed in real-world societal contexts.

Future Directions

Below, I highlight a subset of open research directions that emerge out of the ideas discussed in this dissertation:

Dynamic Coalition Formation and Information Sharing: Building on the theoretical and algorithmic frameworks introduced in **Part I**, which explores interactions among multiple autonomous agents in dynamic environments, future work could focus on frameworks for dynamic collaborations among agents to improve individual and collective outcomes. Recent studies, such as those examining collaborations between strategic EV charging stations [220], demonstrate that poorly designed collaborations can lead to unintended consequences

for both individuals and society. Thus, it is crucial to develop a principled theory of dynamic coalition formation and information sharing, with an emphasis on understanding when and how agents should collaborate under asymmetric information. This is particularly important in applications such as vehicle platooning, distributed energy markets, and robotic swarms for humanitarian assistance and disaster response, where efficient and equitable team formation is vital.

Integration of Dynamic and Resource-Constrained Multi-Agent Learning: Recall that **Parts I** and **II** focus on multi-agent learning in dynamic and resource-constrained (or congested) environments, respectively. An important next step is to combine these insights to design and analyze decentralized algorithms that operate in environments characterized by both dynamic transitions and resource constraints or congestion effects. Such interactions are central to the development of navigation algorithms for transportation networks and other critical infrastructure.

High-Speed, High-Stakes Multi-Agent Planning and Control: Inspired by the results in Chapter 3, another promising direction is the extension of multi-agent planning and control algorithms to high-speed, high-stakes robotic applications, such as multi-drone racing, pursuit-evasion under limited information, and space robotics. These domains require algorithms that can operate under strict time constraints and adversarial interactions, while coordinating multiple physical agents in uncertain environments. Progress in this area will help bridge the gap between high-level AI reasoning and low-level robotic control, unlocking the potential for safe and intelligent robotic autonomy in complex, real-world settings.

Characterizing Convergence and Designing Adaptive Incentive Mechanisms: In **Part I** (Chapters 4-5), this dissertation characterizes the set of convergent policies when agents employ decentralized actor-critic algorithms. Building on this, future work should investigate convergence properties when agents use heterogeneous algorithms (e.g., Q-learning, policy gradients). Furthermore, as discussed in **Part III**, interactions among learning agents can sometimes yield socially inefficient, inequitable and unsafe outcomes. Another key challenge is to develop adaptive incentive mechanisms that align agent behavior with societal objectives in dynamic environments. These mechanisms must be privacy-preserving, robust to un-modeled system dynamics, and provide guarantees not only on long-run equilibria but also on transient dynamics, especially in systems with non-myopic or non-stationary agents. Applications include real-time pricing in traffic and logistics networks and coordination of autonomous service providers in urban environments.

Market Mechanisms for Advanced Air Mobility: While **Part IV** focuses on the design of market mechanisms for advanced air mobility that ensure strategic (network-level) deconfliction between UAV fleets, future research must also address tactical (low-level) deconfliction between individual robots. This involves enabling robots to dynamically negotiate

trajectories, even as conditions change mid-flight. Additionally, future work should develop airspace allocation mechanisms that are both efficient and fair across competing UAV service providers, who may privately manage different airspaces. This direction aligns with NASA's *Sky for All* initiative and underscores the need for scalable, socially aligned market mechanisms in urban air traffic management.

Concluding Words

This dissertation lays the methodological foundation for engineering autonomous systems that reconcile individual decision-making with broader societal objectives. By unifying insights from artificial intelligence, systems theory, and economics, the proposed frameworks address key challenges in multi-agent learning, coordination, and governance. The research directions outlined—from high-speed robotic control to adaptive market design—offer a road map for building autonomous technologies that are both effective and responsible within complex socio-technical systems.

While substantial theoretical and technical challenges remain, the opportunity to shape a future in which autonomous technologies enhance human life is immense. This endeavor extends beyond purely academic inquiry—it is a societal imperative to design and deploy autonomous systems that can be responsibly integrated into diverse real-world multi-agent domains.

Bibliography

[1] Daron Acemoglu and Martin Kaae Jensen. "Aggregate comparative statics". In: *Games and Economic Behavior* 81 (2013), pp. 27–49.

[2] Daron Acemoglu, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. *Networks, shocks, and systemic risk*. Tech. rep. National Bureau of Economic Research, 2015.

[3] Jeffrey L Adler and Mecit Cetin. "A direct redistribution model of congestion pricing". In: *Transportation research part B: methodological* 35.5 (2001), pp. 447–460.

[4] Federal Aviation Administration. *Unmanned Aircraft System (UAS) Traffic Management (UTM) Concept of Operations (ConOps) Version 2.0*. https://www.faa.gov/sites/faa.gov/files/2022-08/UTM_ConOps_v2.pdf. 2022.

[5] Federal Aviation Administration. *Urban Air Mobility Concept of Operations 2.0*. Technical report. Federal Aviation Administration, 2023. URL: https://www.faa.gov/air-taxis/uam_blueprint.

[6] Shipra Agrawal and Navin Goyal. "Analysis of thompson sampling for the multi-armed bandit problem". In: *Conference on learning theory*. JMLR Workshop and Conference Proceedings. 2012, pp. 39–1.

[7] Takashi Akamatsu. "Decomposition of Path Choice Entropy in General Transport Networks". In: *Transportation Science* 31.4 (Nov. 1997), pp. 349–362.

[8] Ahmet Alacaoglu, Luca Viano, Niao He, and Volkan Cevher. "A natural actor-critic framework for zero-sum Markov games". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 307–366.

[9] José Alcalde. "Exchange-proofness or divorce-proofness? Stability in one-sided matching markets". In: *Review of Economic Design* 1 (Feb. 1994), pp. 275–287. DOI: 10.1007/BF02716626.

[10] Robin Allesiardo, Raphaël Féraud, and Odalric-Ambrym Maillard. "The non - stationary stochastic multi-armed bandit problem". In: *International Journal of Data Science and Analytics* 3.4 (2017), pp. 267–283.

[11] Tansu Alpcan and Lacra Pavel. "Nash equilibrium design and optimization". In: *2009 international conference on game theory for networks*. IEEE. 2009, pp. 164–170.

[12] Tansu Alpcan, Lacra Pavel, and Nem Stefanovic. "A control theoretic approach to noncooperative game design". In: *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*. IEEE. 2009, pp. 8575–8580.

[13] Enrico Angelelli, Idil Arsik, Valentina Morandi, Martin Savelsbergh, and Maria Grazia Speranza. "Proactive route guidance to avoid congestion". In: *Transportation Research Part B: Methodological* 94 (2016), pp. 1–21.

[14] Enrico Angelelli, Valentina Morandi, Martin Savelsbergh, and Maria Grazia Speranza. "System optimal routing of traffic flows with user constraints using linear programming". In: *European journal of operational research* 293.3 (2021), pp. 863–879.

[15] Guy Aridor, Yishay Mansour, Aleksandrs Slivkins, and Zhiwei Steven Wu. "Competing bandits: The perils of exploration under competition". In: *arXiv preprint arXiv:2007.10144* (2020).

[16] Richard Arnott and Kenneth Small. "The economics of traffic congestion". In: *American scientist* 82.5 (1994), pp. 446–455.

[17] Gürdal Arslan and Serdar Yüksel. "Decentralized Q-learning for stochastic teams and games". In: *IEEE Transactions on Automatic Control* 62.4 (2016), pp. 1545–1558.

[18] Anil Aswani, Zuo-Jun Shen, and Auyon Siddiq. "Inverse optimization with noisy data". In: *Operations Research* 66.3 (2018), pp. 870–892.

[19] I.C. Athira, C.P. Muneera, K. Krishnamurthy, and M.V.L.R. Anjaneyulu. "Estimation of Value of Travel Time for Work Trips". In: *Transportation Research Procedia* 17 (2016). International Conference on Transportation Planning and Implementation Methodologies for Developing Countries (12th TPMDC) Selected Proceedings, IIT Bombay, Mumbai, India, 10-12 December 2014, pp. 116–123. ISSN: 2352-1465. DOI: https://doi.org/10.1016/j.trpro.2016.11.067. URL: https://www.sciencedirect.com/science/article/pii/S2352146516306810.

[20] Peter Auer. "Using confidence bounds for exploitation-exploration trade-offs". In: *Journal of Machine Learning Research* 3.Nov (2002), pp. 397–422.

[21] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. "Finite-time analysis of the multiarmed bandit problem". In: *Machine learning* 47.2 (2002), pp. 235–256.

[22] Peter Auer, Pratik Gajane, and Ronald Ortner. "Adaptively tracking the best bandit arm with an unknown number of distribution changes". In: (2019), pp. 138–158.

[23] Lawrence M. Ausubel and Peter Cramton. "Auctioning Many Divisible Goods". In: *Journal of the European Economic Association* 2.2-3 (May 2004), pp. 480–493. ISSN: 1542-4766. DOI: 10.1162/154247604323068168. eprint: https://academic.oup.com/jeea/article-pdf/2/2-3/480/10317693/jeea0480.pdf. URL: https://doi.org/10.1162/154247604323068168.

[24] J. A. Bagnell. "Robust supervised learning." In: *AAAI*. Vol. 2. 2005, pp. 714–719.

[25] Yu Bai, Chi Jin, Huan Wang, and Caiming Xiong. "Sample-efficient learning of stackelberg equilibria in general-sum games". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 25799–25811.

[26] Jean-Bernard Baillon and Roberto Cominetti. "Markovian Traffic Equilibrium". In: *Mathematical Programming* (Feb. 2008).

[27] Bharathan Balaji, Sunil Mallya, Sahika Genc, Saurabh Gupta, Leo Dirac, Vineet Khare, Gourav Roy, Tao Sun, Yunzhe Tao, Brian Townsend, et al. "Deepracer: Autonomous racing platform for experimentation with sim2real reinforcement learning". In: *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2020, pp. 2746–2754.

[28] Hamsa Balakrishnan and Bala Chandran. "A distributed framework for traffic flow management in the presence of unmanned aircraft". In: ATM Seminar. 2017.

[29] Michael O. Ball, Frank Berardino, and Mark Hansen. "The use of auctions for allocating airport access rights". In: *Transportation Research Part A: Policy and Practice* 114 (2018), pp. 186–202. ISSN: 0965-8564. DOI: https://doi.org/10.1016/j.tra.2017.09.026. URL: https://www.sciencedirect.com/science/article/pii/S0965856416303287.

[30] Michael O. Ball, Alexander S. Estes, Mark Hansen, and Yulin Liu. "Quantity - Contingent Auctions and Allocation of Airport Slots". In: *Transportation Science* 54.4 (2020), pp. 858–881. DOI: 10.1287/trsc.2020.0995. URL: https://doi.org/10.1287/trsc.2020.0995.

[31] Ian C Barnes, Karen Trapenberg Frick, Elizabeth Deakin, and Alexander Skabardonis. "Impact of peak and off-peak tolls on traffic in san francisco–oakland bay bridge corridor in california". In: *Transportation research record* 2297.1 (2012), pp. 73–79.

[32] Jorge Barrera and Alfredo Garcia. "Dynamic incentives for congestion control". In: *IEEE Transactions on Automatic Control* 60.2 (2014), pp. 299–310.

[33] Tamer Başar. "Affine incentive schemes for stochastic systems with dynamic information". In: *SIAM Journal on Control and Optimization* 22.2 (1984), pp. 199–210.

[34] Leonardo J. Basso and Anming Zhang. "Pricing vs. slot policies when airport profits matter". In: *Transportation Research Part B: Methodological* 44.3 (2010). Economic Analysis of Airport Congestion, pp. 381–391. ISSN: 0191-2615. DOI: https://doi.org/10.1016/j.trb.2009.09.005. URL: https://www.sciencedirect.com/science/article/pii/S0191261509001179.

[35] Soumya Basu, Karthik Abinav Sankararaman, and Abishek Sankararaman. "Beyond log-squared Regret for Decentralized Bandits in Matching Markets". In: *arXiv preprint arXiv:2103.07501* (2021).

[36] Lucas Baudin and Rida Laraki. "Fictitious play and best-response dynamics in identical interest and zero-sum stochastic games". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 1664–1690.

[37] Aleksandar Bauranov and Jasenka Rakas. "Designing airspace for urban air mobility: A review of concepts and approaches". In: *Progress in Aerospace Sciences* 125 (2021), p. 100726. ISSN: 0376-0421. DOI: https://doi.org/10.1016/j.paerosci.2021.100726. URL: https://www.sciencedirect.com/science/article/pii/S0376042121000312.

[38] Ana LC Bazzan. "Opportunities for multiagent systems and multiagent reinforcement learning in traffic control". In: *Autonomous Agents and Multi-Agent Systems* 18.3 (2009), pp. 342–375.

[39] Martin Beckmann, Charles B McGuire, and Christopher B Winsten. *Studies in the Economics of Transportation*. Tech. rep. 1956.

[40] M. E. Ben-Akiva. *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge: MIT Press, 1985.

[41] A. Ben-Tal, D. D. Hertog, A. D. Waegenaere, B. Melenberg, and G. Rennen. "Robust solutions of optimization problems affected by uncertain probabilities". In: *Manag. Sci.* 59 (2013), pp. 341–357.

[42] Michel Benaı¨m. "Dynamics of stochastic approximation algorithms". In: *Seminaire de Probabilites XXXIII*. Springer, 1999, pp. 1–68.

[43] Michel Benaı¨m, Josef Hofbauer, and Sylvain Sorin. "Stochastic approximations and differential inclusions". In: *SIAM Journal on Control and Optimization* 44.1 (2005), pp. 328–348.

[44] David Bernstein. "Congestion pricing with tolls and subsidies". In: *Pacific Rim Trans Tech Conference—Volume II: International Ties, Management Systems, Propulsion Technology, Strategic Highway Research Program*. ASCE. 1993, pp. 145–151.

[45] Josh Bertram and Peng Wei. "An Efficient Algorithm for Self-Organized Terminal Arrival in Urban Air Mobility". In: *AIAA Scitech 2020 Forum*. DOI: 10.2514/6.2020-0660. eprint: https://arc.aiaa.org/doi/pdf/10.2514/6.2020-0660. URL: https://arc.aiaa.org/doi/abs/10.2514/6.2020-0660.

[46] Dimitris Bertsimas, David Brown, and Constantine Caramanis. "Theory and applications of robust optimization". In: *SIAM Review* 53 (Oct. 2010). DOI: 10.1137/080734510.

[47] Dimitris Bertsimas, Vishal Gupta, and Ioannis Ch Paschalidis. "Data-driven estimation in equilibrium using inverse optimization". In: *Mathematical Programming* 153 (2015), pp. 595–633.

[48] Dimitris Bertsimas, Guglielmo Lulli, and Amedeo Odoni. "An Integer Optimization Approach to Large-Scale Air Traffic Flow Management". In: *Operations Research* 59.1 (2011), pp. 211–227. DOI: 10.1287/opre.1100.0899. URL: https://doi.org/10.1287/opre.1100.0899.

[49] Dimitris Bertsimas and Sarah Stock Patterson. "The Air Traffic Flow Management Problem with Enroute Capacities". In: *Operations Research* 46.3 (1998), pp. 406–422. DOI: 10.1287/opre.46.3.406. eprint: https://doi.org/10.1287/opre.46.3.406. URL: https://doi.org/10.1287/opre.46.3.406.

[50] Dimitris Bertsimas and Sarah Stock Patterson. "The Traffic Flow Management Rerouting Problem in Air Traffic Control: A Dynamic Network Flow Approach". In: *Transportation Science* 34.3 (2000), pp. 239–255. DOI: 10.1287/trsc.34.3.239.12300. eprint: https://doi.org/10.1287/trsc.34.3.239.12300. URL: https://doi.org/10.1287/trsc.34.3.239.12300.

[51] Omar Besbes, Yonatan Gur, and Assaf Zeevi. "Optimal exploration-exploitation in multi-armed-bandit problems with non - stationary rewards". In: (2019), pp. 319–337.

[52] Johannes Betz, Hongrui Zheng, Alexander Liniger, Ugo Rosolia, Phillip Karle, Madhur Behl, Venkat Krovi, and Rahul Mangharam. "Autonomous vehicles on the edge: A survey on autonomous vehicle racing". In: *IEEE Open Journal of Intelligent Transportation Systems* 3 (2022), pp. 458–488.

[53] Martin Bichler, Peter Gritzmann, Paul Karaenke, and Michael Ritter. "On Airport Time Slot Auctions: A Market Design Complying with the IATA Scheduling Guidelines". In: *Transportation Science* 57.1 (2023), pp. 27–51.

[54] Ilai Bistritz and Nicholas Bambos. "Online learning for load balancing of unknown monotone resource allocation games". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 968–979.

[55] Avrim Blum, Nika Haghtalab, and Ariel D Procaccia. "Learning optimal commitment to overcome insecurity". In: *Advances in Neural Information Processing Systems* 27 (2014).

[56] Vincenzo Bonifaci, Mahyar Salek, and Guido Schäfer. "Efficiency of restricted tolls in non-atomic network routing games". In: *Algorithmic Game Theory: 4th International Symposium, SAGT 2011, Amalfi, Italy, October 17-19, 2011. Proceedings 4*. Springer. 2011, pp. 302–313.

[57] Vivek Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

[58] Vivek S Borkar. "Reinforcement learning in Markovian evolutionary games". In: *Advances in Complex Systems* 5.01 (2002), pp. 55–72.

[59] Vivek S Borkar. "Stochastic approximation with two time scales". In: *Systems & Control Letters* 29.5 (1997), pp. 291–294.

[60] Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*. Vol. 48. Springer, 2009.

[61] Vivek S Borkar and Sarath Pattathil. "Concentration bounds for two time scale stochastic approximation". In: *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 2018, pp. 504–511.

[62] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[63] Stephen D. Boyles, Kara M. Kockelman, and S. Travis Waller. "Congestion pricing under operational, supply-side uncertainty". In: *Transportation Research Part C: Emerging Technologies* 18.4 (2010), pp. 519–535. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2009.09.006. URL: https://www.sciencedirect.com/science/article/pii/S0968090X09001399.

[64] Yann Bramoullé, Andrea Galeotti, and Brian W Rogers. *The Oxford handbook of the economics of networks*. Oxford University Press, 2016.

[65] Yann Bramoullé and Rachel Kranton. "Public goods in networks". In: *Journal of Economic theory* 135.1 (2007), pp. 478–494.

[66] Mario Bravo, David Leslie, and Panayotis Mertikopoulos. "Bandit learning in concave N-person games". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Montréal, Canada, 2018, pp. 5666–5676.

[67] Luce Brotcorne, Martine Labbé, Patrice Marcotte, and Gilles Savard. "A bilevel model for toll optimization on a multicommodity transportation network". In: *Transportation science* 35.4 (2001), pp. 345–358.

[68] Noam Brown and Tuomas Sandholm. "Superhuman AI for heads-up no-limit poker: Libratus beats top professionals". In: *Science* 359.6374 (2018), pp. 418–424.

[69] Philip N Brown and Jason R Marden. "A study on price-discrimination for robust social coordination". In: *2016 American Control Conference (ACC)*. IEEE. 2016, pp. 1699–1704.

[70] Lukas Brunke. "Learning Model Predictive Control for Competitive Autonomous Racing". In: *ArXiv* abs/2005.00826 (2020). URL: https://api.semanticscholar.org/CorpusID:218487458.

[71] Sébastien Bubeck, Thomas Budzinski, and Mark Sellke. "Cooperative and Stochastic Multi-Player Multi-Armed Bandit: Optimal Regret With Neither Communication Nor Collisions". In: *CoRR* abs/2011.03896 (2020).

[72] Sébastien Bubeck, Yuanzhi Li, Haipeng Luo, and Chen-Yu Wei. "Improved path-length regret bounds for bandits". In: *Conference On Learning Theory*. PMLR. 2019, pp. 508–528.

[73] Swapna Buccapatnam, Jian Tan, and Li Zhang. "Information sharing in distributed stochastic bandits". In: *2015 IEEE Conference on Computer Communications (INFOCOM)*. IEEE. 2015, pp. 2605–2613.

[74] Eric Budish. "The Combinatorial Assignment Problem: Approximate Competitive Equilibrium from Equal Incomes". In: *Journal of Political Economy* 119.6 (2011), pp. 1061–1103. ISSN: 00223808, 1537534X. URL: http://www.jstor.org/stable/10.1086/664613 (visited on 10/07/2024).

[75] Eric Budish, Ruiquan Gao, Abraham Othman, Aviad Rubinstein, and Qianfan Zhang. "Practical algorithms and experimentally validated incentives for equilibrium-based fair division (A-CEEI)". In: *Proceedings of the 24th ACM Conference on Economics and Computation*. EC '23. London, United Kingdom: Association for Computing Machinery, 2023, pp. 337–368. ISBN: 9798400701047. DOI: 10.1145/3580507.3597809. URL: https://doi.org/10.1145/3580507.3597809.

[76] Alexander Buyval, Aidar Gabdulin, Ruslan Mustafin, and Ilya Shimchik. "Deriving overtaking strategy from nonlinear model predictive control for a race car". In: *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE. 2017, pp. 2623–2628.

[77] Dan Calderone and S Shankar Sastry. "Markov Decision Process Routing Games". In: *Proceedings of the 8th International Conference on Cyber-Physical Systems*. 2017, pp. 273–279.

[78] Dan Calderone and Shankar Sastry. "Infinite-horizon Average-cost Markov Decision Process Routing Games". In: *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2017, pp. 1–6.

[79] Joaquin Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009. ISBN: 0262170051.

[80] Ozan Candogan, Asuman Ozdaglar, and Pablo A Parrilo. "Dynamics in near-potential games". In: *Games and Economic Behavior* 82 (2013), pp. 66–90.

[81] Ozan Candogan, Asuman Ozdaglar, and Pablo A Parrilo. "Near-potential games: Geometry and dynamics". In: *ACM Transactions on Economics and Computation (TEAC)* 1.2 (2013), pp. 1–32.

[82] Alan Carlin and R. E. Park. "Marginal Cost Pricing of Airport Runway Capacity". In: *The American Economic Review* 60.3 (1970), pp. 310–319. ISSN: 00028282. URL: http://www.jstor.org/stable/1817981 (visited on 01/03/2024).

[83] Álvaro Cartea, Patrick Chang, José Penalva, and Harrison Waldon. "Algorithms can Learn to Collude: A Folk Theorem from Learning with Bounded Rationality". In: *Available at SSRN 4293831* (2022).

[84] Sarah H Cen and Devavrat Shah. "Regret, stability & fairness in matching markets with bandit learners". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 8938–8968.

[85] Sarah H Cen and Devavrat Shah. "Regret, stability, and fairness in matching markets with bandit learners". In: *arXiv preprint arXiv:2102.06246* (2021).

[86] Shicong Cen, Yuting Wei, and Yuejie Chi. "Fast policy extragradient methods for competitive games with entropy regularization". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 27952–27964.

[87] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[88] Mithun Chakraborty, Kai Yee Phoebe Chua, Sanmay Das, and Brendan Juba. "Coordinated Versus Decentralized Exploration In Multi-Agent Multi-Armed Bandits." In: *IJCAI*. 2017, pp. 164–170.

[89] Siddharth Chandak, Ilai Bistritz, and Nicholas Bambos. "Learning to Control Unknown Strongly Monotone Games". In: *arXiv preprint arXiv:2407.00575* (2024).

[90] Lesi Chen, Jing Xu, and Jingzhao Zhang. "On Bilevel Optimization without Lower-level Strong Convexity". In: *arXiv preprint arXiv:2301.00712* (2023).

[91] P. Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. "ZOO: Zeroth-Order Optimization-based Black-box Attacks to Deep Neural Networks Without Training Substitute Models". In: *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security* (2017).

[92] Yilan Chen, Daniel E Ochoa, Jason R Marden, and Jorge I Poveda. "High-Order Decentralized Pricing Dynamics for Congestion Games: Harnessing Coordination to Achieve Acceleration". In: *2023 American Control Conference (ACC)*. IEEE. 2023, pp. 1086–1091.

[93] Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. "Learning to optimize under non-stationarity". In: (2019), pp. 1079–1087.

[94] Christopher Chin, Karthik Gopalakrishnan, Hamsa Balakrishnan, Maxim Egorov, and Antony Evans. "Protocol-Based Congestion Management for Advanced Air Mobility". In: *Journal of Air Transportation* 31.1 (2023), pp. 35–44.

[95] Christopher Chin, Karthik Gopalakrishnan, Maxim Egorov, Antony Evans, and Hamsa Balakrishnan. "Efficiency and Fairness in Unmanned Air Traffic Flow Management". In: *IEEE Transactions on Intelligent Transportation Systems* 22.9 (2021), pp. 5939–5951. DOI: 10.1109/TITS.2020.3048356.

[96] Christopher Chin, Victor Qin, Karthik Gopalakrishnan, and Hamsa Balakrishnan. "Traffic management protocols for advanced air mobility". In: *Frontiers in Aerospace Engineering* 2 (2023), p. 1176969.

[97] Chih-Yuan Chiu, Chinmay Maheshwari, Pan-Yang Su, and Shankar Sastry. "Dynamic Tolling in Arc-based Traffic Assignment Models". In: *2023 59th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 2023, pp. 1–8.

[98] Simon Clark. "The Uniqueness of Stable Matchings". In: *Contributions to Theoretical Economics* 6 (Feb. 2006), pp. 1283–1283. DOI: 10.2202/1534-5971.1283.

[99] Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

[100] Simon Le Cleac'h, Mac Schwager, and Zachary Manchester. "Algames: A fast solver for constrained dynamic games". In: *arXiv preprint arXiv:1910.09713* (2019).

[101] Adam P Cohen, Susan A Shaheen, and Emily M Farrar. "Urban air mobility: History, ecosystem, market potential, and challenges". In: *IEEE Transactions on Intelligent Transportation Systems* 22.9 (2021), pp. 6074–6087.

[102] Riccardo Colini-Baldeschi, Roberto Cominetti, Panayotis Mertikopoulos, and Marco Scarsini. "When is selfish routing bad? The price of anarchy in light and heavy traffic". In: *Operations Research* 68.2 (2020), pp. 411–434.

[103] Roberto Cominetti, Valerio Dose, and Marco Scarsini. "The price of anarchy in routing games as a function of the demand". In: *Mathematical Programming* (2021), pp. 1–28.

[104] Roberto Cominetti, Francisco Facchinei, and Jean B Lasserre. *Modern Optimization Modeling Techniques*. Springer Science & Business Media, 2012.

[105] Roberto Cominetti, Emerson Melo, and Sylvain Sorin. "A payoff-based learning procedure and its application to traffic games". In: *Games and Economic Behavior* 70.1 (2010), pp. 71–83.

[106] Giacomo Como and Rosario Maggistro. "Distributed Dynamic Pricing of Multiscale Transportation Networks". In: *IEEE Transactions on Automatic Control* (2021).

[107] Lauren Craik and Hamsa Balakrishnan. "Equity impacts of the London congestion charging scheme: an empirical evaluation using synthetic control methods". In: *Transportation research record* 2677.5 (2023), pp. 1017–1029.

[108] Peter Cramton, Yoav Shoham, and Richard Steinberg. "Introduction to combinatorial auctions". In: *Combinatorial auctions* (2006), pp. 1–14.

[109] Peter C Cramton, Yoav Shoham, Richard Steinberg, and Vernon L Smith. *Combinatorial auctions*. Vol. 1. 0. MIT press Cambridge, 2006.

[110] Qiwen Cui, Maryam Fazel, and Simon S Du. "Learning Optimal Tax Design in Nonatomic Congestion Games". In: *arXiv preprint arXiv:2402.07437* (2024).

[111] Qiwen Cui, Zhihan Xiong, Maryam Fazel, and Simon S Du. "Learning in congestion games with bandit feedback". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 11009–11022.

[112] Stella Dafermos. "Sensitivity analysis in variational inequalities". In: *Mathematics of Operations Research* 13.3 (1988), pp. 421–434.

[113] Carlos F Daganzo. "A pareto optimum congestion reduction scheme". In: *Transportation Research Part B: Methodological* 29.2 (1995), pp. 139–154.

[114] Carlos F Daganzo and Yosef Sheffi. "On Stochastic Models of Traffic Assignment". In: *Transportation science* 11.3 (1977), pp. 253–274.

[115] Xiaowu Dai and Michael I Jordan. "Learning strategies in decentralized matching markets under uncertain preferences". In: *Journal of Machine Learning Research* 22.260 (2021), pp. 1–50.

[116] Sanmay Das and Emir Kamenica. "Two-Sided Bandits and the Dating Market." In: *IJCAI*. Vol. 5. Citeseer. 2005, p. 19.

[117] Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. "Independent policy gradient methods for competitive reinforcement learning". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 5527–5540.

[118] P DeCorla-Souza. "Income-Based Equity Impacts of Congestion Pricing". In: *Federal Highway administration* (2008).

[119] Lydia Depillis, Rebecca Lieberman, and Crista Chapman. *How the costs of car ownership add up*. 2023. URL: https://www.nytimes.com/interactive/2023/10/07/business/car-ownership-costs.html.

[120] Robert B. Dial. "A Probabilistic Multipath Traffic Assignment Model which Obviates Path Enumeration". In: *Transportation Research* 5.2 (1971), pp. 83–111. ISSN: 0041-1647.

[121] Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo Jovanovic. "Independent policy gradient for large-scale Markov potential games: Sharper rates, function approximation, and game-agnostic convergence". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 5166–5220.

[122] Aasheesh Kumar Dixit, Garima Shakya, Suresh Kumar Jakhar, and Swaprava Nath. "Algorithmic mechanism design for egalitarian and congestion-aware airport slot allocation". In: *Transportation Research Part E: Logistics and Transportation Review* 169 (2023), p. 102971. ISSN: 1366-5545. DOI: https://doi.org/10.1016/j.tre.2022.102971. URL: https://www.sciencedirect.com/science/article/pii/S1366554522003489.

[123] Nico Dogterom, Dick Ettema, and Martin Dijst. "Tradable credits for managing car travel: a review of empirical research and relevant behavioural approaches". In: *Transport Reviews* 37.3 (2017), pp. 322–343.

[124] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. "Strategic classification from revealed preferences". In: *Proceedings of the 2018 ACM Conference on Economics and Computation*. EC '18. Ithaca, NY, USA: Association for Computing Machinery, 2018, pp. 55–70. ISBN: 9781450358293. DOI: 10.1145/3219166.3219193. URL: https://doi.org/10.1145/3219166.3219193.

[125] Ulrich Doraszelski and Juan F Escobar. "A theory of regular Markov perfect equilibria in dynamic stochastic games: Genericity, stability, and purification". In: *Theoretical Economics* 5.3 (2010), pp. 369–402.

[126] Dmitriy Drusvyatskiy and Lin Xiao. "Stochastic optimization with decision-dependent distributions". In: *arXiv* (2020). eprint: 2011.11173 (math.OC).

[127] Maxim Egorov, Vanessa Kuroda, and Peter Sachs. "Encounter aware flight planning in the unmanned airspace". In: *2019 Integrated Communications, Navigation and Surveillance Conference (ICNS)*. IEEE. 2019, pp. 1–15.

[128] Joakim Ekström, Leonid Engelson, and Clas Rydergren. "Heuristic algorithms for a second-best congestion pricing problem". In: *NETNOMICS: Economic Research and Electronic Networking* 10 (2009), pp. 85–102.

[129] Jonas Eliasson. "Road pricing with limited information and heterogeneous users: A successful case". In: *The annals of regional science* 35 (2001), pp. 595–604.

[130] Jonas Eliasson and Lars-Göran Mattsson. "Equity effects of congestion pricing: quantitative methodology and a case study for Stockholm". In: *Transportation Research Part A: Policy and Practice* 40.7 (2006), pp. 602–620.

[131] European Organisation for the Safety of Air Navigation (EUROCONTROL). *U-space ConOps (edition 3.10): U-space capabilities and services to enable Urban Air Mobility*. Tech. rep. EUROCONTROL, 2022.

[132] Antony D. Evans, Maxim Egorov, and Steven Munn. "Fairness in Decentralized Strategic Deconfliction in UTM". In: *AIAA Scitech 2020 Forum*. DOI: 10.2514/6.2020-2203. eprint: https://arc.aiaa.org/doi/pdf/10.2514/6.2020-2203. URL: https://arc.aiaa.org/doi/abs/10.2514/6.2020-2203.

[133] Benjamin Patrick Evans and Sumitra Ganesh. "Learning and Calibrating Heterogeneous Bounded Rational Market Behaviour with Multi-Agent Reinforcement Learning". In: *arXiv preprint arXiv:2402.00787* (2024).

[134] Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

[135] Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media, 2007.

[136] Federal Aviation Administration. *Urban Air Mobility (UAM) Concept of Operations Version 2.0*. Tech. rep. Washington, DC: Federal Aviation Administration, 2023.

[137] Rui Feng, Huixia Zhang, Bin Shi, Qian Zhong, and Baozhen Yao. "Collaborative road pricing strategy for heterogeneous vehicles considering emission constraints". In: *Journal of Cleaner Production* 429 (2023), p. 139561.

[138] Paolo Ferrari. "Road network toll pricing and social welfare". In: *Transportation Research Part B: Methodological* 36.5 (2002), pp. 471–483.

[139] Tanner Fiez, Benjamin Chasnov, and Lillian J Ratliff. "Convergence of learning dynamics in stackelberg games". In: *arXiv preprint arXiv:1906.01217* (2019).

[140] Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. "Online convex optimization in the bandit setting: gradient descent without a gradient". In: *arXiv preprint cs/0408007* (2004).

[141] Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. "Online convex optimization in the bandit setting: Gradient descent without gradient". In: *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '05. Vancouver, British Columbia: Society for Industrial and Applied Mathematics, 2005, pp. 385–394. ISBN: 0898715857.

[142] Lisa Fleischer, Kamal Jain, and Mohammad Mahdian. "Tolls for heterogeneous selfish users in multicommodity networks and generalized congestion games". In: *45th Annual IEEE Symposium on Foundations of Computer Science*. IEEE. 2004, pp. 277–285.

[143] Mogens Fosgerau, Emma Frejinger, and Anders Karlstrom. "A Link-based Network Route Choice Model with Unrestricted Choice Set". In: *Transportation Research Part B: Methodological* 56 (2013), pp. 70–80.

[144] Roy Fox, Stephen M Mcaleer, Will Overman, and Ioannis Panageas. "Independent natural policy gradient always converges in Markov potential games". In: *AISTATS*. PMLR. 2022, pp. 4414–4425.

[145] Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. "Forward and reverse gradient-based hyperparameter optimization". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1165–1173.

[146] Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. "Bilevel programming for hyperparameter optimization and meta-learning". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 1568–1577.

[147] Karen T Frick, Steve Heminger, and Hank Dittmar. "Bay bridge congestion-pricing project: Lessons learned to date". In: *Transportation Research Record* 1558.1 (1996), pp. 29–38.

[148] David Fridovich-Keil, Ellis Ratner, Lasse Peters, Anca D Dragan, and Claire J Tomlin. "Efficient iterative linear-quadratic approximations for nonlinear multi-player general-sum differential games". In: *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2020, pp. 1475–1481.

[149] Drew Fudenberg and David Levine. *The theory of learning in games*. Vol. 2. MIT press, 1998.

[150] Drew Fudenberg and Jean Tirole. *Game Theory*. MIT press, 1991.

[151] David Gale and Lloyd S Shapley. "College admissions and the stability of marriage". In: *The American Mathematical Monthly* 69.1 (1962), pp. 9–15.

[152] Xiang Gao, B. Jiang, and S. Zhang. "On the information-adaptive variants of the ADMM: An iteration complexity perspective". In: *Journal of Scientific Computing* 76 (2018), pp. 327–363.

[153] Aurélien Garivier and Eric Moulines. "On upper-confidence bound policies for switching bandit problems". In: (2011), pp. 174–188.

[154] Kurtuluş Gemici, Elias Koutsoupias, Barnabé Monnot, Christos Papadimitriou, and Georgios Piliouras. "Wealth inequality and the price of anarchy". In: *arXiv preprint arXiv:1802.09269* (2018).

[155] Saeed Ghadimi and Mengdi Wang. "Approximation methods for bilevel programming". In: *arXiv preprint arXiv:1802.02246* (2018).

[156] Avishek Ghosh, Abishek Sankararaman, Kannan Ramchandran, Tara Javidi, and Arya Mazumdar. "Decentralized Competing Bandits in Non - Stationary Matching Markets". In: *arXiv preprint arXiv:2206.00120* (2022).

[157] Gauthier Gidel, Tony Jebara, and Simon Lacoste-Julien. "Frank-wolfe algorithms for saddle point problems". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Vol. 54. Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA: PMLR, 2017, pp. 362–371. URL: `http://proceedings.mlr.press/v54/gidel17a.html`.

[158] Sreenivas Gollapudi, Kostas Kollias, Chinmay Maheshwari, and Manxi Wu. "Online Learning for Traffic Navigation in Congested Networks". In: *International Conference on Algorithmic Learning Theory (ALT)*. PMLR. 2023, pp. 642–662.

[159] Eric J Gonzales and Eleni Christofa. "Empirical assessment of bottleneck congestion with a constant and peak toll: San Francisco–Oakland Bay Bridge". In: *EURO Journal on Transportation and Logistics* 3.3 (2015), pp. 267–288.

[160] PB Goodwin. "How to make road pricing popular". In: *Economic Affairs* 10.5 (1990), pp. 6–7.

[161] Phil B Goodwin. "The rule of three: a possible solution to the political problem of competing objectives for road pricing". In: *Traffic engineering & control* 30.10 (1989), pp. 495–497.

[162] Stephen Gould, Basura Fernando, Anoop Cherian, Peter Anderson, Rodrigo Santa Cruz, and Edison Guo. "On differentiating parameterized argmin and argmax problems with application to bi-level optimization". In: *arXiv preprint arXiv:1607.05447* (2016).

[163] Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. "On the iteration complexity of hypergradient computation". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 3748–3758.

[164] Panagiotis D Grontas, Giuseppe Belgioioso, Carlo Cenedese, Marta Fochesato, John Lygeros, and Florian Dörfler. "BIG Hype: Best intervention in games via distributed hypergradient descent". In: *IEEE Transactions on Automatic Control* (2024).

[165] Panagiotis D Grontas, Carlo Cenedese, Marta Fochesato, Giuseppe Belgioioso, John Lygeros, and Florian Dörfler. "Designing Optimal Personalized Incentive for Traffic Routing using BIG Hype". In: *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE. 2023, pp. 3142–3147.

[166] Nelson M. Guerreiro, George E. Hagen, Jeffrey M. Maddalon, and Ricky W. Butler. "Capacity and Throughput of Urban Air Mobility Vertiports with a First-Come, First-Served Vertiport Scheduling Algorithm". In: *AIAA AVIATION 2020 FORUM*. DOI: 10.2514/6.2020-2903. URL: https://arc.aiaa.org/doi/abs/10.2514/6.2020-2903.

[167] Hugh Gunn. "Spatial and temporal transferability of relationships between travel demand, trip cost and travel time". In: *Transportation Research Part E: Logistics and Transportation Review* 37.2-3 (2001), pp. 163–189.

[168] Hongyi Guo, Zuyue Fu, Zhuoran Yang, and Zhaoran Wang. "Decentralized single-timescale actor-critic on zero-sum two-player stochastic games". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 3899–3909.

[169] Xiaolei Guo and Hai Yang. "Pareto improving congestion pricing and revenue refunding with multiple user classes". In: *Transportation Research Part B: Methodological* 44.8-9 (2010), pp. 972–982.

[170] Xin Guo, Xinyu Li, and Yufei Zhang. "An α-potential game framework for N-player games". In: *arXiv preprint arXiv:2403.16962* (2024).

[171] Xin Guo, Xinyu Li*, Chinmay Maheshwari*, Shankar Sastry, and Manxi Wu. "Markov α-Potential Games". In: *IEEE Transactions on Automatic Control (Provisionally accepted)* (2025). Pre-print available at https://arxiv.org/abs/2305.12553.

[172] Libin Han, Chong Peng, and Zhenyu Xu. "The effect of commuting time on quality of life: evidence from China". In: *International journal of environmental research and public health* 20.1 (2022), p. 573.

[173] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. "Strategic classification". In: *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*. ITCS '16. Cambridge, Massachusetts, USA: Association for Computing Machinery, 2016, pp. 111–122. ISBN: 9781450340571. DOI: 10.1145/2840728.2840730. URL: https://doi.org/10.1145/2840728.2840730.

[174] Tobias Harks, Ingo Kleinert, Max Klimm, and Rolf H Möhring. "Computing network tolls with support constraints". In: *Networks* 65.3 (2015), pp. 262–285.

[175] Tobias Harks and Julian Schwarz. "A unified framework for pricing in nonconvex resource allocation games". In: *SIAM Journal on Optimization* 33.2 (2023), pp. 1223–1249.

[176] Bingsheng He, Shengjie Xu, and Xiaoming Yuan. "Extensions of ADMM for separable convex optimization problems with linear equality or inequality constraints". In: *Handbook of Numerical Analysis*. Vol. 24. Elsevier, 2023, pp. 511–557.

[177] Alexander Heilmeier, Alexander Wischnewski, Leonhard Hermansdorfer, Johannes Betz, Markus Lienkamp, and Boris Lohmann. "Minimum curvature trajectory planning and control for an autonomous race car". In: *Vehicle System Dynamics* 58.10 (2020), pp. 1497–1527. DOI: 10.1080/00423114.2019.1631455.

[178] Amélie Heliou, Johanne Cohen, and Panayotis Mertikopoulos. "Learning with bandit feedback in potential games". In: *Advances in Neural Information Processing Systems* 30 (2017).

[179] James Herman, Jonathan Francis, Siddha Ganju, Bingqing Chen, Anirudh Koul, Abhinav Gupta, Alexey Skabelkin, Ivan Zhukov, Max Kumskoy, and Eric Nyberg. "Learn-to-race: A multimodal control environment for autonomous racing". In: *proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

[180] Rainer Hettich and Kenneth O Kortanek. "Semi-infinite programming: theory, methods, and applications". In: *SIAM review* 35.3 (1993), pp. 380–429.

[181] Morris W Hirsch. "Systems of differential equations that are competitive or cooperative II: Convergence almost everywhere". In: *SIAM Journal on Mathematical Analysis* 16.3 (1985), pp. 423–439.

[182] Martin Hoefer, Lars Olbrich, and Alexander Skopalik. "Taxing subnetworks". In: *International workshop on internet and network economics*. Springer. 2008, pp. 286–294.

[183] Josef Hofbauer and William H Sandholm. "On the global convergence of stochastic fictitious play". In: *Econometrica* 70.6 (2002), pp. 2265–2294.

[184] Josef Hofbauer and Karl Sigmund. "Evolutionary game dynamics". In: *Bulletin of the American Mathematical Society* 40.4 (2003), pp. 479–519.

[185] Josef Hofbauer and Sylvain Sorin. "Best response dynamics for continuous zero-sum games". In: *Discrete and Continuous Dynamical Systems Series B* 6.1 (2006), p. 215.

[186] Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. "Does distributionally robust supervised learning give robust classifiers?" In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 2029–2037. URL: `http://proceedings.mlr.press/v80/hu18a.html`.

[187] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. "Black-box adversarial attacks with limited queries and information". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 2137–2146. URL: `http://proceedings.mlr.press/v80/ilyas18a.html`.

[188] Meena Jagadeesan, Alexander Wei, Yixin Wang, Michael Jordan, and Jacob Steinhardt. "Learning Equilibria in Matching Markets from Bandit Feedback". In: *Advances in Neural Information Processing Systems* 34 (2021).

[189] Olaf Jahn, Rolf H Möhring, Andreas S Schulz, and Nicolás E Stier-Moses. "System-optimal routing of traffic flows with user constraints in networks with congestion". In: *Operations research* 53.4 (2005), pp. 600–616.

[190] Achin Jain and Manfred Morari. "Computing the racing line using Bayesian optimization". In: *2020 59th IEEE Conference on Decision and Control (CDC)* (2020), pp. 6192–6197.

[191] Prateek Jain, Dheeraj M. Nagaraj, and Praneeth Netrapalli. "SGD without replacement: sharper rates for general smooth convex functions". In: *ICML*. 2019.

[192] Devansh Jalota, Karthik Gopalakrishnan, Navid Azizan, Ramesh Johari, and Marco Pavone. "Online Learning for Traffic Routing under Unknown Preferences". In: *arXiv preprint arXiv:2203.17150* (2022).

[193] Devansh Jalota, Marco Pavone, Qi Qi, and Yinyu Ye. "Fisher markets with linear constraints: Equilibrium properties and efficient distributed algorithms". In: *Games and Economic Behavior* 141 (2023), pp. 223–260.

[194] Devansh Jalota, Kiril Solovey, Karthik Gopalakrishnan, Stephen Zoepf, Hamsa Balakrishnan, and Marco Pavone. "When efficiency meets equity in congestion pricing and revenue refunding schemes". In: *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 2021, pp. 1–11.

[195] Devansh Jalota, Kiril Solovey, Matthew Tsao, Stephen Zoepf, and Marco Pavone. "Balancing fairness and efficiency in traffic routing via interpolated traffic assignment". In: *Autonomous Agents and Multi-Agent Systems* 37.2 (2023), p. 32.

[196] K Ji, J Yang, and Y Liang. "Bilevel Optimization: Convergence Analysis and Enhanced Design. arXiv e-prints, art". In: *arXiv preprint arXiv:2010.07962* (2020).

[197] Yixuan Jia, Maulik Bhatt, and Negar Mehr. "Rapid: Autonomous multi-agent racing using constrained potential dynamic games". In: *2023 European Control Conference (ECC)*. IEEE. 2023, pp. 1–8.

[198] Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. "V-Learning—A Simple, Efficient, Decentralized Algorithm for Multiagent Reinforcement Learning". In: *Mathematics of Operations Research* 49.4 (2024), pp. 2295–2322. DOI: 10.1287/moor.2021.0317.

[199] Ramesh Johari, Vijay Kamble, and Yash Kanoria. "Matching while learning". In: *arXiv preprint arXiv:1603.04549* (2016).

[200] Chanyoung Jung, Seungwook Lee, Hyunki Seong, Andrea Finazzi, and David Hyunchul Shim. "Game-theoretic model predictive control with data-driven identification of vehicle model for head-to-head autonomous racing". In: *arXiv preprint arXiv:2106.04094* (2021).

[201] Juraj Kabzan, Miguel I Valls, Victor JF Reijgwart, Hubertus FC Hendrikx, Claas Ehmke, Manish Prajapat, Andreas Bühler, Nikhil Gosala, Mehak Gupta, Ramya Sivanesan, et al. "AMZ driverless: The full autonomous racing system". In: *Journal of Field Robotics* 37.7 (2020), pp. 1267–1294.

[202] Dvij Kalaria, Qin Lin, and John M Dolan. "Towards Optimal Head-to-head Autonomous Racing with Curriculum Reinforcement Learning". In: *arXiv preprint arXiv: 2308.13491* (2023).

[203] Dvij Kalaria, Qin Lin, and John M. Dolan. "Adaptive Planning and Control with Time-Varying Tire Models for Autonomous Racing Using Extreme Learning Machine". In: *ArXiv* abs/2303.08235 (2023). URL: https://api.semanticscholar.org/CorpusID:257532643.

[204] Vyacheslav V Kalashnikov, Roberto Carlos Herrera Maldonado, José-Fernando Cama cho-Vallejo, and Nataliya I Kalashnykova. "A heuristic algorithm solving bilevel toll optimization problems". In: *The International Journal of Logistics Management* 27.1 (2016), pp. 31–51.

[205] Dileep Kalathil, Naumaan Nayyar, and Rahul Jain. "Decentralized learning for multiplayer multiarmed bandits". In: *IEEE Transactions on Information Theory* 60.4 (2014), pp. 2331–2345.

[206] George Karakostas and Stavros G Kolliopoulos. "Edge pricing of multicommodity networks for heterogeneous selfish users". In: *FOCS*. Vol. 4. 2004, pp. 268–276.

[207] Hamed Karimi, Julie Nutini, and Mark Schmidt. "Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD*. Springer. 2016, pp. 795–811.

[208] Alexander Karpov. "A necessary and sufficient condition for uniqueness consistency in the stable marriage matching problem". In: *Economics Letters* 178 (2019), pp. 63–65.

[209] Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. "Champion-level drone racing using deep reinforcement learning". In: *Nature* 620.7976 (2023), pp. 982–987.

[210] Talha Kavuncu, Ayberk Yaraneri, and Negar Mehr. "Potential ilqr: A potential-minimizing controller for planning multi-agent interactive trajectories". In: *arXiv preprint arXiv:2107.04926* (2021).

[211] Imke C. Kleinbekman, Mihaela A. Mitici, and Peng Wei. "eVTOL Arrival Sequencing and Scheduling for On-Demand Urban Air Mobility". In: *2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC)*. 2018, pp. 1–7. DOI: 10.1109/DASC.2018.8569645.

[212] Robert Kleinberg and Tom Leighton. "The value of knowing a demand curve: Bounds on regret for online posted-price auctions". In: *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.* IEEE. 2003, pp. 594–605.

[213] Robert Kleinberg, Georgios Piliouras, and Éva Tardos. "Multiplicative Updates Outperform Generic no-Regret Learning in Congestion Games". In: *Proceedings of the forty-first annual ACM symposium on Theory of computing*. 2009, pp. 533–542.

[214] Vijay Konda and John Tsitsiklis. "Actor-critic algorithms". In: *Advances in Neural Information Processing Systems* 12 (1999).

[215] Fang Kong, Junming Yin, and Shuai Li. "Thompson Sampling for Bandit Learning in Matching Markets". In: *arXiv preprint arXiv:2204.12048* (2022).

[216] Syrine Krichene, Walid Krichene, Roy Dong, and Alexandre Bayen. "Convergence of Heterogeneous Distributed Learning in Stochastic Routing Games". In: *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 2015, pp. 480–487.

[217] Walid Krichene, Benjamin Drighes, and Alexandre Bayen. "On the Convergence of no-Regret Learning in Selfish Routing". In: *International Conference on Machine Learning*. PMLR. 2014, pp. 163–171.

[218] Walid Krichene, Benjamin Drighès, and Alexandre Bayen. "On the convergence of no-regret learning in selfish routing". In: *International Conference on Machine Learning*. PMLR. 2014, pp. 163–171.

[219] Ramakrishnan Krishnamurthy and Aditya Gopalan. "On Slowly-varying non - stationary Bandits". In: *arXiv preprint arXiv:2110.12916* (2021).

[220] Sukanya Kudva, Kshitij Kulkarni, Chinmay Maheshwari, Anil Aswani, and Shankar Sastry. "Understanding the Impact of Coalitions between EV Charging Stations". In: *2024 IEEE 63rd Conference on Decision and Control (CDC)*. 2024, pp. 6161–6167. DOI: 10.1109/CDC56724.2024.10886418.

[221] Erich Kutschinski, Thomas Uthmann, and Daniel Polani. "Learning competitive pricing strategies by multi-agent reinforcement learning". In: *Journal of Economic Dynamics and Control* 27.11-12 (2003), pp. 2207–2218.

[222] Martine Labbé, Patrice Marcotte, and Gilles Savard. "A bilevel model of taxation and its application to optimal highway pricing". In: *Management science* 44.12-part-1 (1998), pp. 1608–1622.

[223] Tze Leung Lai and Herbert Robbins. "Asymptotically efficient adaptive allocation rules". In: *Advances in applied mathematics* 6.1 (1985), pp. 4–22.

[224] Forrest Laine, David Fridovich-Keil, Chih-Yuan Chiu, and Claire Tomlin. "Multi-hypothesis interactions in game-theoretic motion planning". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 8016–8023.

[225] Chandrashekar Lakshminarayanan and Shalabh Bhatnagar. "A stability criterion for two timescale stochastic approximation schemes". In: *Automatica* 79 (2017), pp. 108–114.

[226] Rida Laraki and Panayotis Mertikopoulos. "Higher order game dynamics". In: *Journal of Economic Theory* 148.6 (2013), pp. 2666–2695.

[227] Torbjörn Larsson and Michael Patriksson. "Side constrained traffic equilibrium models—traffic management through link tolls". In: *Equilibrium and advanced transportation modelling*. Springer, 1998, pp. 125–151.

[228] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[229] S Lawphongpanich and Y Yin. "Pareto - improving congestion pricing for general road networks". In: *Technique Report. Gainesville, FL: Department of industrial and System Engineering, University of Florida* (2007).

[230] Siriphong Lawphongpanich and Donald W Hearn. "An MPEC approach to second-best toll pricing". In: *Mathematical programming* 101.1 (2004), pp. 33–55.

[231] Siriphong Lawphongpanich and Yafeng Yin. "Solving the Pareto - improving toll problem via manifold suboptimization". In: *Transportation Research Part C: Emerging Technologies* 18.2 (2010), pp. 234–246.

[232] Daniel A Lazar and Ramtin Pedarsani. "Optimal tolling for multitype mixed autonomous traffic networks". In: *IEEE Control Systems Letters* 5.5 (2020), pp. 1849–1854.

[233] Daniel A Lazar and Ramtin Pedarsani. "The Role of Differentiation in Tolling of Traffic Networks with Mixed Autonomy". In: *arXiv preprint arXiv:2103.13553* (2021).

[234] Phil LeBeau. *Traffic jams cost US $87 billion in lost productivity in 2018, and Boston and DC have the nation's worst*. 2019.

[235] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. "Gradient descent only converges to minimizers". In: *Conference on learning theory*. PMLR. 2016, pp. 1246–1257.

[236] John M. Lee. *Introduction to Smooth Manifolds*. Springer Science+Business Media New York, 2013.

[237] Christopher Leet and Robert A. Morris. "Combinatorial Auction-Based Strategic Deconfliction of Federated UTM Airspace". In: *AIAA AVIATION FORUM AND ASCEND 2024*. DOI: 10.2514/6.2024-4454. eprint: https://arc.aiaa.org/doi/pdf/10.2514/6.2024-4454. URL: https://arc.aiaa.org/doi/abs/10.2514/6.2024-4454.

[238] Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. "Global convergence of multi-agent policy gradient in markov potential games". In: *arXiv preprint arXiv:2106.01969* (2021).

[239] David S Leslie and Edmund J Collins. "Generalised weakened fictitious play". In: *Games and Economic Behavior* 56.2 (2006), pp. 285–298.

[240] David S Leslie, Steven Perkins, and Zibo Xu. "Best-response dynamics in zero-sum stochastic games". In: *Journal of Economic Theory* 189 (2020), p. 105095.

[241] Joshua Letchford, Vincent Conitzer, and Kamesh Munagala. "Learning and approximating the optimal strategy to commit to". In: *Algorithmic Game Theory: Second International Symposium, SAGT 2009, Paphos, Cyprus, October 18-20, 2009. Proceedings 2*. Springer. 2009, pp. 250–262.

[242] Dan Levin. "Taxation within Cournot oligopoly". In: *Journal of Public Economics* 27.3 (1985), pp. 281–290.

[243] Jianhui Li, Youcheng Niu, Shuang Li, Yuzhe Li, Jinming Xu, and Junfeng Wu. "Inducing Desired Equilibrium in Taxi Repositioning Problem with Adaptive Incentive Design". In: *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE. 2023, pp. 8075–8080.

[244] Jiayang Li, Jing Yu, Qianni Wang, Boyi Liu, Zhaoran Wang, and Yu Marco Nie. "Differentiable Bilevel Programming for Stackelberg Congestion Games". In: *arXiv preprint arXiv:2209.07618* (2022).

[245] Jiayi Li, Matthew Motoki, and Baosen Zhang. "Socially optimal energy usage via adaptive pricing". In: *Electric Power Systems Research* 235 (2024), p. 110640.

[246] Xiaocheng Li, Chunlin Sun, and Yinyu Ye. "Simple and fast algorithm for binary integer and online linear programming". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9412–9421.

[247] Alvin Chong Lim. *Transportation network design problems: An MPEC approach*. The Johns Hopkins University, 2002.

[248] Xi Lin, Yafeng Yin, and Fang He. "Credit-based mobility management considering travelers' budgeting behaviors under uncertainty". In: *Transportation Science* 55.2 (2021), pp. 297–314.

[249] Alexander Liniger, Alexander Domahidi, and Manfred Morari. "Optimization-based autonomous racing of 1: 43 scale RC cars". In: *Optimal Control Applications and Methods* 36.5 (2015), pp. 628–647.

[250] Alexander Liniger and John Lygeros. "A noncooperative game approach to autonomous racing". In: *IEEE Transactions on Control Systems Technology* 28.3 (2019), pp. 884–897.

[251] Nick Littlestone and Manfred K Warmuth. "The weighted majority algorithm". In: *Information and computation* 108.2 (1994), pp. 212–261.

[252] Michael L Littman. "Markov games as a framework for multi-agent reinforcement learning". In: *Machine learning proceedings 1994*. Elsevier, 1994, pp. 157–163.

[253] Boyi Liu, Jiayang Li, Zhuoran Yang, Hoi-To Wai, Mingyi Hong, Yu Nie, and Zhaoran Wang. "Inducing Equilibria via Incentives: Simultaneous Design-and-play ensures Global Convergence". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 29001–29013.

[254] Jennifer Liu. *Commuters in this city spend 119 hours a year stuck in traffic*. 2020.

[255] Keqin Liu and Qing Zhao. "Distributed learning in multi-armed bandit with multiple players". In: *IEEE transactions on signal processing* 58.11 (2010), pp. 5667–5681.

[256] Lydia T Liu, Horia Mania, and Michael Jordan. "Competing bandits in matching markets". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 1618–1628.

[257] Lydia T Liu, Feng Ruan, Horia Mania, and Michael I Jordan. "Bandit learning in decentralized matching markets". In: *Journal of Machine Learning Research* 22.211 (2021), pp. 1–34.

[258] Risheng Liu, Jiaxin Gao, Jin Zhang, Deyu Meng, and Zhouchen Lin. "Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.12 (2021), pp. 10045–10067.

[259] Risheng Liu, Xuan Liu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. "A value-function-based interior-point method for non-convex bi-level optimization". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 6882–6892.

[260] Risheng Liu, Yaohua Liu, Shangzhi Zeng, and Jin Zhang. "Towards gradient-based bilevel optimization with non-convex followers and beyond". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 8662–8675.

[261] Risheng Liu, Pan Mu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. "A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 6305–6315.

[262] Sijia Liu, Songtao Lu, Xiangyi Chen, Yao Feng, Kaidi Xu, Abdullah Al-Dujaili, Mingyi Hong, and Una-May O'Reilly. "Min-max optimization without gradients: Convergence and applications to adversarial ML". In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 6282–6293. URL: `http://proceedings.mlr.press/v119/liu20j.html`.

[263] Xinjie Liu, Lasse Peters, and Javier Alonso-Mora. "Learning to play trajectory games against opponents with unknown objectives". In: *IEEE Robotics and Automation Letters* 8.7 (2023), pp. 4139–4146.

[264] Gábor Lugosi and Abbas Mehrabian. "Multiplayer bandits without observing collision information". In: *Mathematics of Operations Research* (2021).

[265] Shaocong Ma, Ziyi Chen, Shaofeng Zou, and Yi Zhou. "Decentralized robust v-learning for solving markov games with model uncertainty". In: *Journal of Machine Learning Research* 24.371 (2023), pp. 1–40.

[266] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and A. Vladu. "Towards deep learning models resistant to adversarial attacks". In: *ArXiv* (June 2017).

[267] Chinmay Maheshwari, James Cheng, Shankar Sastry, Lillian Ratliff, and Eric Mazumdar. "Follower Agnostic Learning in Stackelberg Games". In: *2024 IEEE 63rd Conference on Decision and Control (CDC)*. IEEE. 2024, pp. 222–228.

[268] Chinmay Maheshwari, Chih-Yuan Chiu, Eric Mazumdar, Shankar Sastry, and Lillian Ratliff. "Zeroth-order methods for convex-concave min-max problems: Applications to decision-dependent risk minimization". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 6702–6734.

[269] Chinmay Maheshwari, Kshitij Kulkarni, Druv Pai, Jiarui Yang, Manxi Wu, and Shankar Sastry. "Congestion Pricing for Efficiency and Equity: Theory and Applications to the San Francisco Bay Area". In: (2024). Accepted for poster presentation in the ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO), 2024. Available at `https://arxiv.org/abs/2401.16844`.

[270] Chinmay Maheshwari, Kshitij Kulkarni, Manxi Wu, and S Shankar Sastry. "Dynamic tolling for inducing socially optimal traffic loads". In: *2022 American Control Conference (ACC)*. IEEE. 2022, pp. 4601–4607.

[271] Chinmay Maheshwari, Kshitij Kulkarni, Manxi Wu, and S Shankar Sastry. "Inducing Social Optimality in Games via Adaptive Incentive Design". In: *IEEE 61st Conference on Decision and Control (CDC)*. IEEE. 2022, pp. 2864–2869.

[272] Chinmay Maheshwari, Kshitij Kulkarni, Manxi Wu, and Shankar Sastry. "Adaptive Incentive Design with Learning Agents". In: *arXiv preprint arXiv:2405.16716* (2024).

[273] Chinmay Maheshwari, Maria G Mendoza, Victoria Tuck, Pan-Yang Su, Victor L Qin, Sanjit Seshia, Hamsa Balakrishnan, and Shankar Sastry. "Privacy-Preserving Mechanisms for Coordinating Airspace Usage in Advanced Air Mobility". In: *Journal on Autonomous Transportation Systems* (2024).

[274] Chinmay Maheshwari, Shankar Sastry, and Eric Mazumdar. "Decentralized, Communication and Coordination-free Learning in Structured Matching Markets". In: *Advances in Neural Information Processing Systems (NeuRIPS)*. Vol. 35. 2022, pp. 15081–15092.

[275] Chinmay Maheshwari, Manxi Wu, Druv Pai, and Shankar Sastry. "Independent and Decentralized Learning in Markov Potential Games". In: *IEEE Transactions on Automatic Control* (2025). Pre-print available at `https://arxiv.org/abs/2205.14590`.

[276] Chinmay Maheshwari, Manxi Wu, and Shankar Sastry. "Convergence of Decentralized Actor-Critic Algorithm in General-Sum Markov Games". In: *IEEE Control Systems Letters* (2024).

[277] Chinmay Maheshwari*, Dvij Kalaria*, and Shankar Sastry. "Alpha-Racer: Real-Time Algorithms for Game-Theoretic Motion Planning and Control in Autonomous Racing using Near-Potential Function". In: *7th Annual Learning for Dynamics and Control Conference (L4DC) 2025*. IEEE. 2025.

[278] Chinmay Maheshwari*, Chih-Yuan Chiu*, Pan-Yang Su, and Shankar Sastry. "Arc-based Traffic Assignment: Equilibrium Characterization and Learning". In: *IEEE Conference on Decision and Control (CDC)*. 2023.

[279] Tien Mai. "A Method of Integrating Correlation Structures for a Generalized Recursive Route Choice Model". In: *Transportation Research Part B: Methodological* 93 (2016), pp. 146–161.

[280] Tien Mai, Mogens Fosgerau, and Emma Frejinger. "A Nested Recursive Logit Model for Route Choice Analysis". In: *Transportation Research Part B: Methodological* 75 (2015), pp. 100–112.

[281] Yishay Mansour, Aleksandrs Slivkins, and Zhiwei Steven Wu. "Competing bandits: Learning under competition". In: *arXiv preprint arXiv:1702.08533* (2017).

[282] Traffic Assignment Manual. *For Application With a Large, High Speed Computer*. 1964.

[283] Weichao Mao and Tamer Başar. "Provably efficient reinforcement learning in decentralized general-sum markov games". In: *Dynamic Games and Applications* 13.1 (2023), pp. 165–186.

[284] Weichao Mao, Tamer Başar, Lin F Yang, and Kaiqing Zhang. "Decentralized Cooperative Multi-Agent Reinforcement Learning with Exploration". In: *arXiv preprint arXiv:2110.05707* (2021).

[285] Patrice Marcotte and DL Zhu. "Existence and computation of optimal tolls in multiclass network equilibrium problems". In: *Operations Research Letters* 37.3 (2009), pp. 211–214.

[286] Jason R Marden. "State based potential games". In: *Automatica* 48.12 (2012), pp. 3075–3088.

[287] Jason R Marden, Gürdal Arslan, and Jeff S Shamma. "Joint strategy fictitious play with inertia for potential games". In: *IEEE Transactions on Automatic Control* 54.2 (2009), pp. 208–220.

[288] A. Mas-Colell, M.D. Whinston, and J.R. Green. *Microeconomic Theory*. Oxford University Press, 2006. ISBN: 9780198089537. URL: `https://books.google.com/books?id=kDtHswEACAAJ`.

[289] Laetitia Matignon, Guillaume J Laurent, and Nadine Le Fort-Piat. "Independent reinforcement learners in cooperative Markov games: A survey regarding coordination problems". In: *The Knowledge Engineering Review* 27.1 (2012), pp. 1–31.

[290] Eric Mazumdar, Lillian J Ratliff, and S Shankar Sastry. "On gradient-based learning in continuous games". In: *SIAM Journal on Mathematics of Data Science* 2.1 (2020), pp. 103–131.

[291] Eric Mazumdar, Lillian J. Ratliff, Michael I. Jordan, and S. Shankar Sastry. "Policy-Gradient Algorithms Have No Guarantees of Convergence in Linear Quadratic Games". In: *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS '20. Auckland, New Zealand, 2020, pp. 860–868. ISBN: 9781450375184.

[292] Peter McCullagh and John A Nelder. *Generalized Linear Models*. Routledge, 2019.

[293] Negar Mehr and Roberto Horowitz. "Pricing traffic networks with mixed vehicle autonomy". In: *2019 American Control Conference (ACC)*. IEEE. 2019, pp. 2676–2682.

[294] Ruta Mehta and Vijay V. Vazirani. "An incentive compatible, efficient market for air traffic flow management". In: *Theoretical Computer Science* 818 (2020). Computing and Combinatorics, pp. 41–50. ISSN: 0304-3975. DOI: https://doi.org/10.1016/j.tcs.2018.09.006. URL: https://www.sciencedirect.com/science/article/pii/S0304397518305711.

[295] Emily Meigs, Francesca Parise, and Asuman Ozdaglar. "Learning Dynamics in Stochastic Routing Games". In: *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 2017, pp. 259–266.

[296] Panayotis Mertikopoulos, Ya-Ping Hsieh, and Volkan Cevher. "A unified stochastic approximation framework for learning in games". In: *Mathematical Programming* 203.1 (2024), pp. 559–609.

[297] Panayotis Mertikopoulos and William H Sandholm. "Learning in games via reinforcement and regularization". In: *Mathematics of Operations Research* 41.4 (2016), pp. 1297–1324.

[298] Panayotis Mertikopoulos and Zhengyuan Zhou. "Learning in games with continuous action sets and unknown payoff functions". In: *Mathematical Programming* 173.1 (2019), pp. 465–507.

[299] David Meunier and Emile Quinet. "Value of time estimations in cost benefit analysis: the French experience". In: *Transportation Research Procedia* 8 (2015), pp. 62–71.

[300] John Miller, Juan Perdomo, and Tijana Zrnic. "Outside the echo chamber: Optimizing the performative risk". In: *arXiv preprint arXiv:2102.08570* (2021).

[301] Yifei Min, Tianhao Wang, Ruitu Xu, Zhaoran Wang, Michael I Jordan, and Zhuoran Yang. "Learn to Match with No Regret: Reinforcement Learning in Markov Matching Markets". In: *arXiv preprint arXiv:2203.03684* (2022).

[302] Eduardo Mojica-Nava, Jorge I Poveda, and Nicanor Quijano. "Stackelberg Population Learning Dynamics". In: *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE. 2022, pp. 6395–6400.

[303] Aryan Mokhtari, A. Ozdaglar, and S. Pattathil. "A Unified Analysis of Extra-gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach". In: *AISTATS*. 2020.

[304] Aryan Mokhtari, A. Ozdaglar, and S. Pattathil. "Convergence Rate of O(1/k) for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems". In: *SIAM J. Optim.* 30 (2020), pp. 3230–3251.

[305] Dov Monderer and Lloyd S Shapley. "Fictitious play property for games with identical interests". In: *Journal of economic theory* 68.1 (1996), pp. 258–265.

[306] Dov Monderer and Lloyd S Shapley. "Potential games". In: *Games and Economic Behavior* 14.1 (1996), pp. 124–143.

[307] Barnabé Monnot, Francisco Benita, and Georgios Piliouras. "How bad is selfish routing in practice?" In: *arXiv preprint arXiv:1703.01599* (2017).

[308] Kazuo Murota. "Discrete Convex Analysis". In: *Hausdorff Institute of Mathematics, Summer School (September 21–25, 2015)* (2015). eprint: `https://kzmurota.fpark.tmu.ac.jp/paper/HIMSummerSchool15Murota.pdf`.

[309] Deepan Muthirayan, Chinmay Maheshwari, Pramod Khargonekar, and Shankar Sastry. "Competing Bandits in Time-varying Matching Markets". In: *Learning for Dynamics and Control (L4DC) Conference*. PMLR. 2023, pp. 1020–1031.

[310] Roger B. Myerson. "Optimal Auction Design". In: *Mathematics of Operations Research* 6.1 (1981), pp. 58–73. DOI: `10.1287/moor.6.1.58`. eprint: `https://doi.org/10.1287/moor.6.1.58`. URL: `https://doi.org/10.1287/moor.6.1.58`.

[311] Katsuhiko Nakamura and Kara Maria Kockelman. "Congestion pricing and roadspace rationing: an application to the San Francisco Bay Bridge corridor". In: *Transportation Research Part A: Policy and Practice* 36.5 (2002), pp. 403–417.

[312] Hongseok Namkoong and John C Duchi. "Stochastic gradient methods for distributionally robust optimization with f-divergences". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. 2016. URL: `https://proceedings.neurips.cc/paper/2016/file/4588e674d3f0faf985047d4c3f13ed0d-Paper.pdf`.

[313] Dheeraj Narasimha, Kiyeob Lee, Dileep Kalathil, and Srinivas Shakkottai. "Multi-Agent Learning via Markov Potential Games in Marketplaces for Distributed Energy Resources". In: *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE. 2022, pp. 6350–6357.

[314] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. 1st ed. Springer Publishing Company, Incorporated, 2014. ISBN: 1461346916.

[315] Yurii Nesterov et al. *Lectures on convex optimization*. Vol. 137. Springer, 2018.

[316] Yurii Nesterov and Vladimir Spokoiny. "Random gradient-free minimization of convex functions". In: *Found. Comput. Math.* 17.2 (Apr. 2017), pp. 527–566. ISSN: 1615-3375. DOI: `10.1007/s10208-015-9296-2`. URL: `https://doi.org/10.1007/s10208-015-9296-2`.

[317] Yu Marco Nie. "Transaction costs and tradable mobility credits". In: *Transportation Research Part B: Methodological* 46.1 (2012), pp. 189–203.

[318] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani. *Algorithmic Game Theory*. Cambridge University Press, 2007.

[319] Gennaro Notomista, Mingyu Wang, Mac Schwager, and Magnus Egerstedt. "Enhancing game-theoretic autonomous car racing using control barrier functions". In: *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2020, pp. 5393–5399.

[320] Giuseppe Nuti, Mahnoosh Mirghaemi, Philip Treleaven, and Chaiyakorn Yingsaeree. "Algorithmic trading". In: *Computer* 44.11 (2011), pp. 61–69.

[321] Steven J O'Hare, Richard D Connors, and David P Watling. "Mechanisms that govern how the price of anarchy varies with travel demand". In: *Transportation Research Part B: Methodological* 84 (2016), pp. 55–80.

[322] Daniel E Ochoa and Jorge I Poveda. "High-performance optimal incentive-seeking in transactive control for traffic congestion". In: *European Journal of Control* 68 (2022), p. 100696.

[323] Amedeo R. Odoni. "The Flow Management Problem in Air Traffic Control". In: *Flow Control of Congested Networks*. Ed. by Amedeo R. Odoni, Lucio Bianco, and Giorgio Szegö. Berlin, Heidelberg: Springer Berlin Heidelberg, 1987, pp. 269–288. ISBN: 978-3-642-86726-2.

[324] Yuki Oyama, Yusuke Hara, and Takashi Akamatsu. "Markovian Traffic Equilibrium Assignment Based on Network Generalized Extreme Value Model". In: *Transportation Research Part B: Methodological* 155 (2022), pp. 135–159.

[325] Yuki Oyama and Eiji Hato. "A Discounted Recursive Logit Model for Dynamic Gridlock Network Analysis". In: *Transportation Research Part C: Emerging Technologies* 85 (2017), pp. 509–527.

[326] Yuki Oyama and Eiji Hato. "Prism-based Path Set Restriction for Solving Markovian Traffic Assignment Problem". In: *Transportation Research Part B: Methodological* 122 (2019), pp. 528–546.

[327] Raymond Palmquist, Daniel Phaneuf, and V Kerry Smith. *Measuring the values for time*. 2007.

[328] Ioannis Panageas and Georgios Piliouras. "Average case performance of replicator dynamics in potential games via computing regions of attraction". In: *Proceedings of the 2016 ACM Conference on Economics and Computation*. 2016, pp. 703–720.

[329] Christos H Papadimitriou. "The complexity of finding Nash equilibria". In: *Algorithmic game theory* 2 (2007), p. 30.

[330] Andrea Papola and Vittorio Marzano. "A Network Generalized Extreme Value Model for Route Choice Allowing Implicit Route Enumeration". In: *Computer-Aided Civil and Infrastructure Engineering* 28.8 (2013), pp. 560–580.

[331] Michael Patriksson. *The traffic assignment problem: models and methods*. Courier Dover Publications, 2015.

[332] Michael Patriksson and R Tyrrell Rockafellar. "A mathematical model and descent algorithm for bilevel traffic management". In: *Transportation Science* 36.3 (2002), pp. 271–291.

[333] Fabian Pedregosa. "Hyperparameter optimization with approximate gradient". In: *International conference on machine learning*. PMLR. 2016, pp. 737–746.

[334] Binghui Peng, Weiran Shen, Pingzhong Tang, and Song Zuo. "Learning optimal strategies to commit to". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 2149–2156.

[335] Marco Percoco. "Heterogeneity in the reaction of traffic flows to road pricing: a synthetic control approach applied to Milan". In: *Transportation* 42 (2015), pp. 1063–1079.

[336] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. "Performative prediction". In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 7599–7609. URL: http://proceedings.mlr.press/v119/perdomo20a.html.

[337] Steven Perkins and David S Leslie. "Asynchronous stochastic approximation with differential inclusions". In: *Stochastic Systems* 2.2 (2013), pp. 409–446.

[338] Julien Perolat, Bilal Piot, and Olivier Pietquin. "Actor-critic fictitious play in simultaneous move multistage games". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2018, pp. 919–928.

[339] Julien Perolat, Bruno Scherrer, Bilal Piot, and Olivier Pietquin. "Approximate dynamic programming for two-player zero-sum Markov games". In: *International Conference on Machine Learning*. PMLR. 2015, pp. 1321–1329.

[340] Thomas Pertuiset and Georgina Santos. "Primary auction of slots at European airports". In: *Research in Transportation Economics* 45 (2014). Pricing and Regulation in the Airline Industry, pp. 66–71. ISSN: 0739-8859. DOI: https://doi.org/10.1016/j.retrec.2014.07.009. URL: https://www.sciencedirect.com/science/article/pii/S0739885914000304.

[341] Lasse Peters, Andrea Bajcsy, Chih-Yuan Chiu, David Fridovich-Keil, Forrest Laine, Laura Ferranti, and Javier Alonso-Mora. "Contingency games for multi-agent interaction". In: *IEEE Robotics and Automation Letters* (2024).

[342] Sock-Yong Phang and Rex S Toh. "Road congestion pricing in Singapore: 1975 to 2003". In: *Transportation Journal* (2004), pp. 16–25.

[343] Jorge I Poveda, Philip N Brown, Jason R Marden, and Andrew R Teel. "A class of distributed adaptive pricing mechanisms for societal systems with limited information". In: *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE. 2017, pp. 1490–1495.

[344] KJ Prabuchandran, Hemanth Kumar AN, and Shalabh Bhatnagar. "Multi-agent reinforcement learning for traffic signal control". In: *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2014, pp. 2529–2534.

[345] HL Prasad, Prashanth LA, and Shalabh Bhatnagar. "Two-timescale algorithms for learning Nash equilibria in general-sum stochastic games". In: *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*. 2015, pp. 1371–1379.

[346] Victor Qin and Hamsa Balakrishnan. "Cost-Aware Congestion Management Protocols for Advanced Air Mobility". In: (2022).

[347] S. J. Rassenti, V. L. Smith, and R. L. Bulfin. "A Combinatorial Auction Mechanism for Airport Time Slot Allocation". In: *The Bell Journal of Economics* 13.2 (1982), pp. 402–417. ISSN: 0361915X. URL: `http://www.jstor.org/stable/3003463` (visited on 01/03/2024).

[348] Lillian J Ratliff and Tanner Fiez. "Adaptive incentive design". In: *IEEE Transactions on Automatic Control* 66.8 (2020), pp. 3871–3878.

[349] Benjamin Recht and Stephen Wright. *Optimization for Data Analysis*. 1st ed. Cambridge University Press, 2021. ISBN: 1316518981.

[350] Jonathan Rosenski, Ohad Shamir, and Liran Szlak. "Multi-player bandits–a musical chairs approach". In: *International Conference on Machine Learning*. PMLR. 2016, pp. 155–163.

[351] Robert W Rosenthal. "A class of games possessing pure-strategy Nash equilibria". In: *International Journal of Game Theory* 2.1 (1973), pp. 65–67.

[352] Ugo Rosolia and Francesco Borrelli. "Learning How to Autonomously Race a Car: A Predictive Control Approach". In: *IEEE Transactions on Control Systems Technology* 28 (2020), pp. 2713–2719.

[353] Ugo Rosolia, Ashwin Carvalho, and Francesco Borrelli. "Autonomous racing using learning model predictive control". In: *2017 American control conference (ACC)*. IEEE. 2017, pp. 5115–5120.

[354] Tim Roughgarden. "Algorithmic game theory". In: *Communications of the ACM* 53.7 (2010), pp. 78–86.

[355] Kaushik Roy and Claire J. Tomlin. "Solving the aircraft routing problem using network flow algorithms". In: *2007 American Control Conference*. 2007, pp. 3330–3335. DOI: `10.1109/ACC.2007.4282854`.

[356] Walter Rudin. *Principles Of Mathematical Analysis*. McGraw-Hill, Inc., 1976.

[357] Walter Rudin et al. *Principles of Mathematical Analysis*. Vol. 3. McGraw-hill New York, 1976.

[358] Yoan Russac, Claire Vernade, and Olivier Cappé. "Weighted linear bandits for non - stationary environments". In: *Advances in Neural Information Processing Systems* 32 (2019).

[359] Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, and Ian Osband. "A Tutorial on Thompson Sampling". In: abs/1707.02038 (2017).

[360] Dorsa Sadigh, Shankar Sastry, Sanjit A Seshia, and Anca D Dragan. "Planning for autonomous cars that leverage effects on human actions." In: *Robotics: Science and systems*. Vol. 2. Ann Arbor, MI, USA. 2016, pp. 1–9.

[361] William H Sandholm. *Population games and evolutionary dynamics*. MIT press, 2010.

[362] William H. Sandholm. *Population Games And Evolutionary Dynamics*. Economic Learning and Social Evolution, 2010.

[363] Abishek Sankararaman, Soumya Basu, and Karthik Abinav Sankararaman. "Dominate or Delete: Decentralized Competing Bandits in Serial Dictatorship". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 1252–1260.

[364] Abishek Sankararaman, Ayalvadi Ganesh, and Sanjay Shakkottai. "Social learning in multi agent multi armed bandits". In: *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 3.3 (2019), pp. 1–35.

[365] Shankar Sastry. *Nonlinear Systems: Analysis, Stability, and Control*. Springer, 1999.

[366] Shankar Sastry. *Nonlinear Systems: Analysis, Stability, and Control*. Vol. 10. Springer Science & Business Media, 2013.

[367] Muhammed Sayin, Kaiqing Zhang, David Leslie, Tamer Basar, and Asuman Ozdaglar. "Decentralized Q-learning in zero-sum Markov games". In: *Advances in Neural Information Processing Systems* 34 (2021).

[368] Muhammed O Sayin, Francesca Parise, and Asuman Ozdaglar. "Fictitious play in zero-sum stochastic games". In: *SIAM Journal on Control and Optimization* 60.4 (2022), pp. 2095–2114.

[369] Muhammed O Sayin, Kaiqing Zhang, and Asuman Ozdaglar. "Fictitious Play in Markov Games with Single Controller". In: *Proceedings of the 23rd ACM Conference on Economics and Computation*. 2022, pp. 919–936.

[370] AE Scheidegger. "Stability of motion, by nn krasovskii. translated from the (1959) russian ed. by jl brenner. stanford university press, 1963." In: *Canadian Mathematical Bulletin* 7.1 (1964), pp. 151–151.

[371] Alexander Schrijver. *Theory of Linear and Integer Programming*. USA: John Wiley & Sons, Inc., 1998. ISBN: 0471908541.

[372] Wilko Schwarting, Alyssa Pierson, Sertac Karaman, and Daniela Rus. "Stochastic dynamic games in belief space". In: *IEEE Transactions on Robotics* 37.6 (2021), pp. 2157–2172.

[373] Pier Giuseppe Sessa, Ilija Bogunovic, M. Kamgarpour, and A. Krause. "Learning to play sequential games versus unknown opponents". In: *ArXiv* abs/2007.05271 (2020).

[374] Pier Giuseppe Sessa, Ilija Bogunovic, Maryam Kamgarpour, and Andreas Krause. "Learning to play sequential games versus unknown opponents". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 8971–8981.

[375] Sven Seuken, Paul Friedrich, and Ludwig Dierks. "Market Design for Drone Traffic Management". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 11. 2022, pp. 12294–12300.

[376] Amirreza Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. "Truncated back-propagation for bilevel optimization". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 1723–1732.

[377] Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and D. Kuhn. "Distributionally robust logistic regression". In: *NeurIPS*. 2015.

[378] Mehran Shakarami, Ashish Cherukuri, and Nima Monshizadeh. "Dynamic interventions with limited knowledge in network games". In: *IEEE Transactions on Control of Network Systems* (2023).

[379] Mehran Shakarami, Ashish Cherukuri, and Nima Monshizadeh. "Steering the aggregative behavior of noncooperative agents: a nudge framework". In: *Automatica* 136 (2022), p. 110003.

[380] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. "Safe, multi-agent, reinforcement learning for autonomous driving". In: *arXiv preprint arXiv:1610.03295* (2016).

[381] Quan Shao, Mengxue Shao, and Yang Lu. "Terminal area control rules and eVTOL adaptive scheduling model for multi-vertiport system in urban air Mobility". In: *Transportation Research Part C: Emerging Technologies* 132 (2021), p. 103385. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103385. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21003843.

[382] Lloyd S Shapley. "Stochastic games". In: *Proceedings of the National Academy of Sciences* 39.10 (1953), pp. 1095–1100.

[383] Yosef Sheffi and Warren Powell. "A Comparison of Stochastic and Deterministic Traffic Assignment over Congested Networks". In: *Transportation Research Part B: Methodological* 15.1 (1981), pp. 53–64.

[384] Aman Sinha, Matthew O'Kelly, Hongrui Zheng, Rahul Mangharam, John Duchi, and Russ Tedrake. "Formulazero: Distributionally robust online adaptation via offline population synthesis". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 8992–9004.

[385] Brent Skorup. "Auctioning airspace". In: *NCJL & Tech.* 21 (2019), p. 79.

[386] Aleksandrs Slivkins et al. "Introduction to multi-armed bandits". In: *Foundations and Trends® in Machine Learning* 12.1-2 (2019), pp. 1–286.

[387] Kenneth A Small. "Using the revenues from congestion pricing". In: *Transportation* 19 (1992), pp. 359–381.

[388] David Roger Smart. *Fixed point theorems*. Vol. 66. Cup Archive, 1980.

[389] MJ Smith. "The marginal cost taxation of a transportation network". In: *Transportation Research Part B: Methodological* 13.3 (1979), pp. 237–242.

[390] Yunlong Song, Mats Steinweg, Elia Kaufmann, and Davide Scaramuzza. "Autonomous drone racing with deep reinforcement learning". In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2021, pp. 1205–1212.

[391] Ziang Song, Song Mei, and Yu Bai. "When Can We Learn General-Sum Markov Games with a Large Number of Players Sample-Efficiently?" In: *International Conference on Learning Representations*. 2021.

[392] Ziqi Song, Yafeng Yin, and Siriphong Lawphongpanich. "Nonnegative Pareto - improving tolls with multiclass network equilibria". In: *Transportation Research Record* 2091.1 (2009), pp. 70–78.

[393] Daouda Sow, Kaiyi Ji, Ziwei Guan, and Yingbin Liang. "A constrained optimization approach to bilevel optimization with multiple inner minima". In: *arXiv preprint arXiv:2203.01123* (2022).

[394] J. Spall. "A one-measurement form of simultaneous perturbation stochastic approximation". In: *Autom.* 33 (1997), pp. 109–112.

[395] James C Spall. "A one-measurement form of simultaneous perturbation stochastic approximation". In: *Automatica* 33.1 (1997), pp. 109–112.

[396] Riccardo Spica, Eric Cristofalo, Zijian Wang, Eduardo Montijano, and Mac Schwager. "A real-time game theoretic planner for autonomous two-player drone racing". In: *IEEE Transactions on Robotics* 36.5 (2020), pp. 1389–1403.

[397] Pan-Yang Su, Chinmay Maheshwari, Victoria Marie Tuck, and Shankar Sastry. "Incentive compatible vertiport reservation in advanced air mobility: An auction-based approach". In: *2024 IEEE 63rd Conference on Decision and Control (CDC)*. IEEE. 2024, pp. 7720–7727.

[398] Lingfeng Sun, Pin-Yun Hung, Changhao Wang, Masayoshi Tomizuka, and Zhuo Xu. "Distributed Multi-agent Interaction Generation with Imagined Potential Games". In: *arXiv preprint arXiv:2310.01614* (2023).

[399] Luying Sun, Peng Wei, and Weijun Xie. "Fair and Risk-Averse Urban Air Mobility Resource Allocation Under Uncertainties". In: *Available at SSRN 4343979* (2023).

[400] Youbang Sun, Tao Liu, Ruida Zhou, PR Kumar, and Shahin Shahrampour. "Provably fast convergence of independent natural policy gradient for Markov Potential Games". In: *Advances in Neural Information Processing Systems* 36 (2023).

[401] Brian Swenson, Ryan Murray, and Soummya Kar. "On best-response dynamics in potential games". In: *SIAM Journal on Control and Optimization* 56.4 (2018), pp. 2734–2767.

[402] Vladislav B Tadić, Sean P Meyn, and Roberto Tempo. "Randomized algorithms for semi-infinite programming problems". In: *2003 European Control Conference (ECC)*. IEEE. 2003, pp. 3011–3015.

[403] Ming Tan. "Multi-agent reinforcement learning: Independent vs. cooperative agents". In: *Proceedings of the tenth international conference on machine learning*. 1993.

[404] Rishabh Saumil Thakkar, Aryaman Singh Samyal, David Fridovich-Keil, Zhe Xu, and Ufuk Topcu. "Hierarchical control for cooperative teams in competitive autonomous racing". In: *IEEE Transactions on Intelligent Vehicles* (2024).

[405] Rishabh Saumil Thakkar, Aryaman Singh Samyal, David Fridovich-Keil, Zhe Xu, and Ufuk Topcu. "Hierarchical control for head-to-head autonomous racing". In: *arXiv preprint arXiv:2202.12861* (2022).

[406] Thomas C Thomas and Gordon I Thompson. "The value of time for commuting motorists as a function of their income level and amount of time saved". In: *Highway Research Record* (1970).

[407] William R. Thompson. "On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples". In: *Biometrika* 25.3/4 (1933), pp. 285–294.

[408] John N Tsitsiklis. "Asynchronous stochastic approximation and Q-learning". In: *Machine Learning* 16.3 (1994), pp. 185–202.

[409] Chun-Chen Tu, Pai-Shun Ting, P. Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. "AutoZOOM: autoencoder-based zeroth order optimization method for attacking black-box neural networks". In: *AAAI*. 2019.

[410] Sergio Valcarcel Macua, Javier Zazo, and Santiago Zazo. "Learning Parametric Closed-Loop Policies for Markov Potential Games". In: *6th International Conference on Learning Representations (ICLR)*. 2018.

[411] Hal R Varian. "Equity, envy, and efficiency". In: *Journal of Economic Theory* 9.1 (1974), pp. 63–91. ISSN: 0022-0531. DOI: https://doi.org/10.1016/0022-0531(74)90075-1. URL: https://www.sciencedirect.com/science/article/pii/0022053174900751.

[412] Erik T Verhoef. "Second-best congestion pricing in general networks. Heuristic algorithms for finding second-best optimal toll levels and toll points". In: *Transportation Research Part B: Methodological* 36.8 (2002), pp. 707–729.

[413] Richard W. Vesel. "Racing line optimization @ race optimal". In: *ACM SIGEVOlution* 7 (2015), pp. 12–20.

[414] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georg iev, et al. "Grandmaster level in StarCraft II using multi-agent reinforcement learning". In: *Nature* 575.7782 (2019), pp. 350–354.

[415] Ben Wang, Zilong Deng, Xuan Ni, Kevin B Smith, Max Z Li, and Romesh Saigal. "Learning-Driven Airspace Congestion Pricing for Advanced Air Mobility". In: *AIAA SCITECH 2023 Forum*. 2023, p. 0547.

[416] Mingyu Wang, Negar Mehr, Adrien Gaidon, and Mac Schwager. "Game-theoretic planning for risk-aware interactive agents". In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2020, pp. 6998–7005.

[417] Mingyu Wang, Zijian Wang, John Talbot, J Christian Gerdes, and Mac Schwager. "Game Theoretic Planning for Self-Driving Cars in Competitive Scenarios." In: *Robotics: Science and Systems*. 2019, pp. 1–9.

[418] Mingyu Wang, Zijian Wang, John Talbot, J Christian Gerdes, and Mac Schwager. "Game-theoretic planning for self-driving cars in multivehicle competitive scenarios". In: *IEEE Transactions on Robotics* 37.4 (2021), pp. 1313–1325.

[419] Z. Wang, K. Balasubramanian, Shiqian Ma, and Meisam Razaviyayn. "Zeroth-order algorithms for nonconvex minimax problems with improved complexities". In: *ArXiv* abs/2001.07819 (2020).

[420] Zijian Wang, Riccardo Spica, and Mac Schwager. "Game theoretic motion planning for multi-robot racing". In: *Distributed Autonomous Robotic Systems: The 14th International Symposium*. Springer. 2019, pp. 225–238.

[421] Zijian Wang, Tim Taubner, and Mac Schwager. "Multi-agent sensitivity enhanced iterative best response: A real-time game theoretic planner for drone racing in 3D environments". In: *Robotics Auton. Syst.* 125 (2020), p. 103410. URL: `https://api.semanticscholar.org/CorpusID:211050763`.

[422] John Glen Wardrop. "Some Theoretical Aspects of Road Traffic Research." In: *Proceedings of the institution of civil engineers* 1.3 (1952), pp. 325–362.

[423] William G Waters. "The value of travel time savings and the link with income: implications for public project evaluation". In: *International Journal of Transport Economics/Rivista internazionale di economia dei trasporti* (1994), pp. 243–253.

[424] Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. "Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive Markov games". In: *Conference on Learning Theory*. PMLR. 2021, pp. 4259–4299.

[425] Chen-Yu Wei and Haipeng Luo. "More adaptive algorithms for adversarial bandits". In: *Conference On Learning Theory*. PMLR. 2018, pp. 1263–1291.

[426] Peng Wei, Paul Krois, Joseph Block, Paul Cobb, Gano Chatterji, and Cherie Kurian. "Arrival Management for High-density Vertiport and Terminal Airspace Operations". In: (2023).

[427] Richard Wheeler and Kumpati Narendra. "Decentralized learning in finite Markov chains". In: *IEEE Transactions on Automatic Control* 31.6 (1986), pp. 519–526.

[428] Grady Williams, Brian Goldfain, Paul Drews, James M Rehg, and Evangelos A Theodorou. "Autonomous racing with autorally vehicles and differential games". In: *arXiv preprint arXiv:1707.04540* (2017).

[429] Salomón Wollenstein-Betech, Chuangchuang Sun, Jing Zhang, and Ioannis Ch Paschalidis. "Joint estimation of od demands and cost functions in transportation networks from data". In: *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE. 2019, pp. 5113–5118.

[430] Stephen J Wright and Benjamin Recht. *Optimization for data analysis*. Cambridge University Press, 2022.

[431] Di Wu, Yafeng Yin, Siriphong Lawphongpanich, and Hai Yang. "Design of more equitable congestion pricing and tradable credit schemes for multimodal transportation networks". In: *Transportation Research Part B: Methodological* 46.9 (2012), pp. 1273–1287.

[432] Pengcheng Wu, Xuxi Yang, Peng Wei, and Jun Chen. "Safety Assured Online Guidance With Airborne Separation for Urban Air Mobility Operations in Uncertain Environments". In: *IEEE Transactions on Intelligent Transportation Systems* 23.10 (2022), pp. 19413–19427. DOI: 10.1109/TITS.2022.3163657.

[433] Peter R Wurman, Samuel Barrett, Kenta Kawamoto, James MacGlashan, Kaushik Subramanian, Thomas J Walsh, Roberto Capobianco, Alisa Devlic, Franziska Eckert, Florian Fuchs, et al. "Outracing champion Gran Turismo drivers with deep reinforcement learning". In: *Nature* 602.7896 (2022), pp. 223–228.

[434] Tetsuo Yai, Seiji Iwakura, and Shigeru Morichi. "Multinomial Probit with Structured Covariance for Route Choice Behavior". In: *Transportation Research Part B: Methodological* 31.3 (1997), pp. 195–207.

[435] Hai Yang and Hai-Jun Huang. *Mathematical and Economic Theory of Road Pricing*. Emerald Group Publishing Limited, 2005.

[436] Hai Yang and William HK Lam. "Optimal road tolls under conditions of queueing and congestion". In: *Transportation Research Part A: Policy and Practice* 30.5 (1996), pp. 319–332.

[437] Mao Ye, Bo Liu, Stephen Wright, Peter Stone, and Qiang Liu. "Bome! bilevel optimization made easy: A simple first-order approach". In: *arXiv preprint arXiv: 2209. 08709* (2022).

[438] Bora Yongacoglu, Gürdal Arslan, and Serdar Yüksel. "Asynchronous decentralized Q-learning: Two timescale analysis by persistence". In: *arXiv preprint arXiv:2308.03239* (2023).

[439] Bora Yongacoglu, Gürdal Arslan, and Serdar Yüksel. "Decentralized learning for optimality in stochastic dynamic teams and games with local control and global state information". In: *IEEE Transactions on Automatic Control* 67.10 (2021), pp. 5230–5245.

[440] Bora Yongacoglu, Gürdal Arslan, and Serdar Yüksel. "Satisficing paths and independent multiagent reinforcement learning in stochastic games". In: *SIAM Journal on Mathematics of Data Science* 5.3 (2023), pp. 745–773.

[441] Hyejin Youn, Michael T Gastner, and Hawoong Jeong. "Price of anarchy in transportation networks: efficiency and optimality control". In: *Physical review letters* 101.12 (2008), p. 128701.

[442] Yaodong Yu, Tianyi Lin, Eric V. Mazumdar, and Michael I. Jordan. "Fast distributionally robust learning with variance reduced min-max optimization". In: *ArXiv* abs/2104.13326 (2021).

[443] Dajun Yue and Fengqi You. "Stackelberg-game-based modeling and optimization for supply chain design and operations: A mixed integer bilevel programming framework". In: *Computers & Chemical Engineering* 102 (2017), pp. 81–95.

[444] Brian Hu Zhang, Gabriele Farina, Ioannis Anagnostides, Federico Cacciamani, Stephen Marcus McAleer, Andreas Alexander Haupt, Andrea Celli, Nicola Gatti, Vincent Conitzer, and Tuomas Sandholm. "Steering No-Regret Learners to Optimal Equilibria". In: *arXiv preprint arXiv:2306.05221* (2023).

[445] Jing Zhang, Sepideh Pourazarm, Christos G Cassandras, and Ioannis Ch Paschalidis. "The price of anarchy in transportation networks by estimating user cost functions from actual traffic data". In: *2016 IEEE 55th conference on decision and control (cdc)*. IEEE. 2016, pp. 789–794.

[446] Kai Zhang and Stuart Batterman. "Air pollution and health risks due to vehicle traffic". In: *Science of the total Environment* 450 (2013), pp. 307–316.

[447] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. "Multi-agent reinforcement learning: A selective overview of theories and algorithms". In: *Handbook of reinforcement learning and control* (2021), pp. 321–384.

[448] Lin Zhang, Haining Du, and Linda Lee. "Congestion Pricing Study of San Francisco–Oakland Bay Bridge in California". In: *Transportation research record* 2221.1 (2011), pp. 83–95.

[449] Runyu Zhang, Jincheng Mei, Bo Dai, Dale Schuurmans, and Na Li. "On the Global Convergence Rates of Decentralized Softmax Gradient Play in Markov Potential Games". In: *Advances in Neural Information Processing Systems*. 2022.

[450] Runyu Zhang, Zhaolin Ren, and Na Li. "Gradient play in stochastic games: stationary points, convergence, and sample complexity". In: *arXiv preprint arXiv:2106.00198* (2021).

[451] Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou. "A simple approach for non - stationary linear bandits". In: (2020), pp. 746–755.

[452] Hongrui Zheng, Zhijun Zhuang, Johannes Betz, and Rahul Mangharam. "Game-theoretic objective space planning". In: *arXiv preprint arXiv:2209.07758* (2022).

[453] Dao-Li Zhu, Hai Yang, Chang-Min Li, and Xiao-Lei Wang. "Properties of the multiclass traffic network equilibria under a tradable credit scheme". In: *Transportation Science* 49.3 (2015), pp. 519–534.

[454] Edward L Zhu and Francesco Borrelli. "A sequential quadratic programming approach to the solution of open-loop generalized nash equilibria". In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2023, pp. 3211–3217.

[455] Edward L Zhu and Francesco Borrelli. "A Sequential Quadratic Programming Approach to the Solution of Open-Loop Generalized Nash Equilibria for Autonomous Racing". In: *arXiv preprint arXiv:2404.00186* (2024).

[456] Maëlle Zimmermann and Emma Frejinger. "A Tutorial on Recursive Models for Analyzing and Predicting Path Choice Behavior". In: *EURO Journal on Transportation and Logistics* 9.2 (2020), p. 100004.

Appendix A

Appendix for Chapter 2

Here, we provide additional material supplementing the content of Chapter 2.

A.1 Proofs in Section 2.3

Proof of Proposition 2.3.1

Let Φ be a potential function of MPG $\mathcal{G}$. Using Definition 2.3.1, it suffices to show $\Phi \in \mathcal{F}^{\mathcal{G}}$. First, we claim that for every $s \in S, \pi, \pi' \in \Pi$,

$$|\Phi(s,\pi) - \Phi(s,\pi')| \leqslant \sum_{i=1}^{N} |V_i(s,\tilde{\pi}^{(i)}) - V_i(s,\tilde{\pi}^{(i+1)})|, \tag{A.1}$$

where for any $i \in I$, $\tilde{\pi}^{(i)} = (\pi'_1, \pi'_2, ..\pi'_{i-1}, \pi_i, \pi_{i+1}, ..., \pi_N)$ with the understanding that $\tilde{\pi}^{(1)} = \pi$ and $\tilde{\pi}^{(N+1)} = \pi'$. To prove this claim, note that

$$\begin{aligned} |\Phi(s,\pi) - \Phi(s,\pi')| &= \left| \sum_{i=1}^{N} \Phi(s,\tilde{\pi}^{(i)}) - \Phi(s,\tilde{\pi}^{(i+1)}) \right| \\ &\leqslant \sum_{i=1}^{N} \left| V_i(s,\tilde{\pi}^{(i)}) - V_i(s,\tilde{\pi}^{(i+1)}) \right|, \end{aligned}$$

which follows from Definition 2.3.1 as $\tilde{\pi}^{(i)}$ and $\tilde{\pi}^{(i+1)}$ only differ at player i's policy. By (A.1), for any $s \in S$, $\pi, \pi' \in \Pi$,

$$|\Phi(s,\pi) - \Phi(s,\pi')| \leqslant 2N \max_{i \in I} \|V_i\|_\infty \leqslant \frac{2N}{1-\gamma} \max_{i \in I} \|u_i\|_\infty.$$

We note without loss of generality, $\min_{\pi \in \Pi} \Phi(s,\pi) = 0$ for every $s \in S$ (cf. Definition 2.3.1). Therefore,

$$\|\Phi\|_\infty \leqslant \frac{2N}{1-\gamma} \max_{i \in I} \|u_i\|_\infty.$$

To show that $\mathbf{\Phi}$ lies in a uniformly equicontinuous set $\mathcal{F}_{\mathcal{G}}$, we next show that $\mathbf{\Phi}$ is uniformly continuous. Note that for each $s \in S$ and $i \in I_N$, $V_i(s,\cdot) : \Pi \to \mathbb{R}$ is a continuous function [440, Lemma 2.10]. Given that Π is compact and $|S| < \infty$, for every $\epsilon > 0$ there exists $\bar{\delta}(\epsilon) > 0$ such that $\max\limits_{i \in I_N, s \in S} |V_i(s,\pi) - V_i(s,\pi')| \leqslant \epsilon/N$ for any $\pi, \pi' \in \Pi$ satisfying $\mathbf{d}(\pi,\pi') \leqslant \bar{\delta}(\epsilon)$. Consequently, from (A.1), we conclude that for any $\epsilon > 0$, $|\mathbf{\Phi}(s,\pi) - \mathbf{\Phi}(s,\pi')| \leqslant \epsilon$ for any $\pi, \pi' \in \Pi$ satisfying $\mathbf{d}(\pi,\pi') \leqslant \bar{\delta}(\epsilon)$.

Proof of Proposition 2.3.2

The proof of Proposition 2.3.2 relies on the following lemma.

Lemma A.1.1. *If there exists some $\zeta > 0$ such that for all $s, s' \in S$, $|P(s'|s,w) - P(s'|s,w')| \leqslant \zeta \|w - w'\|_1$. Then for any $i \in I, \pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$,*

$$\|P^{\pi_i,\pi_{-i}} - P^{\pi_i',\pi_{-i}}\|_\infty \leqslant 2\zeta |S| \max_{a_i \in A_i} |a_i|/N. \tag{A.2}$$

Proof. For any $i \in I$, $\pi \in \Pi$, $\pi_i' \in \Pi_i$, and $s, s' \in S$,

$$\begin{aligned}
&P^{\pi_i,\pi_{-i}}(s'|s) - P^{\pi_i',\pi_{-i}}(s'|s) \\
&= \mathbb{E}_{\substack{a_{-i} \sim \pi_{-i} \\ a_i \sim \pi_i}} \Big[P(s'|s, w(a_i,a_{-i})) - P(s'|s, w(a_i,a_{-i})) \Big] \\
&\leqslant \mathbb{E}_{a_{-i} \sim \pi_{-i}} \Big[P(s'|s, w(\bar{a}_i,a_{-i})) - P(s'|s, w(\underline{a}_i,a_{-i})) \Big],
\end{aligned} \tag{A.3}$$

where the first equation is due to the structure of transition function,

$$\bar{a}_i \in \arg\max\nolimits_{a_i \in A_i} P(s'|s, w(a_i,a_{-i})),$$

and

$$\underline{a}_i \in \arg\min\nolimits_{a_i \in A_i} P(s'|s, w(a_i,a_{-i})).$$

By (A.3) and the Lipschitz property of the transition matrix in Lemma A.1.1,

$$\begin{aligned}
&\sum_{s' \in S} |P^{\pi_i,\pi_{-i}}(s'|s) - P^{\pi_i',\pi_{-i}}(s'|s)| \\
&\overset{(a)}{=} \frac{\zeta |S|}{N} \mathop{\mathbb{E}}_{a_{-i} \sim \pi_{-i}} \left[\sum_{e \in E} |\mathbb{1}(e \in \bar{a}_i) - \mathbb{1}(e \in \underline{a}_i)| \right] \\
&= \frac{2\zeta |S| \max_{a_i \in A_i} |a_i|}{N}, \quad \forall\ s \in S,
\end{aligned}$$

where (a) follows by (2.6). □

Proof of Proposition 2.3.2

Recall that for any $s \in S$, the stage game is a potential game with a potential function

$$\varphi(s,a) = 1/N \sum_{e\in E} \sum_{j=1}^{w_e(a)N} c_e\left(s, j/N\right).$$

Under this notation, we can equivalently write (2.7) as

$$\Psi(s,\pi) = \varphi(s,\pi) + \gamma \sum_{s'\in S} P^{\pi}(s'|s)\Psi(s',\pi). \tag{A.4}$$

For the rest of the proof, fix arbitrary $\pi_i, \pi_i' \in \Pi_i, \pi_{-i} \in \Pi_{-i}$ and denote $\pi = (\pi_i, \pi_{-i})$, $\pi' = (\pi_i', \pi_{-i})$. By (A.4),

$$\Psi(s,\pi) - \Psi(s,\pi') = \varphi(s,\pi) - \varphi(s,\pi') + \gamma \sum_{s'\in S} \left(P^{\pi}(s'|s)\Psi(s',\pi) - P^{\pi'}(s'|s)\Psi(s',\pi')\right). \tag{A.5}$$

Additionally, recall that $V_i(s,\pi) = u_i(s,\pi) + \gamma \sum_{s'\in S} P^{\pi}(s'|s)V_i(s',\pi)$. Consequently,

$$V_i(s,\pi) - V_i(s,\pi') = u_i(s,\pi) - u_i(s,\pi') + \gamma \sum_{s'\in S} \left(P^{\pi}(s'|s)V_i(s',\pi) - P^{\pi'}(s'|s)V_i(s',\pi')\right). \tag{A.6}$$

Subtracting (A.5) from (A.6), we obtain

$$\begin{aligned}
&V_i(s,\pi) - V_i(s,\pi') - \left(\Psi(s,\pi) - \Psi(s,\pi')\right) \\
=&\gamma \sum_{s'\in S} P^{\pi}(s'|s)\left(V_i(s',\pi) - \Psi(s',\pi)\right) - \gamma \sum_{s'\in S} P^{\pi'}(s'|s)\left(V_i(s',\pi') - \Psi(s',\pi')\right) \\
=&\gamma \sum_{s'\in S} P^{\pi}(s'|s)\left(V_i(s',\pi) - V_i(s',\pi') + \Psi(s',\pi') - \Psi(s',\pi)\right) \\
&- \gamma \sum_{s'\in S} \left(P^{\pi'}(s'|s) - P^{\pi}(s'|s)\right)\left(V_i(s',\pi') - \Psi(s',\pi')\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\max_{s\in S} |V_i(s,\pi) - V_i(s,\pi') - (\Psi(s,\pi) - \Psi(s,\pi'))| \\
&\leqslant \gamma \max_{s\in S} |V_i(s,\pi) - V_i(s,\pi') - \left(\Psi(s,\pi) - \Psi(s,\pi')\right)| \\
&+ \gamma \max_{s'\in S} |\Psi(s',\pi) - V_i(s',\pi)| \max_{s\in S} \sum_{s'\in S} \left|P^{\pi}(s'|s) - P^{\pi'}(s'|s)\right|.
\end{aligned} \tag{A.7}$$

Rearranging terms leads to

$$\begin{aligned}(\text{A.7}) &\leqslant \frac{\gamma}{1-\gamma}\max_{s'\in S}|\Psi(s',\pi)-V_i(s',\pi)|\|P^{\pi}-P^{\pi'}\|_{\infty}\\ &\leqslant \frac{2\gamma\zeta|S|\max_{a_i\in A_i}|a_i|}{(1-\gamma)N}\max_{s'\in S}|\Psi(s',\pi)-V_i(s',\pi)|. \end{aligned} \tag{A.8}$$

where the last inequality follows from Lemma A.1.1. Finally, since

$$\begin{aligned}u_i(s^k,a^k) &= \sum_{e\in E} c_e(s^k,w_e^k)\mathbb{1}(e\in a_i^k) \leqslant \sum_{e\in E} c_e(s^k,w_e^k)\\ &\leqslant \varphi(s^k,a^k),\end{aligned}$$

then for any $s'\in S$,

$$\begin{aligned}|\Psi(s',\pi)-V_i(s',\pi)| &\leqslant \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k\left|\varphi(s^k,a^k)-u_i(s^k,a^k)\right|\right]\\ \leqslant \left|\mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k\varphi(s^k,a^k)\right]\right| &\leqslant \sup_{s,\pi}\Psi(s,\pi).\end{aligned}$$

Plugging the above inequality into (A.8) finishes the proof.

Proof of Proposition 2.3.3

Throughout the proof, let us fix arbitrary $i\in I,\pi_i,\pi_i'\in\Pi_i,\pi_{-i}\in\Pi_{-i}$, and define $\pi=(\pi_i,\pi_{-i}),\pi'=(\pi_i',\pi_{-i})$. We show that for every $i\in I,\pi_i,\pi_i'\in\Pi_i,\pi_{-i}\in\Pi_{-i}$,

$$\max_{s\in S}|V_i(s,\pi)-V_i(s,\pi')-(\Psi(s,\pi)-\Psi(s,\pi'))|\leqslant\frac{2\kappa}{(1-\gamma)^2},$$

where $\Psi(s,\pi):=\mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty}\gamma^k r(s^k,a^k)|s^0=s\right]$. Note that

$$\Psi(s,\pi)=r(s,\pi)+\gamma\sum_{s'\in S}P^{\pi}(s'|s)\Psi(s',\pi). \tag{A.9}$$

By (A.9), for any $s\in S$,

$$\begin{aligned}\Psi(s,\pi)-\Psi(s,\pi') &= r(s,\pi)-r(s,\pi')\\ &+\gamma\sum_{s'\in S}\left(P^{\pi}(s'|s)\Psi(s',\pi)-P^{\pi'}(s'|s)\Psi(s',\pi')\right).\end{aligned} \tag{A.10}$$

Similarly, for any $s\in S$,

$$\begin{aligned}V_i(s,\pi)-V_i(s,\pi') &= u_i(s,\pi)-u_i(s,\pi')\\ &+\gamma\sum_{s'\in S}P^{\pi}(s'|s)V_i(s',\pi)-P^{\pi'}(s'|s)V_i(s',\pi').\end{aligned} \tag{A.11}$$

Consequently,

$$\begin{aligned}
&V_i(s,\pi) - V_i(s,\pi') - \Big(\Psi(s,\pi) - \Psi(s,\pi')\Big) \\
=&u_i(s,\pi) - u_i(s,\pi') - \Big(r(s,\pi) - r(s,\pi')\Big) \\
&- \gamma \sum_{s'\in S} \Big(P^{\pi'}(s'|s) - P^{\pi}(s'|s)\Big)\Big(V_i(s',\pi') - \Psi(s',\pi')\Big) \\
&+ \gamma \sum_{s'\in S} P^{\pi}(s'|s)\Big(V_i(s',\pi) - V_i(s',\pi') + \Psi(s',\pi') - \Psi(s',\pi)\Big).
\end{aligned}$$

Since $|u_i(s,\pi) - u_i(s,\pi') - (r(s,\pi) - r(s,\pi'))| \leqslant 2\|\xi_i\|_\infty \leqslant 2\kappa$, then

$$\begin{aligned}
&\max_{s\in S} |V_i(s,\pi) - V_i(s,\pi') - \Big(\Psi(s,\pi) - \Psi(s,\pi')\Big)| \qquad \text{(A.12)} \\
\leqslant& 2\kappa + 2\gamma \max_{s'\in S} |\Psi(s',\pi) - V_i(s',\pi)| \\
&+ \gamma \max_{s\in S} |V_i(s,\pi) - V_i(s,\pi') - \Big(\Psi(s,\pi) - \Psi(s,\pi')\Big)|.
\end{aligned}$$

Rearranging terms in above inequality, we obtain

$$\text{(A.12)} \leqslant \frac{2\kappa}{1-\gamma} + \frac{2\gamma}{1-\gamma} \max_{s'\in S} |\Psi(s',\pi) - V_i(s',\pi)|. \qquad \text{(A.13)}$$

Finally, note that $|\Psi(s',\pi) - V_i(s',\pi)| = \left|\sum_{k=0}^{\infty} \gamma^k \xi_i(s,\pi(s^k))\right| \leqslant \kappa/(1-\gamma)$. Plugging this inequality into (A.13) completes the proof.

A.2 Proofs in Section 2.4

Proof of Proposition 2.4.1

To prove Proposition 2.4.1, we first need the following lemma.

Lemma A.2.1 (Lemma B.1 in [440]). *Fix $i \in I$ and $K \in \mathbb{N}$. For any $s \in S$ and $\omega = \left(\tilde{s}^k, \tilde{a}^k\right)_{k=0}^{K} \in (S\times A)^{K+1}$, the mapping $\Pi \ni \pi \mapsto \mathbb{E}_\pi\left[\mathbb{1}\left((s^k,a^k)_{k=0}^K = \omega\right) \mid s^0 = s\right]$ is continuous.*

Proof of Proposition 2.4.1

Fix $\epsilon > 0$ and define $M := N \max_{i\in I} \|u_i\|_\infty$. Choose $K \in \mathbb{N}$ large enough that $\frac{\gamma^K \cdot M}{1-\gamma} < \frac{\epsilon}{4}$ and $\tilde{\epsilon} := \frac{(1-\gamma)\epsilon}{2M|S|^{K+1}|A|^{K+1}}$. Since Π is compact and $S\times A$ is finite, Lemma A.2.1 ensures that there exists $\delta(\epsilon)$ such that for any $\pi,\pi' \in \Pi$ with $\mathbf{d}(\pi,\pi') \leqslant \delta(\epsilon)$, and $\omega \in (S\times A)^{K+1}, s \in S$,

$$\left|\mathbb{E}_\pi\left[\mathbb{1}\left((s^k,a^k)_{k=0}^K = \omega\right) \mid s^0 = s\right] - \mathbb{E}_{\pi'}\left[\mathbb{1}\left((s^k,a^k)_{k=0}^K = \omega\right) \mid s^0 = s\right]\right| \leqslant \tilde{\epsilon}. \qquad \text{(A.14)}$$

From (2.8), we note that for any $\Psi \in \tilde{\mathcal{F}}^{\mathcal{G}}$, there exists $\phi : S \times A \to \mathbb{R}$ such that for any $\pi, \pi' \in \Pi, s \in S$,

$$\begin{aligned}&\left|\Psi(s,\pi) - \Psi(s,\pi')\right| \\ &\leqslant \left|\mathbb{E}_{\pi}\left[\sum_{k=0}^{K} \gamma^k \phi\left(s^k, a^k\right) \mid s_0 = s\right] - \mathbb{E}_{\pi'}\left[\sum_{k=0}^{K} \gamma^k \phi\left(s^k, a^k\right) \mid s_0 = s\right]\right| + \frac{\epsilon}{2}. \end{aligned} \tag{A.15}$$

Define a function $\varphi : (S \times A)^{K+1} \to \mathbb{R}$ such that for every $\left(\tilde{s}^k, \tilde{a}^k\right)_{k=0}^{K} \in (S \times A)^{K+1}$,

$$\varphi\left(\tilde{s}^0, \tilde{a}^0, \cdots, \tilde{s}^K, \tilde{a}^K\right) := \sum_{k=0}^{K} \gamma^k \phi\left(\tilde{s}^k, \tilde{a}^k\right).$$

Thus, for any $\pi \in \Pi$,

$$\begin{aligned}&\mathbb{E}_{\pi}\left[\sum_{k=0}^{K} \gamma^k \phi\left(s^k, a^k\right) \mid s^0 = s\right] \\ &= \sum_{\omega \in (S\times A)^{K+1}} \varphi(\omega) \mathbb{E}_{\pi}\left[\mathbb{1}\left(\left(s^k, a^k\right)_{t=0}^{K} = \omega\right) \middle| s^0 = s\right],\end{aligned}$$

Thus, by applying the above equation and (A.14) to (A.15), we obtain that for any $s \in S, \pi, \pi' \in \Pi$ satisfying $\mathbf{d}(\pi, \pi') \leqslant \delta(\epsilon)$,

$$\begin{aligned}\left|\Psi(s,\pi) - \Psi(s,\pi')\right| &\leqslant \|\varphi\|_{\infty} |S|^{K+1} |A|^{K+1} \tilde{\epsilon} + \frac{\epsilon}{2} \\ &\leqslant \frac{M|S|^{K+1}|A|^{K+1}\tilde{\epsilon}}{1-\gamma} + \frac{\epsilon}{2} \leqslant \epsilon.\end{aligned}$$

Since we chose arbitrary $\Psi \in \tilde{\mathcal{F}}^{\mathcal{G}}$, and δ is independent of the choice of Ψ, then $\tilde{\mathcal{F}}^{\mathcal{G}}$ is equi-continuous. Thus, $\tilde{\mathcal{F}}^{\mathcal{G}} \subseteq \mathcal{F}^{\mathcal{G}}$.

Proof of Lemma 2.5.2

To prove Lemma 2.5.2, we define $\pi_{i\sim j} \coloneqq \{\pi_k\}_{k=i+1}^{j-1}$ as the joint policy for players from $i+1$ to $j-1$; $\pi_{<i} \coloneqq \{\pi_k\}_{k=1}^{i-1}$, and $\pi_{>j} \coloneqq \{\pi_k\}_{k=j+1}^{N}$ are defined similarly. Next, we recall a useful result from [121].

Lemma A.2.2 (Lemma 2 in [121]). *For any function $f : \Pi \to \mathbb{R}$, and any two policies $\pi, \pi' \in \Pi$,*

$$\begin{aligned} f(\pi') - f(\pi) &= \sum_{i=1}^{N} \Big(f(\pi_i', \pi_{-i}) - f(\pi) \Big) \\ &+ \sum_{i=1}^{N} \sum_{j=i+1}^{N} \Big(f(\pi_{<i, i\sim j}, \pi'_{>j}, \pi_i', \pi_j') - f(\pi_{<i, i\sim j}, \pi'_{>j}, \pi_i, \pi_j') \\ &- f(\pi_{<i, i\sim j}, \pi'_{>j}, \pi_i', \pi_j) + f(\pi_{<i, i\sim j}, \pi'_{>j}, \pi_i, \pi_j) \Big). \end{aligned} \tag{A.16}$$

Next, we state a result that lower bounds the improvement in value function of each player in each step of Algorithm 1.

Lemma A.2.3. *Consider a Markov game $\mathcal{G}$ with initial state distribution ν, let $\pi^{(t+1)}$ and $\pi^{(t)}$ be consecutive policies in Algorithm 1. Then we have,*

$$\begin{aligned} &(i)\, V_i(\nu, \pi^{(t+1)}) - V_i(\nu, \pi^{(t)}) \geqslant -\frac{4\eta^2 \bar{A}^2 N^2}{(1-\gamma)^5} + \frac{1}{2\eta(1-\gamma)} \\ &\quad \cdot \sum_{i \in I, s \in S} d_\nu^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s) \left\| \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\|^2; \end{aligned}$$

$$\begin{aligned} &(ii)\, V_i(\nu, \pi^{(t+1)}) - V_i(\nu, \pi^{(t)}) \geqslant \frac{1}{2\eta(1-\gamma)} \left(1 - \frac{4\eta \kappa_\nu^3 \bar{A} N}{(1-\gamma)^4} \right) \\ &\quad \cdot \sum_{i=1}^{N} \sum_{s \in S} d_\nu^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(s) \left\| \pi_i^{(t+1)}(s) - \pi_i^{(t)}(s) \right\|^2. \end{aligned}$$

Proof. This result directly follows from [121, Lemma 3]. Specifically, the proof of [121, Lemma 3] is established by lower-bounding the difference $\Phi(\nu, \pi^{(t+1)}) - \Phi(\nu, \pi^{(t)})$ for a Markov potential game with potential function Φ. At its core, the proof relies on the key property of Markov potential games, which allows the difference in potential functions to be expressed as the difference in value functions for each player. The remainder of the proof focuses on lower-bounding the difference in value functions at each step of the policy update process in Algorithm 1, which is precisely what we require. We omit details due to space constraints. □

Proof of Lemma 2.5.2

For ease of exposition, let $\pi' = \pi^{(t+1)}$ and $\pi = \pi^{(t)}$. By Definition 2.2.2, $|V_i(\nu, \pi_i', \pi_{-i}) - V_i(\nu, \pi_i, \pi_{-i}) - (\Phi(\nu, \pi_i', \pi_{-i}) - \Phi(\nu, \pi_i, \pi_{-i}))| \leqslant \alpha$ for any $\nu, i \in I, \pi_i, \pi_i' \in \Pi_i$ and $\pi_{-i} \in \Pi_{-i}$. Apply Lemma A.2.2 with $f(\cdot) = V_i(\nu, \cdot) - \Phi(\nu, \cdot)$ respectively. Since each term in

(A.16) only differs in one player's policy, we obtain

$$\begin{aligned}&|V_i(\nu,\pi')-V_i(\nu,\pi)-(\Phi(\nu,\pi')-\Phi(\nu,\pi'))|\\&\leqslant\sum_{i=1}^{N}\alpha+\sum_{i=1}^{N}\sum_{j=i+1}^{N}\alpha\leqslant N^2\alpha.\end{aligned}$$

The proof follows by the above inequality and Lemma A.2.3.

Proofs in Section 2.5

Proof of Lemma 2.5.3

Fix arbitrary $i\in I,\mu\in\Delta(S),\pi_i,\pi_i'\in\Pi_i,\pi_{-i}\in\Pi_{-i}$. We define $\pi=(\pi_i,\pi_{-i}),\pi'=(\pi_i',\pi_{-i})\in\Pi$. Note that

$$\begin{aligned}&\tilde{V}_i(\mu,\pi)-\tilde{V}_i(\mu,\pi')\\&=\mathbb{E}_\pi\bigg[\sum_{k=0}^{\infty}\gamma^k\bigg(u_i(s^k,a^k)-\tau\sum_{j\in I}\nu_j(s^k,\pi_j)-\tilde{V}_i(s^k,\pi')+\tilde{V}_i(s^k,\pi')\bigg)\bigg]-\tilde{V}_i(\mu,\pi')\\&=\mathbb{E}_\pi\bigg[\sum_{k=0}^{\infty}\gamma^k\bigg(u_i(s^k,a^k)-\tau\sum_{j\in I}\nu_j(s^k,\pi_j)-\tilde{V}_i(s^k,\pi')\bigg)\bigg]+\mathbb{E}_\pi\left[\sum_{k=1}^{\infty}\gamma^k\tilde{V}_i(s^k,\pi')\right].\quad\text{(A.17)}\end{aligned}$$

Note that

$$\mathbb{E}_\pi\big[\sum_{k=1}^{\infty}\gamma^k\tilde{V}_i(s^k,\pi')\big]=\gamma\mathbb{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^k\tilde{V}_i(s^{k+1},\pi')\right].$$

Therefore,

$$\begin{aligned}\text{(A.17)}&=\mathbb{E}_\pi\Big[\sum_{k=0}^{\infty}\gamma^k\Big(u_i\left(s^k,a^k\right)-\tau\sum_{j\in I_N}\nu_j\left(s^k,\pi_j\right)-\tilde{V}_i\left(s^k,\pi'\right)+\gamma\tilde{V}_i\left(s^{k+1},\pi'\right)\Big)\Big]\\&=\mathbb{E}_\pi\bigg[\sum_{k=0}^{\infty}\gamma^k\bigg(u_i(s^k,a^k)-\tau\sum_{j\in I}\nu_j(s^k,\pi_j')+\gamma\sum_{s'\in S}P(s'|s^k,a^k)\tilde{V}_i(s',\pi')-\tilde{V}_i(s^k,\pi')\\&\qquad+\tau\sum_{j\in I}\nu_j(s^k,\pi_j')-\tau\sum_{j\in I}\nu_j(s^k,\pi_j)\bigg)\bigg]\\&=\mathbb{E}_{\pi_i}\bigg[\sum_{k=0}^{\infty}\gamma^k\bigg(\tilde{Q}_i^{\pi'}(s^k,a_i^k)-\tilde{V}_i(s^k,\pi')+\tau\sum_{j\in I}\nu_j(s^k,\pi_j')-\tau\sum_{j\in I}\nu_j(s^k,\pi_j)\bigg)\bigg].\end{aligned}\quad\text{(A.18)}$$

We can continue the above calculations by noting that $\pi'_j = \pi_j$ for all $j \neq i$ and $\tilde{V}_i(s', \pi') = \pi'_i(s')^\top \tilde{\mathbf{Q}}_i^{\pi'}(s')$,

$$\text{(A.18)}=\frac{1}{1-\gamma}\sum_{s'\in S} d_\mu^\pi(s')\Big(\pi_i(s')-\pi'_i(s')\Big)^\top \tilde{\mathbf{Q}}_i^{\pi'}(s') + \tau\nu_i(s',\pi'_i) - \tau\nu_i(s',\pi_i)\Big).$$

Proof of Lemma 2.5.4

From the definition of smoothed infinite horizon utility (2.16), we note that for every $i \in I, \pi_i \in \Pi_i, \pi_{-i} \in \Pi_{-i}, s \in S$,

$$\tilde{V}_i(s,\pi_i,\pi_{-i}) = V_i(s,\pi_i,\pi_{-i}) - \tau\mathbb{E}_\pi\left[\sum_{k=0}^\infty \gamma^k \sum_{j\in I}\nu_j(s^k,\pi_j)|s_0=s\right]. \tag{A.19}$$

Using (A.19), it holds that for any $\mu \in \Delta(S)$ and $\pi \in \Pi$,

$$\begin{aligned}|\tilde{V}_i(\mu,\pi) - V_i(\mu,\pi)| &= \tau\left|\mathbb{E}_{\mu,\pi}\left[\sum_{k=0}^\infty \gamma^k\sum_{j\in I}\nu_j(s^k,\pi_j)\right]\right| \\ &\leqslant \frac{\tau N \max_{s,\pi_i}\nu_i(s,\pi_i)}{1-\gamma} = \frac{\tau N\log(\bar{A})}{1-\gamma}.\end{aligned} \tag{A.20}$$

The desired result follows from triangle inequality and (A.20).

Proof of Lemma 2.5.5

The proof of Lemma 2.5.5 requires the following technical lemmas.

Lemma A.2.4. *If $\mathcal{G}$ is a Markov α-potential game with Φ as its α-potential function, then for any $s \in S, i \in I, \pi'_i, \pi_i \in \Pi_i, \pi_{-i} \in \Pi_{-i}$, $\Big|(\tilde{\Psi}(s,\pi'_i,\pi_{-i}) - \tilde{\Psi}(s,\pi_i,\pi_{-i})) - (\tilde{V}_i(s,\pi'_i,\pi_{-i}) - \tilde{V}_i(s,\pi_i,\pi_{-i}))\Big| \leqslant \alpha$, where*

$$\tilde{\Psi}(s,\pi) := \Phi(s,\pi) - \tau\mathbb{E}_\pi[\sum_{j\in I}\sum_{k=0}^\infty \gamma^k\nu_j(s^k,\pi_j) \mid s^0 = s].$$

Proof. To ease the notation, for function $f : S\times\Pi \to \mathbb{R}$, we write $f(s,\cdot)$ as $f^s(\cdot)$. By (A.19) and the definition of $\tilde{\Psi}$ in Lemma 2.5.5, we have for all $s \in S, i \in I, \pi'_i, \pi_i \in \Pi_i, \pi_{-i} \in \Pi_{-i}$,

$$\begin{aligned}&|\tilde{\Psi}^s(\pi'_i,\pi_{-i}) - \tilde{\Psi}^s(\pi_i,\pi_{-i}) - (\tilde{V}_i^s(\pi'_i,\pi_{-i}) - \tilde{V}_i^s(\pi_i,\pi_{-i}))| \\ =&|\Phi^s(\pi'_i,\pi_{-i}) - \Phi^s(\pi_i,\pi_{-i}) - (V_i^s(\pi'_i,\pi_{-i}) - V_i^s(\pi_i,\pi_{-i}))|,\end{aligned}$$

which is bounded by α using Definition 2.2.3. □

Lemma A.2.5. *For any $i \in I, s \in S, \pi_i' \in \Pi_i, t \in [T]$, it hold that*

$$\sum_{a_i \in A_i} \tilde{Q}_i^{(t)}(s, a_i) \left(BR_i^{(t)}(a_i|s) - \pi_i'(a_i|s) \right) \geqslant \tau \sum_{a_i \in A_i} \log \left(BR_i^{(t)}(a_i|s) \right) \left(BR_i^{(t)}(a_i|s) - \pi_i'(a_i|s) \right).$$

Proof. Fix arbitrary $i \in I, s \in S$, and $t \in [T]$. Next, note that the optimization problem in (2.18) is a strongly concave optimization problem. By the first order conditions of constrained optimality, for all $\pi_i' \in \Pi_i$,

$$\left(\tilde{\mathbf{Q}}_i^{(t)}(s) - \tau \nabla_{\pi_i(s)} \nu_i(s, \mathrm{BR}_i^{(t)}(s)) \right)^\top (\mathrm{BR}_i^{(t)}(s) - \pi_i'(s)) \geqslant 0.$$

Note that $\nabla_{\pi_i(a_i|s)} \nu_i(s, \pi_i) = 1 + \log(\pi_i(a_i|s))$ for every $a_i \in A_i$. Therefore, for every $\pi_i' \in \Pi_i$,

$$\sum_{a_i \in A_i} \tilde{Q}_i^{(t)}(s, a_i) \left(\mathrm{BR}_i^{(t)}(a_i|s) - \pi_i'(a_i|s) \right) \geqslant \tau \sum_{a_i \in A_i} \left(1 + \log \left(\mathrm{BR}_i^{(t)}(a_i|s) \right) \right) \left(\mathrm{BR}_i^{(t)}(a_i|s) - \pi_i'(a_i|s) \right).$$

The result follows by noting that $\sum_{a_i \in A_i} \mathrm{BR}_i^{(t)}(a_i|s) = \sum_{a_i \in A_i} \pi_i'(a_i|s) = 1$. □

Lemma A.2.6. *For any $i \in I, s \in S, \pi_i, \pi_i' \in \Pi_i$,*

$$\nu_i(s, \pi_i) - \nu_i(s, \pi_i') \geqslant \frac{1}{2} \|\pi_i(s) - \pi_i'(s)\|^2 + \sum_{a_i \in A_i} \left(\log(\pi_i'(a_i|s)) \right) \left(\pi_i(a_i|s) - \pi_i'(a_i|s) \right).$$

Proof. Fix arbitrary $i \in I, s \in S$. To prove the lemma, we first claim that the mapping $\Delta(A_i) \ni \pi \mapsto \nu_i(s, \pi)$ is 1-strongly convex. This can be observed by computing the Hessian, which is a $\mathbb{R}^{A_i \times A_i}$ diagonal matrix with (a_i, a_i) entry as $1/\pi(a_i|s)$. Since $\pi(a_i|s) \leqslant 1$, it follows that the diagonal entries of the Hessian matrix are all greater than 1. Thus, $\nu_i(s, \cdot)$ is 1-strongly convex function. The result follows by noting that for any κ-strongly convex function f,

$$f(y) \geqslant f(x) + \nabla f(x)^\top (y - x) + \frac{\kappa}{2} \|y - x\|^2.$$

□

Lemma A.2.7. *For any $i \in I, t \in [T], a \in A_{\bar{i}(t)}$, there exists $0 \leqslant t^* \leqslant t$ such that*

$$\tau |\log(\pi_{\bar{i}(t)}^{(t)}(a|\bar{s}^{(t)}))| \leqslant 2 \|\tilde{\mathbf{Q}}_{\bar{i}(t)}^{(t^*)}(\bar{s}^{(t)})\|_\infty + \tau \log(|A_{\bar{i}(t)}|).$$

Proof. Recall that in Algorithm 2, at any time step $t \in [T]$, player $\bar{i}^{(t)}$ updates her policy at time $t+1$ in the state $\bar{s}^{(t)}$, while policies for other players and other states remain unchanged. Fix arbitrary $t \in [T]$. Let $0 \leqslant t^* \leqslant t$ be the latest time step when player $\bar{i}^{(t)}$ updated its policy in state $\bar{s}^{(t)}$ before time t. Note that $t^* = 0$ if t is the first time when player $\bar{i}^{(t)}$ is updating its policy in state $\bar{s}^{(t)}$. Naturally, $\bar{i}^{(t)} = \bar{i}^{(t^*)}$ and $\bar{s}^{(t)} = \bar{s}^{(t^*)}$. Consequently, for every $a \in A_{\bar{i}^{(t)}}$,

$$\pi^{(t)}_{\bar{i}^{(t)}}(a|\bar{s}^{(t)}) = \mathrm{BR}^{(t^*)}_{\bar{i}^{(t)}}(a|\bar{s}^{(t)}) = \frac{\exp(\tilde{Q}^{(t^*)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}, a))}{\sum_{a' \in A_{\bar{i}^{(t)}}} \exp(\tilde{Q}^{(t^*)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}, a'))}.$$

Consequently, for every $a \in A_{\bar{i}^{(t)}}$,

$$\begin{aligned}
\pi^{(t)}_{\bar{i}^{(t)}}(a|\bar{s}^{(t)}) &\geqslant \frac{\exp(\tilde{Q}^{(t^*)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}, \underline{a})/\tau)}{|A_{\bar{i}^{(t)}}| \exp(\tilde{Q}^{(t^*)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}, \bar{a})/\tau)} \\
&= \frac{1}{|A_{\bar{i}^{(t)}}|} \exp\left(\left(\tilde{Q}^{(t^*)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}, \underline{a}) - \tilde{Q}^{(t^*)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}, \bar{a})\right)/\tau\right),
\end{aligned}$$

with $\bar{a} \in \arg\max_{a \in A_{\bar{i}^{(t)}}} \tilde{Q}^{(t^*)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}, a)$ and $\underline{a} \in \arg\min_{a \in A_{\bar{i}^{(t)}}} \tilde{Q}^{(t^*)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}, a)$. Since $\pi^{(t)}_{\bar{i}^{(t)}}(a|\bar{s}^{(t)}) \leqslant 1$, it follows that for every $a \in A_{\bar{i}^{(t)}}$,

$$\begin{aligned}
&|\log(\pi^{(t)}_{\bar{i}^{(t)}}(a|\bar{s}^{(t)}))| \\
&\leqslant \log(|A_{\bar{i}^{(t)}}|) + \frac{1}{\tau}\left(\tilde{Q}^{(t^*)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}, \bar{a}) - \tilde{Q}^{(t^*)}_{\bar{i}^{(t)}}(\bar{s}^{(t)}, \underline{a})\right) \\
&\leqslant \log(|A_{\bar{i}^{(t)}}|) + \frac{2}{\tau}\|\tilde{\mathbf{Q}}^{(t^*)}_{\bar{i}^{(t)}}(\bar{s}^{(t)})\|_\infty.
\end{aligned}$$

Lemma A.2.8. *For any $t \in [T], i \in I, s \in S$, it holds that $\|\tilde{\mathbf{Q}}^{(t)}_i(s)\|_\infty \leqslant C\frac{1+\tau N \log(\bar{A})}{1-\gamma}$, where $C := \max_{i \in I} \|u_i\|_\infty$.*

Proof. First, we note that for any $s \in S$, $\pi \in \Pi$,

$$\begin{aligned}
|\tilde{V}_i(s, \pi)| &\leqslant \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k |u_i(s^k, a^k) - \tau \sum_{j \in I_N} \nu_j(s^k, \pi_j)|\right] \\
&\leqslant \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k \left(|u_i(s^k, a^k)| + \tau N \log(\bar{A})\right)\right] \\
&\leqslant C\frac{1 + \tau N \log(\bar{A})}{(1-\gamma)}.
\end{aligned}$$

By (2.17), we note that for every $i \in I, s \in S, a_i \in A_i$,

$$|\tilde{Q}_i^{(t)}(s,a_i)| \leqslant \mathop{\mathbb{E}}_{a_{-i}\sim\pi_{-i}}\Big[|u_i(s,a_i,a_{-i}) - \tau \sum_{j\in I_N} \nu_j(s,\pi_j)| + \gamma \sum_{s'\in S} P(s'|s,a_i,a_{-i})\Big|\tilde{V}_i(s',\pi)\Big|\Big]$$
$$\leqslant C \mathop{\mathbb{E}}_{a_{-i}\sim\pi_{-i}}\left[(1+\tau N\log(\bar{A}))\left(1+\frac{\gamma}{1-\gamma}\right)\right].$$

Proof of Lemma 2.5.5

(1) Fix $t \in [T]$. To ease the notation, let $\pi'_* := \pi^{(t+1)}_{\bar{i}(t)}$, $\pi_* := \pi^{(t)}_{\bar{i}(t)}$, $\pi_{-*} := \pi^{(t)}_{-\bar{i}(t)}$, $\nu_* := \nu_{\bar{i}(t)}$, Q_* denote $\tilde{Q}^{(t)}_{\bar{i}(t)}$, $\mathbf{Q}_*$ denote $\tilde{\mathbf{Q}}^{(t)}_{\bar{i}(t)}$. Note that by (2.19) and (2.21),

$$\Omega^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)}) = \sum_{a\in A_{\bar{i}(t)}} \left(\pi'_*(a|\bar{s}^{(t)}) - \pi_*(a|\bar{s}^{(t)})\right) Q_*(\bar{s}^{(t)},a) + \tau\nu_*(\bar{s}^{(t)},\pi_*) - \tau\nu_*(\bar{s}^{(t)},\pi'_*)$$
$$\leqslant \sum_{a\in A_{\bar{i}(t)}} \left(\pi'_*(a|\bar{s}^{(t)}) - \pi_*(a|\bar{s}^{(t)})\right) Q_*(\bar{s}^{(t)},a)$$
$$+ \tau \sum_{a\in A_{\bar{i}(t)}} \log(\pi_*(a|\bar{s}^{(t)}))\left(\pi_*(a|\bar{s}^{(t)}) - \pi'_*(a|\bar{s}^{(t)})\right)$$
$$\leqslant \sum_{a\in A_{\bar{i}(t)}} \left(\left|\pi'_*(a|\bar{s}^{(t)}) - \pi'_*(a|\bar{s}^{(t)})\right| \cdot \left|Q_*(\bar{s}^{(t)},a) - \tau\log(\pi_*(a|\bar{s}^{(t)}))\right|\right), \tag{A.21}$$

where the first inequality follows from convexity of $\nu_i(s,\cdot)$. By Cauchy-Schwarz inequality and noting that $\max_{i\in I}|A_i| \leqslant \bar{A}$,

$$\text{(A.21)} \leqslant \sqrt{\bar{A}} \max_{a\in A_{\bar{i}(t)}} \left|Q_*(\bar{s}^{(t)},a) - \tau\log(\pi_*(a|\bar{s}^{(t)}))\right| \left\|\pi'_*(\bar{s}^{(t)}) - \pi_*(\bar{s}^{(t)})\right\|_2$$
$$\leqslant \sqrt{\bar{A}}\left(\max_{a\in A_{\bar{i}(t)}} \left|Q_*(\bar{s}^{(t)},a)\right| + \max_{a\in A_{\bar{i}(t)}} \tau\left|\log(\pi_*(a|\bar{s}^{(t)}))\right|\right) \left\|\pi'_*(\bar{s}^{(t)}) - \pi_*(\bar{s}^{(t)})\right\|_2.$$

Note that Lemma A.2.7 implies that there exists $\hat{t} \leqslant t$ such that

$$\max_{a\in A_{\bar{i}(t)}} \tau\left|\log(\pi_*(a|\bar{s}^{(t)}))\right| \leqslant 2\|\tilde{\mathbf{Q}}^{(\hat{t})}_{\bar{i}(t)}(\bar{s}^{(t)})\|_\infty + \tau\log(\bar{A}).$$

Consequently, it follows that

$$\Omega^{(t)}_{\bar{i}(t)}(\bar{s}^{(t)}) \leqslant \sqrt{\bar{A}}\Big(\|\mathbf{Q}_*(\bar{s}^{(t)})\|_\infty + 2\|\tilde{\mathbf{Q}}^{(\hat{t})}_{\bar{i}(t)}(\bar{s}^{(t)})\|_\infty + \tau\log(\bar{A})\Big) \cdot \left\|\pi'_*(\bar{s}^{(t)}) - \pi_*(\bar{s}^{(t)})\right\|_2$$
$$\leqslant 4C\frac{1+\tau N\log(\bar{A})}{1-\gamma}\sqrt{\bar{A}}\left\|\pi'_*(\bar{s}^{(t)}) - \pi_*(\bar{s}^{(t)})\right\|_2,$$

where the last inequality follows from Lemma A.2.8. This concludes the proof for Lemma 2.5.5 1).
(2) Here, we show that

$$\sum_{t=1}^{T-1} \|\pi'_*(\bar{s}^{(t)}) - \pi_*(\bar{s}^{(t)})\|_2^2 \leqslant \frac{2}{\tau\bar{\mu}} \left(\tilde{\Psi}(\mu, \pi^{(T)}) - \tilde{\Psi}(\mu, \pi^{(0)}) + \alpha T\right).$$

To see this, note that for any $t \in [T]$,

$$\begin{aligned}
&\tilde{\Psi}(\mu, \pi^{(t+1)}) - \tilde{\Psi}(\mu, \pi^{(t)}) = \tilde{\Psi}(\mu, \pi'_*, \pi_{-*}) - \tilde{\Psi}(\mu, \pi_*, \pi_{-*}) \\
&\overset{(i)}{\geqslant} \tilde{V}_{\bar{i}^{(t)}}(\mu, \pi'_*, \pi_{-*}) - \tilde{V}_{\bar{i}^{(t)}}(\mu, \pi_*, \pi_{-*}) - \alpha \\
&\overset{(ii)}{=} \frac{1}{1-\gamma} \sum_{s\in S} d_\mu^{\pi'_*, \pi_{-*}}(s) \left(\left(\pi'_*(s) - \pi_*(s)\right)^\top \mathbf{Q}_*(s) + \tau\nu_*(s, \pi_*) - \tau\nu_*(s, \pi'_*) \right) - \alpha \\
&\overset{(iii)}{=} \frac{1}{1-\gamma} d_\mu^{\pi'_*, \pi_{-*}}(\bar{s}^{(t)}) \left((\pi'_*(\bar{s}^{(t)}) - \pi_*(\bar{s}^{(t)})^\top) \cdot \mathbf{Q}_*(\bar{s}^{(t)}) + \tau\nu_*(\bar{s}^{(t)}, \pi_*) - \tau\nu_*(\bar{s}^{(t)}, \pi'_*) \right) - \alpha,
\end{aligned} \tag{A.22}$$

where (i) follows from Lemma A.2.4, (ii) follows from Lemma 2.5.3, and (iii) holds because $\pi'_*(s) = \pi_*(s)$ for all $s \neq \bar{s}^{(t)}$. Next, from Algorithm 2, note that $\pi'_*(\bar{s}^{(t)}) = \mathrm{BR}^{(t)}_{\bar{i}^{(t)}}(\bar{s}^{(t)})$. Consequently, using Lemma A.2.5, we obtain

$$\begin{aligned}
&\text{(A.22)} \\
&\geqslant \frac{\tau d_\mu^{\pi'_*, \pi_{-*}}(\bar{s}^{(t)})}{1-\gamma} \Big(\log(\pi'_*(\bar{s}^{(t)}))^\top \cdot \left(\pi'_*(\bar{s}^{(t)}) - \pi_*(\bar{s}^{(t)})\right) + \nu_*(\bar{s}^{(t)}, \pi_*) - \nu_*(\bar{s}^{(t)}, \pi'_*) \Big) - \alpha.
\end{aligned} \tag{A.23}$$

Furthermore, using Lemma A.2.6, we obtain

$$\begin{aligned}
\text{(A.23)} &\geqslant \frac{\tau}{2(1-\gamma)} d_\mu^{\pi'_*, \pi'_{-*}(\bar{s}^{(t)})} \|\pi'_*(\bar{s}^{(t)}) - \pi_*(\bar{s}^{(t)})\|_2^2 - \alpha \\
&\overset{(a)}{\geqslant} \frac{\tau\bar{\mu}}{2} \|\pi'_*(\bar{s}^{(t)}) - \pi_*(\bar{s}^{(t)})\|_2^2 - \alpha,
\end{aligned}$$

where (a) follows from $d_\mu^{\pi_i^{(t+1)}, \pi_{-i}^{(t)}}(\bar{s}^{(t)}) \geqslant (1-\gamma)\bar{\mu}$. Summing the above inequality over all $t \in [T]$ yields:

$$\begin{aligned}
\tilde{\Psi}(\mu, \pi^{(T)}) - \tilde{\Psi}(\mu, \pi^{(0)}) &= \sum_{t\in[T]} \tilde{\Psi}(\mu, \pi^{(t+1)}) - \tilde{\Psi}(\mu, \pi^{(t)}) \\
&\geqslant \frac{\tau\bar{\mu}}{2} \sum_{t\in[T]} \|\pi'_*(\bar{s}^{(t)}) - \pi_*(\bar{s}^{(t)})\|_2^2 - \alpha T.
\end{aligned}$$

Finally to conclude Lemma 2.5.5 2), note that

$$\begin{aligned}
&\sum_{t\in[T]} \|\pi'_*(\bar{s}^{(t)}) - \pi_*(\bar{s}^{(t)})\|_2^2 \\
&\leqslant \frac{2}{\tau\bar{\mu}}\left(\tilde{\Psi}(\mu,\pi^{(T)}) - \tilde{\Psi}(\mu,\pi^{(0)}) + \alpha T\right) \\
&\leqslant \frac{2}{\tau\bar{\mu}}\left(|\Phi(\mu,\pi^{(T)}) - \Phi(\mu,\pi^{(0)})| + \frac{2\tau N\log(\bar{A})}{1-\gamma} + \alpha T\right),
\end{aligned}$$

where the last inequality follows by noting that for any $\pi,\pi' \in \Pi$ and any $\mu \in \Delta(S)$,

$$\begin{aligned}
|\tilde{\Psi}(\mu,\pi) - \tilde{\Psi}(\mu,\pi')| &\leqslant \left|\Phi(\mu,\pi) - \Phi(\mu,\pi')\right| \\
&\quad + \tau\left|\mathbb{E}_\pi\left[\sum_{\substack{j\in I\\ t\in\mathbb{N}}} \gamma^t \nu_j(s^t,\pi_j)\right]\right| + \tau\left|\mathbb{E}_{\pi'}\left[\sum_{\substack{j\in I\\ t\in\mathbb{N}}} \gamma^t \nu_j(s^t,\pi_j)\right]\right| \\
&\leqslant |\Phi(\mu,\pi) - \Phi(\mu,\pi')| + 2\tau\frac{N\log(\bar{A})}{1-\gamma}.
\end{aligned}$$

A.3 Algorithms to solve semi-infinite linear programming

In this section, we present an algorithm based on the stochastic gradient method from [402] to solve the semi-infinite linear programming problem (2.10). Denote $C \coloneqq N\max_{i\in I}\|u_i\|_\infty$ and define

$$\begin{aligned}
g(\phi,y;\pi,\pi') = \max\Bigg\{ &\max_{i\in I}\left|\sum_{s',a'}(d^s(s',a';\pi_i,\pi_{-i}) - d^s(s',a';\pi'_i,\pi_{-i}))(\phi-u_i)(s',a')\right| - y, \\
&\max_{s\in S,a\in A}|\phi(s,a)| - C\Bigg\},
\end{aligned} \tag{A.24}$$

which ensures that constraint (C1) in (2.10) can be rewritten as $g(\phi,y;\pi,\pi') \leqslant 0, \forall \pi,\pi' \in \Pi$. Let $h:\mathbb{R}\to\mathbb{R}$ be a convex differentiable function such that

$$h(x) = 0 \text{ for all } x \leqslant 0, \text{ and } h(x) > 0 \text{ for all } x > 0.$$

A candidate choice of h is $h(x) = (\max\{0,x\})^2$. Finally, we consider step-size schedules $\{\eta_t\}_{t=1}^\infty$ and $\{\beta_t\}_{t=1}^\infty$ such that

$$\lim_{t\to\infty}\beta_t = \infty, \sum_{t=1}^\infty \eta_t^2\beta_t^2 < \infty, \sum_{t=1}^\infty \eta_t = \infty, \text{ and } \eta_t > 0, \beta_t < \beta_{t+1} \text{ for all } t \geqslant 0. \tag{A.25}$$

Theorem 4 in [402] shows that with probability 1, $(y^{(t)},\phi^{(t)})$ almost surely converges to a solution of (2.10).

Algorithm 14 Algorithm to solve (2.10) [402]

Input: $y^{(0)} \in \mathbb{R}_+, \phi^{(0)} \in \mathbb{R}^{S\times A}$, $\{\eta_t\}_{t=1}^{\infty}$ and $\{\beta_t\}_{t=1}^{\infty}$ satisfying (A.25).
for $t = 0, 1, 2, ..., T-1$ **do**
Sample π, π' in Π from uniform distribution and calculate $g(\phi^{(t)}, y^{(t)}; \pi, \pi')$ in (A.24).
Update $\phi^{(t)}$ with

$$\phi^{(t+1)} = \phi^{(t)} - \eta_{t+1}\beta_{t+1}h'\left(g\left(\phi^{(t)}, y^{(t)}; \pi, \pi'\right)\right) \cdot \nabla_{\phi} g\left(\phi^{(t)}, y^{(t)}; \pi, \pi'\right), \tag{A.26}$$

and update $y^{(t)}$ with

$$y^{(t+1)} = y^{(t)} - \eta_{t+1}\Big(1 + \beta_{t+1}h'\left(g\left(\phi^{(t)}, y^{(t)}; \pi, \pi'\right)\right) \cdot \nabla_{y} g\left(\phi^{(t)}, y^{(t)}; \pi, \pi'\right)\Big).$$

end for

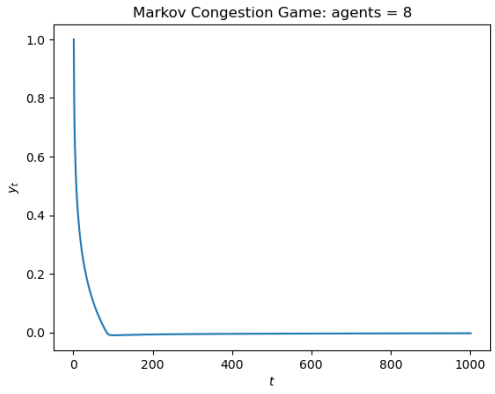


Figure A.1: Find the game elasticity parameter α for MCG using Algorithm 14. ($\phi^{(0)} = \phi^*$, $\eta_t = \frac{1}{t}, \beta_t = t^{0.4999}$, $\forall t \geqslant 1$.)

State-wise potential games. Algorithm 14 iteratively updates the variables $y \in \mathbb{R}$ and $\phi \in \mathbb{R}^{S\times A}$. However, this method may be slow as the dimension of ϕ scales with $|S| \cdot |A|$. For MCGs, where each state is a static potential game, one can utilize the game structure to accelerate the convergence of algorithm.

For an MCG $\mathcal{G}_{\mathsf{mcg}}$, there exists a function $\phi^* : S \times A \to \mathbb{R}$ such that for every $i \in I, s \in S, a_i, a_i' \in A_i, a_{-i} \in A_{-i}$, $|\phi^*(s, a_i, a_{-i}) - \phi^*(s, a_i', a_{-i}) - (u_i(s, a_i, a_{-i}) - u_i(s, a_i', a_{-i}))| = 0$. Then one can input $\phi^{(0)} = \phi^*$ and omit the update of $\phi^{(t)}$ in (A.26) in Algorithm (14). Figure A.1 shows the empirical performance of Algorithm 14 for the Markov congestion game. Note that with the setting in Section 2.6, $y^{(t)}$ converges to 0, which suggests that $\mathcal{G}_{\mathsf{mcg}}$ may be an MPG, at least for some model parameters.

Appendix B

Appendix for Chapter 3

Here, we provide additional material supplementing the content of Chapter 3.

B.1 Description of Single-agent Racing Line

Race drivers follow a racing line for specific maneuvers. This line can be used as a reference path by the motion planner to assign time-optimal trajectories while avoiding collision. The racing line is minimum-time or minimum-curvature. They are similar, but the minimum-curvature path additionally allows the highest cornering speeds given the maximum legitimate lateral acceleration [177].

There are many proposed solutions to finding the optimal racing line, including nonlinear optimization [352, 177], genetic algorithm-based search [413] and Bayesian optimization [190]. However, for our work, we calculate the minimum-curvature optimal line, which is close to the optimal racing line as proposed by [177]. The race track is represented by a sequence of tuples (x_i,y_i,w_i), $i \in \{0, ..., N-1\}$, where (x_i,y_i) denotes the coordinate of the center location and w_i denotes the lane width at the i-th point. The output racing line consists of a tuple of seven variables: coordinates x and y, longitudinal displacement s, longitudinal velocity v_x, acceleration a_x, heading angle ψ, and curvature κ. It is obtained by minimizing the following cost:

$$\begin{aligned} \min_{\eta_1 \dots \eta_N} \quad & \sum_{n=0}^{N-1} \kappa_i^2(n) \\ \text{s.t.} \quad & \eta_i \in \left[-\frac{w_i}{2} + \frac{w_{veh}}{2}, \frac{w_i}{2} - \frac{w_{veh}}{2} \right] \end{aligned} \tag{B.1}$$

where the vehicle width is w_{veh}, and η_i is the lateral displacement with respect to the reference center line.

To create a velocity profile, we need to consider the vehicle's constraints on both longitudinal and lateral acceleration [177]. Our approach involves generating a library of velocity

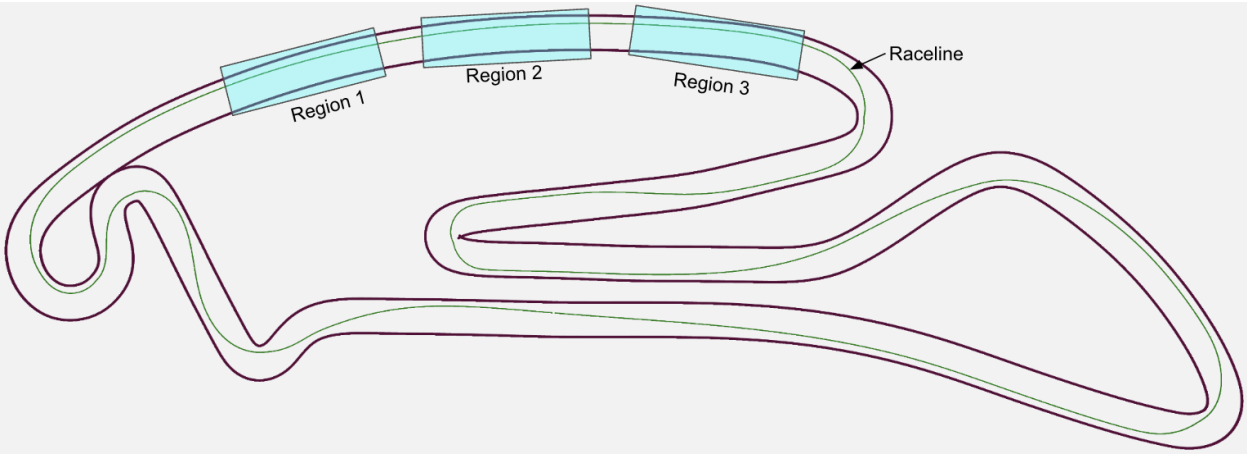


Figure B.1: Track and the starting position regions

profiles, each tailored to specific lateral acceleration limits determined by the friction coefficients for the front (μ_f) and rear (μ_r) tires, as well as the vehicle's mass (m) and the gravitational constant (g). In particular, we produce a set of velocity profiles covering a range of maximum lateral forces corresponding to the friction μ_{eff} within the interval $[\mu_{min}, \mu_{max}]$. This library allows us to retrieve a velocity profile that matches a given value of μ. Interpolation is necessary when we encounter a friction value that falls within the valid range but is not explicitly present in the library.

An example of a racing line calculated for the racetrack used in our numerical study in Section 3.3 is shown in Figure B.1.

B.2 Dynamic Bicycle Model

For any car i, we denote its mass by m^i, its moment of inertia in the vertical direction about the center of mass by I_z^i, the distance between the center of mass (COM) and its front wheel by l_f^i, and the distance from the COM to the rear wheel l_r^i. Also, κ_t^i denotes the inverse of radius of curvature of the track at $p_{x,t}^i$. Using these notations, the dynamics of car i is defined below:

$$\begin{bmatrix} p_{x,t+1}^i \\ p_{y,t+1}^i \\ \phi_{t+1}^i \\ \tilde{v}_{x,t+1}^i \\ \tilde{v}_{y,t+1}^i \\ \omega_{t+1}^i \end{bmatrix} = \begin{bmatrix} p_{x,t}^i \\ p_{y,t}^i \\ \phi_t^i \\ v_{x,t}^i \\ v_{y,t}^i \\ \omega_t^i \end{bmatrix} + \Delta t \begin{bmatrix} v_{x,t}^i \\ v_{y,t}^i \\ \omega_t^i - \frac{\kappa_t^i}{(1-\kappa_t^i p_{y,t}^i)}(\tilde{v}_{x,t}^i \cos(\phi_t^i) - \tilde{v}_{y,t}^i \sin(\phi_t^i)) \\ \frac{1}{m^i}(F_{r,x,t}^i - F_{f,y,t}^i \sin(\delta_t^i) + m^i \tilde{v}_{y,t}^i \omega_t^i) \\ \frac{1}{m^i}(F_{r,y,t}^i + F_{f,y,t}^i \cos(\delta_t^i) - m^i \tilde{v}_{x,t}^i \omega_t^i) \\ \frac{1}{I_z^i}(F_{f,y,t}^i l_f^i \cos(\delta_t^i) - F_{r,y,t}^i l_r^i) \end{bmatrix}, \tag{B.2}$$

where (i) $v^i_{x,t} = \frac{1}{(1-\kappa^i_t p^i_{y,t})}(\tilde{v}^i_{x,t}\cos(\phi^i_t) - \tilde{v}^i_{y,t}\sin(\phi^i_t))$, $v^i_{y,t} = \tilde{v}^i_{x,t}\sin(\phi^i_t) + \tilde{v}^i_{y,t}\cos(\phi^i_t)$ are the velocities in frenet frame; (ii) $\tilde{v}^i_{x,t}, \tilde{v}^i_{y,t}$ are velocities in body frame; (iii) $F^i_{r,x,t} = (C_1 - C_2\tilde{v}^i_{x,t})d^i_t - C_3 - C_4(\tilde{v}^i_{x,t})^2$ is the longitudinal force on the rear tire at time t. Here, C_1 and C_2 are parameters that govern the longitudinal force generated on the car in response to the throttle command, while C_3 and C_4 are parameters that account for the friction and drag forces acting on the car; (iv) $F^i_{f,y,t} = D_f \sin(C_f \tan^{-1}(B_f \alpha^i_{f,t}))$ is the lateral force on the front tire depending on the slipping angle $\alpha^i_{f,t}$, which is given by $\alpha^i_{f,t} = \delta^i_t - \tan^{-1}\left(\frac{\omega^i_t l_f + \tilde{v}^i_{y,t}}{\tilde{v}^i_{x,t}}\right)$. Here B_f, C_f, D_f are the parameters of Pacejka tire model; and (v) $F^i_{r,y,t} = D^i_r \sin(C^i_r \tan^{-1}(B^i_r \alpha^i_{r,t}))$ is the lateral force on the rear tire depending on the slipping angle $\alpha^i_{r,t}$, which is given by $\alpha^i_{r,t} = \tan^{-1}\left(\frac{\omega^i_t l_r - \tilde{v}^i_{y,t}}{\tilde{v}^i_{x,t}}\right)$. Here B^i_r, C^i_r, D^i_r are the parameters of Pacejka tire model.

B.3 Hyperparameters

Network architecture

We use a simple feed-forward deep neural network with ReLU activation except for the last layer to represent the value function and the potential function. The network for value function consists of 3 hidden layers with $(128, 128, 64)$ hidden features on each layer. The network for potential function consists of 3 hidden layers with $(384, 384, 192)$ hidden features on each layer.

Training

We use a learning rate of 0.0001 and train for 50000 epochs (both value functions and potential function). Each race consists of 500 time steps with $\Delta t = 0.1s$, hence $50s$ race.

B.4 Self-play RL training

We use standard PPO training parameters as available in stable_baselines3 with batch size 1024, number of epochs 5, learning rate 0.0005, $\gamma = 0.99$ and 8 environments in parallel. The observation used is the same as the joint state input used for our work for fair comparison. The reward design used is also the same as the utility used in our work. We train for $100K$ time-steps for each iteration of self-play RL where we switch agents for training for total of 99 times i.e. 33 cycles of training for 3 agents

B.5 Iterated Best Response (IBR) hyperparameters

We use 6 iterations for iterated best response with the same utility as the one used in our work and with horizon length of $2s$ with 20 time-steps of length $\Delta t = 0.1s$. The solve time with the following parameters is $0.1s$ which is comparable to the compute time required by our algorithm.

Appendix C

Appendix for Chapter 4

This chapter is organized as follows. In Section C.1 we review the theory of two-timescale asynchronous stochastic approximation from [337]. In Section C.2 we present the proofs of technical lemmas presented in Section 4.2.

C.1 Review of Two-timescale asynchronous stochastic approximation

In this section we review the results from [337] on the theory of two-timescale asynchronous stochastic approximation. Note that we do not state their results in full generality but only to the extent necessary for this chapter.

Let $\{x^t\}_{t=1}^{\infty}, \{y^t\}_{t=1}^{\infty}$ be the stochastic approximation updates. Let $x^t \in \mathbb{R}^X$, $y^t \in \mathbb{R}^Y$ for all $t \in \{1, 2, ..\}$. Let $\bar{X} \subset [X]$ (resp. $\bar{Y} \subset [Y]$) be the elements of x update (resp. y update) that have positive probability of being updated in the asynchronous update process. At iterate t, let $\bar{X}^t \subset \bar{X}$ and $\bar{Y}^t \subset \bar{Y}$ be the elements that are updated. Let

$$\tilde{n}^t(\mathfrak{i}) = \sum_{p=1}^{t} \mathbb{1}(\mathfrak{i} \in \bar{X}^p), \quad n^t(\mathfrak{j}) = \sum_{p=1}^{t} \mathbb{1}(\mathfrak{j} \in \bar{Y}^p),$$

for every $\mathfrak{i} \in [X]$ and $\mathfrak{j} \in [Y]$. Consider the following asynchronous stochastic approximation updates indexed by $t \in \{1, 2, ..\}$

$$\begin{aligned} x^t(\mathfrak{i}) &\in x^{t-1}(\mathfrak{i}) + \alpha_{\mathfrak{i}}(\tilde{n}^t(\mathfrak{i}))\mathbb{1}(\mathfrak{i} \in \bar{X}^t)[F(\mathfrak{i}; x^{t-1}, y^{t-1}) + \tilde{M}^t(\mathfrak{i}) + d^t(\mathfrak{i})], \quad \forall\, \mathfrak{i} \in [X] \\ y^t(\mathfrak{j}) &\in y^{t-1}(\mathfrak{j}) + \beta_{\mathfrak{j}}(n^t(\mathfrak{j}))\mathbb{1}(\mathfrak{j} \in \bar{Y}^t)[G(\mathfrak{j}; x^{t-1}, y^{t-1}) + M^t(\mathfrak{j}) + e^t(\mathfrak{j})], \quad \forall\, \mathfrak{j} \in [Y], \end{aligned} \tag{C.1}$$

where

(i) for any $x \in \mathbb{R}^X, y \in \mathbb{R}^Y$, $F(x, y) = (F(\mathfrak{i}; x, y))_{\mathfrak{i} \in [X]} \subset \mathbb{R}^X$ and

$$G(x, y) = (G(\mathfrak{j}; x, y))_{\mathfrak{j} \in [Y]} \subset \mathbb{R}^Y$$

are set-valued maps;

(ii) $\{\tilde{M}^t = (\tilde{M}^t(\mathfrak{i}))_{\mathfrak{i}\in[X]}\}, \{M^t = (M^t(\mathfrak{j}))_{\mathfrak{j}\in[Y]}\}$ be martingale difference processes defined on $\mathbb{R}^X, \mathbb{R}^Y$ respectively;

(iii) $\{d^t = (d^t(\mathfrak{i}))_{\mathfrak{i}\in\mathbb{R}^X}, e^t = (e^t(\mathfrak{j}))_{\mathfrak{j}\in\mathbb{R}^Y}\}$ are asymptotically negligible error terms;

(iv) For every $\mathfrak{i} \in [X], \mathfrak{j} \in [Y]$, $\{\alpha_\mathfrak{i}(n)\}_{n=0}^\infty, \{\beta_\mathfrak{j}(n)\}_{n=0}^\infty$ are the step sizes;

(v) $x^0 \in \mathbb{R}^X, y^0 \in \mathbb{R}^Y$ are initialized at some values.

For every $t \in \{1, 2, ...\}$, define

$$\begin{aligned}
\bar{\alpha}^t &= \max_{\mathfrak{i}\in\bar{X}^t} \alpha_\mathfrak{i}(\tilde{n}^t(\mathfrak{i})), \quad \mu^t(\mathfrak{i}) = \frac{\alpha_\mathfrak{i}(\tilde{n}^t(\mathfrak{i}))}{\bar{\alpha}^t}\mathbb{1}(\mathfrak{i} \in \bar{X}^t) \\
\bar{\beta}^t &= \max_{\mathfrak{j}\in\bar{Y}^t} \beta_\mathfrak{j}(\tilde{n}^t(\mathfrak{j})), \quad \sigma^t(\mathfrak{j}) = \frac{\beta_\mathfrak{j}(\tilde{n}^t(\mathfrak{j}))}{\bar{\beta}^t}\mathbb{1}(\mathfrak{j} \in \bar{Y}^t) \\
\tilde{\mathcal{D}}^t &= \mathsf{diag}([\mu^t(1), \mu^t(2), ..., \mu^t(X)]), \\
\mathcal{D}^t &= \mathsf{diag}([\sigma^t(1), \sigma^t(2), ..., \sigma^t(Y)]).
\end{aligned}$$

Using these notations we can concisely write (C.1) as

$$\begin{aligned}
x^t &\in x^{t-1} + \bar{\alpha}^t\tilde{\mathcal{D}}^t\left(F(x^{t-1}, y^{t-1}) + \tilde{M}^t + d^t\right) \\
y^t &\in y^{t-1} + \bar{\beta}^t\mathcal{D}^t\left(G(x^{t-1}, y^{t-1}) + M^t + e^t\right).
\end{aligned} \tag{C.2}$$

We now state some assumption that are crucial to study the asymptotic property of the stochastic approximation (C.2). First, we introduce some important notations. Define $\bar{H} \subset \bar{X} \times \bar{Y}$ such that if $i \in \bar{X}, j \in \bar{Y}$ then $(i, j) \in \bar{H}$ if and only if i, j have positive probability of occurring simultaneously. At iterate t, $\bar{H}^t \in \bar{H}$ be the updated component in $[X] \times [Y]$. Furthermore $z^t = (x^t, y^t)$ be the joint update. Let

$$\mathcal{F}^t = \sigma(\{\bar{H}^m\}_m, \{z^m\}_m, \{\tilde{n}^m(\mathfrak{i})\}, \{n^m(\mathfrak{j})\}\ \forall\ m \leqslant t, \mathfrak{i} \in [X], \mathfrak{j} \in [Y])$$

be sigma-algebra containing all information upto iterate t. For any positive integer K and a positive scalar η, define $\Omega_K^\eta = \{\mathsf{diag}(\omega(1), ..., \omega(K)) : \omega(i) \in [\eta, 1]\ \forall i = 1, 2, .., K\}$.

Next, we present the assumptions required in [337] to study the asymptotic behavior of two-timescale asynchronous stochastic approximation update (C.1).

Assumption C.1.1. *Let the following assumptions hold*

(A1) For compact sets $\tilde{\mathcal{S}} \subset \mathbb{R}^X, \mathcal{S} \subset \mathbb{R}^Y$, $x^t \in \tilde{\mathcal{S}}, y^t \in \mathcal{S}$ for all $t \in \{0, 1, ...\}$.

(A2) $\{d^t\}, \{e^t\}$ are bounded sequence such that $\lim_{t\to\infty} d^t = \lim_{t\to\infty} e^t = 0$.

(A3) Following must be true for the stepsizes:

(i) For every $\mathfrak{i} \in [X], \mathfrak{j} \in [Y]$, $\sum_n \alpha_\mathfrak{i}(n) = \infty, \sum_n \beta_\mathfrak{j}(n) = \infty$, $\lim_{n\to\infty} \alpha_\mathfrak{i}(n) = \lim_{n\to\infty} \beta_\mathfrak{j}(n) = 0$ and $\{\alpha_\mathfrak{i}(n)\}, \{\beta_\mathfrak{j}(n)\}$ are non-increasing sequences.

(ii) *For any* $\lambda \in (0,1)$, $\mathfrak{i} \in [X]$ *and* $\mathfrak{j} \in [Y]$ *it holds that* $\sup_n \alpha_{\mathfrak{i}}([\lambda n])/\alpha_{\mathfrak{i}}(n) < A_\lambda < \infty$, $\sup_n \beta_{\mathfrak{j}}([\lambda n])/\beta_{\mathfrak{j}}(n) < A_\lambda < \infty$.

(iii) *For every* $\mathfrak{i} \in [X], \mathfrak{j} \in [Y]$ *it holds that* $\lim_{n\to\infty} \beta_{\mathfrak{j}}(n)/\alpha_{\mathfrak{i}}(n) = 0$.

(iv) *For every* $\mathfrak{i}, \mathfrak{i}' \in [X]$, $\mathfrak{j}, \mathfrak{j}' \in [Y]$, *there exists* $0 < \xi^{\alpha}_{\mathfrak{i}\mathfrak{i}'} < \zeta^{\alpha}_{\mathfrak{i}\mathfrak{i}'} < \infty$, *and* $0 < \xi^{\beta}_{\mathfrak{j}\mathfrak{j}'} < \zeta^{\beta}_{\mathfrak{j}\mathfrak{j}'} < \infty$ *such that* $\frac{\alpha_{\mathfrak{i}}(n)}{\alpha_{\mathfrak{i}'}(n)} \in [\xi^{\alpha}_{\mathfrak{i}\mathfrak{i}'}, \zeta^{\alpha}_{\mathfrak{i}\mathfrak{i}'}]$ *and* $\frac{\beta_{\mathfrak{j}}(n)}{\beta_{\mathfrak{j}'}(n)} \in [\xi^{\beta}_{\mathfrak{j}\mathfrak{j}'}, \zeta^{\beta}_{\mathfrak{j}\mathfrak{j}'}]$ *for all* n.

(A4) *The maps* $F(\cdot,\cdot), G(\cdot,\cdot)$ *are such that*

(i) $F : \tilde{\mathcal{S}} \times \mathcal{S} \rightrightarrows \mathcal{S}$ *is upper semi-continuous, for every* $z \in \tilde{\mathcal{S}} \times \mathcal{S}$, $F(z)$ *is non-empty, compact, convex subset of* $\mathcal{S}$, *and* $\sup_{t\in F(z)} \|t\| \leqslant c(1 + \|z\|)$ *where* c *is a constant independent of* z.

(ii) $G : \tilde{\mathcal{S}} \times \mathcal{S} \rightrightarrows \tilde{\mathcal{S}}$ *is upper semi-continuous.* $G(x,\cdot)$ *is non-empty, convex and compact and satisfy* $\sup_{t\in G(x,y)} \|t\| \leqslant c(1 + \|y\|)$ *where* c *is a constant independent of* x, y.

(A5) *(i)* *for all* $z \in \tilde{\mathcal{S}} \times \mathcal{S}$ *and* $h^{t-1}, h^t \in \bar{H}$,

$$\mathcal{P}\left(\bar{H}^t = h^t | \mathcal{F}^{t-1}\right) = \mathcal{P}(\bar{H}^t = h^t | \bar{H}^{t-1} = h^{t-1}, z^{t-1} = z)$$

(ii) *For any* $z \in \tilde{\mathcal{S}} \times \mathcal{S}$ *the transition probability*

$$\mathcal{P}(z; h^t, h^{t-1}) := \mathcal{P}(\bar{H}^t = h^t | \bar{H}^{t-1} = h^{t-1}, z^{t-1} = z) \tag{C.3}$$

form aperiodic and irreducible Markov chain over $\bar{H}$ *and for every* $\mathfrak{i} \in X$ *and* $\mathfrak{j} \in Y$ *there exists* $h, h' \in \bar{H}$ *such that* $\mathfrak{i} \in h$ *and* $\mathfrak{j} \in h'$.

(iii) *the map* $z \mapsto \mathcal{P}(z; h^t, h^{t-1})$ *is Lipschitz.*

(A6) *For some* $q \geqslant 2$, $\sum_n \alpha_{\mathfrak{i}}(n)^{1+q/2} < \infty$ *and* $\sup_t \mathbb{E}\left[\|\tilde{M}^t\|^q\right] < \infty$ *for every* $\mathfrak{i} \in [X]$. *For some* $q' \geqslant 2$, $\sum_n \beta_{\mathfrak{j}}(n)^{1+q'/2} < \infty$ *and* $\sup_t \mathbb{E}\left[\|M^t\|^q\right] < \infty$ *for every* $\mathfrak{j} \in [Y]$.

(A7) *For all* $y \in \mathcal{S}$ *and every* $\phi > 0$ *the differential inclusion*

$$\frac{d}{d\tau} x^\tau \in \Omega^{\phi}_X \cdot F(x^\tau, y),$$

has unique global attractor $\Lambda(y)$, *where* $\Lambda : \mathbb{R}^Y \to \mathbb{R}^X$ *is bounded, continuous and single-valued for all* $y \in \mathcal{S}$.

Theorem C.1.1 (Fast-timescale convergence)**.** *[337, Corollary 4.4] Under assumption (A1)-(A7) in Assumption C.1.1, with probability 1,*

$$(x^t, y^t) \to \{(\Lambda(y), y) : y \in \mathcal{S}\} \quad as \quad t \to \infty.$$

Next, we present the corresponding convergence results for the slow updates, $\{y^t\}$. Prior to that, we define the linearly interpolated trajectory $\{\bar{y}^\tau\}_{\tau\in\mathbb{R}_+}$ defined as

$$\bar{y}^{\bar{\tau}^t+s} = y^t + s\frac{y^{t+1}-y^t}{\bar{\beta}^{t+1}}, \quad s\in[0,\bar{\beta}^{t+1}),$$

where $\bar{\tau}^t = \sum_{p=0}^{t}\bar{\beta}^t$.

Define $G^\Lambda : \mathbb{R}^Y \to \mathbb{R}^Y$ as $G^\Lambda(y) = G(\Lambda(y), y)$. Furthermore, let $\bar{G}^{\Lambda,\eta}(y) = \Omega_Y^\eta G^\Lambda(y)$, where $\eta = \kappa/A_\kappa$ for some $0 < \kappa \leqslant \min_{z\in\tilde{\mathcal{S}}\times\mathcal{S},\mathfrak{j}\in[Y]}\psi_z(\mathfrak{j})$, and $\psi_z \in \Delta(Y)$ is the marginal of the stationary distribution of the Markov chain on $\bar{H}$ with transition kernel $\mathcal{P}(z;h,h')$ [337, Lemma A.1, Appendix A.3]. Consider the following differential inclusion

$$\dot{y}^\tau \in \bar{G}^{\Lambda,\eta}(y^\tau). \tag{C.4}$$

Theorem C.1.2 (Slow-timescale convergence)**.** *[337, Theorem 4.7] If the conditions (A1)-(A7) in Assumption C.1.1 are satisfied then $\{\bar{y}^\tau\}_{\tau\in\mathbb{R}_+}$ is an asymptotic pseudo-trajectory to* (C.4)*.*

Remark C.1.1. *Note that [337] assume that for every $\mathfrak{i},\mathfrak{i}' \in [X], \mathfrak{j},\mathfrak{j}' \in [Y]$ it holds that $\alpha_{\mathfrak{i}}(\cdot) = \alpha_{\mathfrak{i}'}(\cdot)$ and $\beta_{\mathfrak{j}}(\cdot) = \beta_{\mathfrak{j}'}(\cdot)$. However, they easily generalize under the setting of heterogeneous step sizes considered here due to Assumption (A3)-(iv). Indeed, Theorem C.1.1 (resp. Theorem C.1.2) follow similar to [337] if we fix a $\tilde{\mathfrak{i}} \in [X]$ (resp. $\tilde{\mathfrak{j}} \in [Y]$) and bound the relative evolution of step sizes at fast (resp. slow) timescale $\mathfrak{i} \neq \tilde{\mathfrak{i}}$ (resp. $\mathfrak{j} \neq \tilde{\mathfrak{j}}$ with respect to $\tilde{\mathfrak{i}}, \tilde{\mathfrak{j}}$ using Assumption (A3)-(iv).*

C.2 Remaining Proofs

For clear presentation, we define $Q_i(s,\pi_i';\pi) = \pi_i'(s)^\top Q_i(s;\pi)$. Recall, for any $\pi\in\Pi, \theta \in (0,1)^{|I|}$, we define $\pi^{(\theta)} \in \Pi$ such that for every $s\in S, i\in I$, $\pi_i^{(\theta)}(s) := (1-\theta_i)\pi_i(s) + \theta_i(1/|A_i|)\cdot\mathbb{1}_{A_i}$ to be a perturbed version of policy π due to exploration parameter θ.

Proof of Lemma 4.2.1

The proof follows by verifying that Assumption C.1.1 (A1)-(A7) are satisfied and then evoking Theorem C.1.1. Towards that goal, we first verify Assumption C.1.1 (A1)-(A7).

Before verifying the conditions for two-timescale asynchronous stochastic approximation stated in Section C.1, we introduce some notations. For any $\pi\in\Pi$, $i\in I, a_i\in A_i, s\in S$, we define

$$\mathcal{T}_i^\pi \tilde{q}_i(s,a_i) \coloneqq u_i(s,a_i,\pi_{-i}) + \gamma\sum_{s'} P(s'|s,a_i,\pi_{-i})\pi_i(s')^\top \tilde{q}_i(s'), \tag{C.5}$$

which is analogous to Bellman operator in the setup of this chapter. Furthermore, for any $\pi \in \Pi$, we define

$$\hat{\mathcal{T}}_i^{\pi} \tilde{q}_i(s, a_i) := u_i(s, a_i, a_{-i}) + \gamma \pi_i(s')^{\top} \tilde{q}_i(s'), \tag{C.6}$$

where $a_{-i} \sim \pi_{-i}(s)$ and $s' \sim P(\cdot|s, a_i, \pi_{-i})$. Moreover, for any $s \in S, i \in I, a_i \in A_i$, we define

$$\bar{\text{br}}_i(s; q) = \arg\max_{\pi \in \Delta(A_i)} \pi^{\top} q_i(s), \tag{C.7}$$

where for every $i \in I, s \in S$, $q_i(s) \in \mathbb{R}^{|A_i|}$. Using the above notations, we re-write (4.4)-(4.5) as

$$\tilde{q}_i^t(s, a_i) = \tilde{q}_i^{t-1}(s, a_i) + \alpha_i(\tilde{n}_i^t(s, a_i)) \mathbb{1}\{(s, a_i) = (s^{t-1}, a_i^{t-1})\} \tag{C.8a}$$

$$\cdot (\mathcal{T}_i^{(\pi_i^{t-1}, \pi_{-i}^{t-1,(\theta)})} \tilde{q}_i^{t-1}(s, a_i) - \tilde{q}_i^{t-1}(s, a_i) + \tilde{M}_i^t(s, a_i)), \tag{C.8b}$$

$$\pi_i^t(s) \in \pi_i^{t-1}(s) + \beta_i(n^t(s)) \mathbb{1}\{s = s^{t-1}\} \left(\bar{\text{br}}_i(s; q^{t-1}) - \pi_i^{t-1}(s) \right), \tag{C.8c}$$

for all $(s, a_i) \in S \times A_i$, $i \in I$, where

$$\tilde{M}_i^t(s, a_i) = \hat{\mathcal{T}}_i^{(\pi_i^{t-1}, \pi_{-i}^{t-1,(\theta)})} \tilde{q}_i^{t-1}(s, a_i) - \mathcal{T}_i^{(\pi_i^{t-1}, \pi_{-i}^{t-1,(\theta)})} \tilde{q}_i^{t-1}(s, a_i). \tag{C.9}$$

Note that $\mathbb{E}[\tilde{M}_i^t(s, a_i)|\mathcal{F}^{t-1}] = 0$ where $\mathcal{F}^{t-1} = \sigma(\{(s^m, a^m)\}_m, \{\tilde{q}_i^m\}_m, \{\pi_i^m\}_m : m \leqslant t-1, i \in I)$ is the sigma-algebra comprising of history till stage $t-1$. Consequently, $\{\tilde{M}_i^t\}$ is a martingale difference sequence. The updates (C.8a)-(C.8c) are now cast in the same formulation as in (C.1). The asynchronous q-estimate updates (C.8a) and the policy updates (C.8c) both have $|\Pi_{i=1}^I (S \times A_i)|$ components.

We now verify Assumption C.1.1 (A1)-(A7) one by one

(i) First, we show that (A1) in Assumption C.1.1 is satisfied with $(\tilde{q}^t, \pi^t)$ update (C.8a)-(C.8c). Let $\bar{u} = \max_{i,s,a} |u_i(s, a)|$. Moreover let $\bar{R} = \max\{\bar{u}/(1-\gamma), \max_i \|\tilde{q}_i^0\|_\infty\}$. Then we claim that $\|\tilde{q}_i^t\|_\infty \leqslant \bar{R}$ for all $t = \{0, 1, 2, ..\}$. We show this by induction. It holds for $t = 0$ by construction. Suppose it holds for $t = m-1$ for some m then we show that it also holds for $t = m$. Indeed, we note from (C.8a) that $\tilde{q}_i^t$ is a convex combination[1] of $\tilde{q}_i^{t-1}$ and $\mathcal{T}_i^{(\pi_i^{t-1}, \pi_{-i}^{t-1,(\theta)})} \tilde{q}_i^{t-1}(s, a_i) + \tilde{M}_i^t(s, a_i)$. Using (C.6) and (C.9) we see that

$$\begin{aligned} &\|\mathcal{T}_i^{(\pi_i^{t-1}, \pi_{-i}^{t-1,(\theta)})} \tilde{q}_i^{t-1} + \tilde{M}_i^t\|_\infty \\ &= \|\hat{\mathcal{T}}_i^{(\pi_i^{t-1}, \pi_{-i}^{t-1,(\theta)})} \tilde{q}_i^{t-1}\|_\infty \\ &\leqslant \bar{u} + \delta \bar{R} \leqslant (1-\gamma)\bar{R} + \gamma \bar{R} = \bar{R}. \end{aligned}$$

[1]This is because we assume that $\alpha(n) \in (0, 1)$ in Assumption 4.2.2.

This shows that $\|\tilde{q}_i^t\|_\infty \leqslant \bar{R}$. Moreover note that $\pi^t \in \Pi$ which is product simplex and is always compact.
(ii) Since we do not have any asymptotically negligible error terms in the asynchronous updates, AssumptionC.1.1-(A2) is immediately satisfied
(iii) Next we note Assumption C.1.1-(A3) is satisfied due to Assumption 4.2.2.
(iv) Now we show Assumption C.1.1-(A4) is satisfied. First, we concisely write the mean fields of (C.8a)-(C.8c) as follows

$$\begin{aligned} F((s,a_i);\tilde{q},\pi,\theta) &\coloneqq \mathcal{T}_i^{(\pi_i,\pi_{-i}^{(\theta)})}\tilde{q}_i(s,a_i) - \tilde{q}_i(s,a_i), \\ G((s,a_i);\tilde{q},\pi) &= \bar{\text{br}}_i(s,a_i;q) - \pi_i(s,a_i), \end{aligned}$$

for every $s \in S$, $i \in I$, $a_i \in A_i$. Define $F(\tilde{q},\pi,\theta) = (F((s,a_i);\tilde{q},\pi,\theta))_{s\in S,i\in I,a_i\in A_i}$, $G(\tilde{q},\pi) = (G((s,a_i);\tilde{q},\pi))_{s\in S,i\in I,a_i\in A_i}$. We note that both F, G are continuous as demanded in Assumption C.1.1-(A4). Furthermore, observe that

$$\begin{aligned} \|F(\tilde{q},\pi,\theta)\|_\infty &\leqslant \|\mathcal{T}_i^{(\pi_i,\pi_{-i}^{(\theta)})}\tilde{q}_i\|_\infty + \|\tilde{q}\|_\infty \\ &\leqslant \bar{u} + \delta\|\tilde{q}\|_\infty + \|\tilde{q}\|_\infty \leqslant \tilde{c}(1+\|\tilde{q}\|_\infty), \end{aligned}$$

where $\tilde{c} = \max\{\bar{u}, 1+\delta\}$. Also note that $\sup_{w\in G(\tilde{q},\pi)}\|w\|_\infty \leqslant 1 + \|\pi\|_\infty$. Thus we conclude that Assumption C.1.1-(A4) is satisfied.
(v) We now verity Assumption C.1.1-(A5). Consider $h, h' \in \bar{H}$ such that

$$h = ((s,a_1),(s,a_2),..(s,a_I)), \quad h' = ((s',a_1'),(s,a_2'),...(s',a_I')).$$

Moreover, let $z = (\tilde{q},\pi)$ then,

$$\mathcal{P}(z;h,h') = P(s'|s,a)\prod_{i\in I}\pi_i^{(\theta)}(s',a_i'), \tag{C.10}$$

where $a = (a_i)_{i\in I}$ and the function $\mathcal{P}(z;h,h')$ is defined in (C.3). Since $\theta > 0$, all actions have positive probability of being selected. That is, for every $k \in \mathbb{N}$ we have $\pi_i^{t,(\theta)}(s,a_i) > \theta/|A_i|$ for all $s \in S, i \in I, a_i \in A_i$. Moreover, we impose Assumption 4.2.1 on transition matrix which ensures that every state is visited with some non-zero probability. Thus, Assumption C.1.1(A5)-(i) and C.1.1(A5)-(ii) are satisfied. Finally Assumption C.1.1(A5)-(iii) is satisfied by noting that (C.10) is Lipschitz in π and therefore in z.
(vi) Assumption C.1.1-(A6) is satisfied by noting that (a) $\tilde{M}$ is a bounded martingale difference sequence and (b) the step size condition in Assumption 4.2.2-(ii) holds.
(vii) For any $\phi > 0$, $\pi \in \Pi$ consider the differential equation

$$\frac{d}{d\tau}\tilde{q}_i^\tau = \Omega_{A_i}^\phi\left(\mathcal{T}_i^{(\pi_i,\pi_{-i}^{(\theta)})}\tilde{q}_i^\tau - \tilde{q}_i^\tau\right), \quad \forall\ i \in I, \tag{C.11}$$

where $\Omega_{A_i}^{\phi} = \{\mathsf{diag}(\omega(1), ..., \omega(|A_i|)) : \omega(k) \in [\phi, 1] \ \forall k = 1, 2, .., |A_i|\}$. In order to verify Assumption C.1.1-(A7), we show that (C.11) has unique global attractor for every $\pi \in \Pi$.

We show that $\mathcal{T}_i^{\pi}$ is a contraction for every $\pi \in \Pi$. For any $q, \bar{q}$,

$$\begin{aligned}&\mathcal{T}_i^{\pi} q_i(s, a_i) - \mathcal{T}_i^{\pi} \bar{q}_i(s, a_i) \\ =&\gamma \sum_{s'} P(s'|s, a_i, \pi_{-i}) \sum_{a_i' \in A_i} \pi_i(s', a_i') \left(q_i(s', a_i') - \bar{q}_i(s', a_i') \right).\end{aligned}$$

Thus, for every $s \in S, i \in I, a_i \in A_i$, we have $|\mathcal{T}_i^{\pi} q_i(s, a_i) - \mathcal{T}_i^{\pi} \bar{q}_i(s, a_i)| \leqslant \gamma \|q_i - \bar{q}_i\|_{\infty}$. Consequently, $\|\mathcal{T}_i^{\pi} q_i - \mathcal{T}_i^{\pi} \bar{q}_i\|_{\infty} \leqslant \gamma \|q_i - \bar{q}_i\|_{\infty}$, and $\mathcal{T}_i^{\pi}$ is a contraction mapping.

Consequently, $\mathcal{T}_i^{(\pi_i, \pi_{-i}^{(\theta)})}$ is a contraction and (C.11) has a unique global attractor, which is the fixed point of the mapping $\mathcal{T}_i^{(\pi_i, \pi_{-i}^{(\theta)})}$ [60, Chapter 7.4]. Moreover, from the definition it follows that $Q_i(\cdot, \cdot; (\pi_i, \pi_{-i}^{(\theta)}))$ is the fixed point of $\mathcal{T}_i^{(\pi_i, \pi_{-i}^{(\theta)})}$. That is, for every $s \in S, i \in I, a_i \in A_i$, $\mathcal{T}_i^{(\pi_i, \pi_{-i}^{(\theta)})} Q_i(s, a_i; (\pi_i, \pi_{-i}^{(\theta)})) = Q_i(s, a_i; (\pi_i, \pi_{-i}^{(\theta)}))$. Additionally, $\|Q_i(\cdot, \cdot; (\pi_i, \pi_{-i}^{(\theta)}))\|_{\infty} \leqslant \frac{u_{\max}}{1-\gamma}$, and $Q_i(\cdot, \cdot; \pi)$ is also continuous in π. Thus, Assumption C.1.1-(A7) is satisfied. Finally, the claim in Lemma 4.2.1 follows by Theorem C.1.1.

Proof of Lemma 4.2.3

We prove (a)-(d) in sequence

(a) We claim that for any integer $K \geqslant 0$, $\mu \in \gamma(S), \pi \in \Pi, s \in S, i \in I, a_i \in A_i$,

$$\frac{\partial V_i(\mu, \pi)}{\partial \pi_i(s, a_i)} = \mathbb{E}\left[\sum_{k=0}^{K} \gamma^k \mathbb{1}(s^k = s) \right] Q_i(s, a_i; \pi) + \gamma^{K+1} \mathbb{E}\left[\frac{\partial V_i(s^{K+1}, \pi)}{\partial \pi_i(s, a_i)} \right], \tag{C.12}$$

where $s_0 \sim \mu, a^{k-1} \sim \pi(s^{k-1}), s^k \sim P(\cdot|s^{k-1}, a^{k-1})$. We prove this claim by induction.

Indeed, this holds for $K = 0$ by noting that

$$\begin{aligned}
&\frac{\partial V_i(\mu,\pi)}{\partial \pi_i(s,a_i)} = \frac{\partial}{\partial \pi_i(s,a_i)} \sum_{\bar{s}\in S} \mu(\bar{s}) \left(\sum_{\bar{a}_i \in A_i} \pi_i(\bar{s},\bar{a}_i) Q_i(\bar{s},\bar{a}_i;\pi) \right) \\
=&\frac{\partial}{\partial \pi_i(s,a_i)} \left(\sum_{\bar{s}} \mu(\bar{s}) \sum_{\bar{a}_i} \pi_i(\bar{s},\bar{a}_i) \left(u_i(\bar{s},\bar{a}_i,\pi_{-i}(\bar{s})) + \gamma \sum_{s'} P(s'|\bar{s},\bar{a}_i,\pi_{-i}(\bar{s})) V_i(s',\pi) \right) \right) \\
=&\mu(s) \left(u_i(s,a_i,\pi_{-i}(s)) + \gamma \sum_{s'} P(s'|s,a_i,\pi_{-i}(s)) V_i(s',\pi) \right) \\
&\quad + \gamma \sum_{\bar{s}} \mu(\bar{s}) \sum_{s'} P(s'|\bar{s},\pi) \frac{\partial V_i(s',\pi)}{\partial \pi_i(s,a_i)} \\
=&\mu(s) Q_i(s,a_i;\pi) + \gamma \sum_{\bar{s}} \mu(\bar{s}) \sum_{s'} P(s'|\bar{s},\pi) \frac{\partial V_i(s',\pi)}{\partial \pi_i(s,a_i)} \\
=&\mathbb{E}\left[\mathbb{1}(s^0 = s) \right] Q_i(s,a_i;\pi) + \gamma \mathbb{E}\left[\frac{\partial V_i(s^1,\pi)}{\partial \pi_i(s,a_i)} \right]
\end{aligned}$$

We now suppose that the claim holds for some integer K and then show that it holds for $K+1$, that is we have

$$\begin{aligned}
&\frac{\partial V_i(\mu,\pi)}{\partial \pi_i(s,a_i)} = \mathbb{E}\left[\sum_{k=0}^{K} \gamma^k \mathbb{1}(s^k = s) \right] Q_i(s,a_i;\pi) + \gamma^{K+1} \mathbb{E}\left[\frac{\partial V_i(s^{K+1},\pi)}{\partial \pi_i(s,a_i)} \right] \\
&= \mathbb{E}\left[\sum_{k=0}^{K} \gamma^k \mathbb{1}(s^k = s) \right] Q_i(s,a_i;\pi) + \gamma^{K+1} \mathbb{E}\left[\frac{\partial}{\partial \pi_i(s,a_i)} \left(\sum_{a_i} \pi_i(s^{K+1},a_i) Q_i(s^{K+1},a_i;\pi) \right) \right] \\
&= \mathbb{E}\left[\sum_{k=0}^{K} \gamma^k \mathbb{1}(s^k = s) \right] Q_i(s,a_i;\pi) \\
&\quad + \gamma^{K+1} \left(\mathbb{1}(s^{K+1} = s) \cdot Q_i(s,a_i;\pi) + \gamma \left(\sum_{s'} P(s'|s^{K+1},\pi) \frac{\partial V_i(s',\pi)}{\partial \pi_i(s,a_i)} \right) \right) \\
&= \mathbb{E}\left[\sum_{k=0}^{K+1} \gamma^k \mathbb{1}(s^k = s) \right] Q_i(s,a_i;\pi) + \gamma^{K+2} \mathbb{E}\left[\frac{\partial V_i(s^{K+2},\pi)}{\partial \pi_i(s,a_i)} \right].
\end{aligned}$$

This completes the proof of (C.12). Now if we let $K \to \infty$ in (C.12) then we obtain

$$\begin{aligned}
\frac{\partial V_i(\mu,\pi)}{\partial \pi_i(s,a_i)} &= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \mathbb{1}(s^k = s) \right] Q_i(s,a_i;\pi) \\
&= \sum_{s^0 \in S} \mu(s^0) \sum_{k=0}^{\infty} \mathsf{Pr}(s^k = s|s^0) Q_i(s,a_i;\pi) \\
&= \frac{1}{1-\gamma} d_\mu^\pi(s) Q_i(s,a_i;\pi).
\end{aligned}$$

(b) For any initial state distribution μ and joint policy $\pi = (\pi_i, \pi_{-i}), \pi' = (\pi'_i, \pi_{-i}) \in \Pi$,

$$\begin{aligned}
&V_i(\mu, \pi) - V_i(\mu, \pi') \\
&= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k u_i(s^k, a^k)\right] - V_i(\mu, \pi') \\
=&\mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \Big(u_i(s^k, a^k) - V_i(s^k, \pi') + V_i(s^k, \pi')\Big)\right] - V_i(\mu, \pi') \\
=&\mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \left(u_i(s^k, a^k) - V_i(s^k, \pi')\right)\right] \\
&+ \mathbb{E}\left[V_i(s^0, \pi')\right] + \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^k V_i(s^k, \pi')\right] - V_i(\mu, \pi'),
\end{aligned}$$

where $s^0 \sim \mu, a^{k-1} \sim \pi(s^{k-1}), s^k \sim P(\cdot|s^{k-1}, a^{k-1})$. We note that

$$\begin{aligned}
&\mathbb{E}\left[V_i(s^0, \pi')\right] = V_i(\mu, \pi'), \\
&\mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^k V_i(s^k, \pi')\right] = \gamma \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k V_i(s^{k+1}, \pi')\right].
\end{aligned}$$

Therefore,

$$\begin{aligned}
&V_i(\mu, \pi) - V_i(\mu, \pi') \\
&= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \left(u_i(s^k, a^k) - V_i(s^k, \pi')\right)\right] + \gamma \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k V_i(s^{k+1}, \pi')\right] \\
&= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \Big(u_i(s^k, a^k) - V_i(s^k, \pi') + \gamma V_i(s^{k+1}, \pi')\Big)\right] \\
&= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \Big(u_i(s^k, a^k) + \gamma V_i(s^{k+1}, \pi') - V_i(s^k, \pi')\Big)\right] \\
&= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \Big(u_i(s^k, a^k) + \gamma \sum_{s'} P(s'|s^k, a^k) V_i(s', \pi') - V_i(s^k, \pi')\Big)\right].
\end{aligned}$$

Thus, we conclude that

$$\begin{aligned}
&V_i(\mu,\pi) - V_i(\mu,\pi') \\
&= \mathbb{E}\Bigg[\sum_{k=0}^{\infty}\gamma^k\Big(Q_i(s^k,a_i^k;\pi') - V_i(s^k,\pi')\Big)\Bigg] \\
&= \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k\Gamma_i(s^k,a_i^k;\pi')\right] \\
&= \frac{1}{1-\gamma}\sum_{s'} d_\mu^\pi(s')\Gamma_i(s',\pi_i;\pi').
\end{aligned}$$

(c) For every $s \in S$, define $\gamma V_i(s,\pi) := V_i(s,\pi_i,\pi_{-i}^{(\theta)}) - V_i(s,\pi_i,\pi_{-i})$. Since $V_i(s,\pi_i,\pi_{-i}) = u_i(s,\pi_i,\pi_{-i}) + \gamma\sum_{s'\in S}P(s'|s,\pi_i,\pi_{-i})V_i(s',\pi_i,\pi_{-i})$, we note that

$$\begin{aligned}
&\gamma V_i(s,\pi) = u_i(s,\pi_i,\pi_{-i}^{(\theta)}) - u_i(s,\pi_i,\pi_{-i}) \\
&+\gamma\sum_{s'\in S}P(s'|s,\pi_i,\pi_{-i}^{(\theta)})V_i(s',\pi_i,\pi_{-i}^{(\theta)}) \\
&-\gamma\sum_{s'\in S}P(s'|s,\pi_i,\pi_{-i})V_i(s',\pi_i,\pi_{-i}) \\
&= u_i(s,\pi_i,\pi_{-i}^{(\theta)}) - u_i(s,\pi_i,\pi_{-i}) \\
&+\gamma\sum_{s'\in S}\Big(P(s'|s,\pi_i,\pi_{-i}^{(\theta)}) - P(s'|s,\pi_i,\pi_{-i})\Big)V_i(s',\pi_i,\pi_{-i}^{(\theta)}) \\
&+\gamma\sum_{s'\in S}P(s'|s,\pi_i,\pi_{-i})\Big(V_i(s',\pi_i,\pi_{-i}^{(\theta)}) - V_i(s',\pi_i,\pi_{-i})\Big),
\end{aligned}$$

where the last equality is obtained by adding and subtracting the term

$$\gamma\sum_{s'}P(s'|s,\pi)V_i(s',\pi_i,\pi_{-i}^{(\theta)}).$$

Next, we note that

$$\begin{aligned}
&|\gamma V_i(s,\pi)| \leqslant |u_i(s,\pi_i,\pi_{-i}^{(\theta)}) - u_i(s,\pi_i,\pi_{-i})| \\
&+\gamma\Bigg|\sum_{s'\in S}\Big(P(s'|s,\pi_i,\pi_{-i}^{(\theta)}) - P(s'|s,\pi_i,\pi_{-i})\Big)V_i(s',\pi_i,\pi_{-i}^{(\theta)})\Bigg| \\
&+\gamma|\gamma V_i(s',\pi)|. \qquad \text{(C.13)}
\end{aligned}$$

First, for every $s \in S, \pi \in \Pi$, we bound the term $|u_i(s,\pi_i,\pi_{-i}^{(\theta)}) - u_i(s,\pi_i,\pi_{-i})|$ in (C.13). To bound this, we define a notation, for every $i \in I$,

$$\pi_{-i}^{(\theta_{[1:k]})} := \begin{cases}\pi_j^{(\theta)} & \text{if } j \in \{1,2,...k\}\backslash\{i\},\\ \pi_j & \text{otherwise.}\end{cases}$$

If $k = 0$ (resp. $k = |I|$) then $\pi_{-i}^{(\theta_{[1:k]})}$ is π_{-i} (resp. $\pi_{-i}^{(\theta)}$). Using this notation, we obtain

$$
\begin{aligned}
&|u_i(s, \pi_i, \pi_{-i}^{(\theta)}) - u_i(s, \pi_i, \pi_{-i})| \\
&= |\sum_{a_{-i} \in A_{-i}} u_i(s, \pi_i, a_{-i})(\mathsf{Pr}^{\pi_{-i}^{(\theta)}}(a_{-i}|s) - \mathsf{Pr}^{\pi_{-i}}(a_{-i}|s))| \\
&= |\sum_{a_{-i} \in A_{-i}} u_i(s, \pi_i, a_{-i}) \\
&\quad \cdot \left(\sum_{k \in I \setminus \{i\}} (\mathsf{Pr}^{\pi_{-i}^{(\theta_{[1:k]})}}(a_{-i}|s) - \mathsf{Pr}^{\pi_{-i}^{(\theta_{[1:k-1]})}}(a_{-i}|s)) \right)| \\
&\leqslant u_{\max} \sum_{a_{-i} \in A_{-i}} \sum_{k \in I \setminus \{i\}} |((1-\theta_k)\pi_k(a_k|s)\mathsf{Pr}^{\pi_{-ik}^{(\theta_{[1:k-1]})}}(a_{-ik}|s) \\
&+ \theta_k \frac{1}{|A_k|} \mathsf{Pr}^{\pi_{-ik}^{(\theta_{[1:k-1]})}}(a_{-ik}|s) - \pi_k(a_k|s)\mathsf{Pr}^{\pi_{-ik}^{(\theta_{[1:k-1]})}}(a_{-ik}|s))| \\
&= u_{\max} \sum_{a_{-i} \in A_{-i}} \sum_{k \in I \setminus \{i\}} |(-\theta_k \pi_k(a_k|s)\mathsf{Pr}^{\pi_{-ik}^{(\theta_{[1:k-1]})}}(a_{-ik}|s) \\
&+ \theta_k \frac{1}{|A_k|} \mathsf{Pr}^{\pi_{-ik}^{(\theta_{[1:k-1]})}}(a_{-ik}|s))| \\
&\leqslant u_{\max} \sum_{k \in I \setminus \{i\}} \theta_k \sum_{a_{-i} \in A_{-i}} \mathsf{Pr}^{\pi_{-i}^{(\theta_{[1:k-1]})}}(a_{-i}|s) \\
&\qquad + u_{\max} \sum_{k \in I \setminus \{i\}} \theta_k \sum_{a_{-i} \in A_{-i}} \frac{1}{|A_k|} \mathsf{Pr}^{\pi_{-ik}^{(\theta_{[1:k-1]})}}(a_{-ik}|s) \\
&= 2u_{\max} \sum_{k \in I \setminus \{i\}} \theta_k, \qquad \text{(C.14)}
\end{aligned}
$$

where the last equality is using the fact that $\mathsf{Pr}^{\pi_{-i}^{(\theta_{[1:k-1]})}}(a_{-i}|s)$ and $\mathsf{Pr}^{\pi_{-ik}^{(\theta_{[1:k-1]})}}(a_{-ik}|s)$ are probability distribution on A_{-i} and A_{-ik} respectively.

Next, we bound the second term in (C.13). Note that

$$\begin{aligned}
&\left| \sum_{s' \in S} \Big(P(s'|s, \pi_i, \pi_{-i}^{(\theta)}) - P(s'|s, \pi_i, \pi_{-i}) \Big) V_i(s', \pi_i, \pi_{-i}^{(\theta)}) \right| \\
&\leqslant \sum_{s' \in S} \left| \Big(P(s'|s, \pi_i, \pi_{-i}^{(\theta)}) - P(s'|s, \pi_i, \pi_{-i}) \Big) \right| \bar{V}_i \\
&\leqslant \sum_{s' \in S} \left| \sum_{k \in I \setminus \{i\}} \Big(P(s'|s, \pi_i, \pi_{-i}^{(\theta_{[1:k]})}) - P(s'|s, \pi_i, \pi_{-i}^{(\theta_{[1:k-1]})}) \Big) \right| \bar{V}_i \\
&= \bar{V}_i \sum_{s' \in S} \left| \sum_{k \in I \setminus \{i\}} \Big((1 - \theta_k) P(s'|s, \pi_i, \pi_k, \pi_{-ik}^{(\theta_{[1:k-1]})}) \right. \\
&\quad \left. + \theta_k P(s'|s, \pi_i, \pi_k^\circ, \pi_{-ik}^{(\theta_{[1:k-1]})}) - P(s'|s, \pi_i, \pi_k, \pi_{-ik}^{(\theta_{[1:k-1]})}) \Big) \right| \\
&= \bar{V}_i \sum_{s' \in S} \left| \sum_{k \in I \setminus \{i\}} (-\theta_k P(s'|s, \pi_i, \pi_k, \pi_{-ik}^{(\theta_{[1:k-1]})}) \right. \\
&\quad \left. + \theta_k P(s'|s, \pi_i, \pi_k^\circ, \pi_{-ik}^{(\theta_{[1:k-1]})})) \right| \\
&\leqslant 2 \bar{V}_i \sum_{k \in I \setminus \{i\}} \theta_k, \qquad \text{(C.15)}
\end{aligned}$$

where $\bar{V}_i = \max_{s'} |V_i(s', \pi_i, \pi_{-i}^{(\theta)})|$ and the last inequality is using the fact that

$$P(s'|s, \pi_i, \pi_k, \pi_{-ik}^{(\theta_{[1:k-1]})})$$

and

$$P(s'|s, \pi_i, \pi_k^\circ, \pi_{-ik}^{(\theta_{[1:k-1]})})$$

are probability distributions on S.

Combining (C.13), (C.14), and (C.15), we obtain

$$\begin{aligned}
\max_{s,\pi} |\gamma V_i(s, \pi)| &\leqslant \frac{\sum_{k \in I \setminus \{i\}} \theta_k}{1 - \gamma} \Big(2 u_{\max} + \frac{2 \gamma u_{\max}}{(1 - \gamma)} \Big) \\
&\leqslant \frac{2 \sum_{k \in I \setminus \{i\}} \theta_k}{(1 - \gamma)^2} u_{\max}.
\end{aligned}$$

This concludes the proof.

(d) For any $s \in S, i \in I, a_i \in A_i, \pi \in \Pi$, we note that

$$
\begin{aligned}
&|Q_i(s, a_i; \pi_i, \pi_{-i}) - Q_i(s, a_i; \pi_i, \pi_{-i}^{(\theta)})| \\
&\leqslant |u_i(s, a_i, \pi_{-i}) - u_i(s, a_i, \pi_{-i}^{(\theta)})| \\
&\quad + \gamma \Bigg| \sum_{s'} P(s'|s, a_i, \pi_{-i}) V_i(s'; \pi) \\
&\qquad - \sum_{s'} P(s'|s, a_i, \pi_{-i}^{(\theta)}) V_i(s'; \pi_i, \pi_{-i}^{(\theta)}) \Bigg| \\
&\overset{(C.14)}{\leqslant} 2 \sum_{k \in I \setminus \{i\}} \theta_k u_{\max} + \gamma \Bigg| \sum_{s'} P(s'|s, a_i, \pi_{-i}) V_i(s'; \pi) \\
&\qquad - \sum_{s'} P(s'|s, a_i, \pi_{-i}^{(\theta)}) V_i(s'; \pi_i, \pi_{-i}^{(\theta)}) \Bigg| \\
&\leqslant 2 \sum_{k \in I \setminus \{i\}} \theta_k u_{\max} + \gamma \Bigg| \sum_{s'} P(s'|s, a_i, \pi_{-i}) \Big(V_i(s'; \pi) \\
&\qquad - V_i(s'; \pi_i, \pi_{-i}^{(\theta)}) \Big) \Bigg| + \gamma \Bigg| \sum_{s'} \Big(P(s'|s, a_i, \pi_{-i}) \\
&\quad - P(s'|s, a_i, \pi_{-i}^{(\theta)}) \Big) V_i(s'; \pi_i, \pi_{-i}^{(\theta)}) \Bigg| \\
&\overset{(i)}{\leqslant} 2 \sum_{k \in I \setminus \{i\}} \theta_k u_{\max} + \gamma \frac{2 \sum_{k \in I \setminus \{i\}} \theta_k}{(1-\gamma)^2} u_{\max} \\
&\quad + \gamma \Bigg| \sum_{s'} \Big(P(s'|s, a_i, \pi_{-i}) - P(s'|s, a_i, \pi_{-i}^{(\theta)}) \Big) V_i(s'; \pi_i, \pi_{-i}^{(\theta)}) \Bigg| \\
&\overset{(C.15)}{\leqslant} u_{\max} \sum_{k \in I \setminus \{i\}} \theta_k \left(2 + \gamma \frac{2}{(1-\gamma)^2} + \gamma \frac{2}{(1-\gamma)} \right) \\
&= \frac{2 \sum_{k \in I \setminus \{i\}} \theta_k}{(1-\gamma)^2} u_{\max},
\end{aligned}
$$

where (i) is due to Lemma 4.2.3-(c). This completes the proof.

Proof of Corollary 4.2.1

First, we note that for every $\tilde{\epsilon} > 0$, there exists $\epsilon > 0$ such that $\epsilon + h_\epsilon = \tilde{\epsilon}$. This claim follows by noting that Assumption 4.2.3 guarantees that the map $\epsilon \in \mathbb{R}_+ \mapsto \epsilon + h_\epsilon \in \mathbb{R}_+$ is continuous function that goes to zero as ϵ approaches 0. Furthermore, since h_ϵ is non-decreasing in ϵ, we note that $\epsilon + h_\epsilon$ is increasing in ϵ. Next, we show that every $\tilde{\epsilon}, \tilde{\epsilon}'$ such that $0 < \tilde{\epsilon} < \tilde{\epsilon}'$, there exist positive scalars $0 < \epsilon < \epsilon'$ such that $\epsilon + h_\epsilon = \tilde{\epsilon}$ and $\epsilon' + h_{\epsilon'} = \tilde{\epsilon}'$.

We show this by contradiction. Suppose for some $0 < \tilde{\epsilon} < \tilde{\epsilon}'$ it holds that $\epsilon + h_\epsilon = \tilde{\epsilon}$ and $\epsilon' + h_{\epsilon'} = \tilde{\epsilon}'$ with $\epsilon \geqslant \epsilon'$. This implies that $\tilde{\epsilon} = \epsilon + h_\epsilon \geqslant \epsilon' + h_{\epsilon'} = \tilde{\epsilon}'$, which contradicts the fact that $\tilde{\epsilon} < \tilde{\epsilon}'$.

Next, we show that for every $\tilde{\epsilon} > 0$, the sequence of policies $\{\pi^t\}_{t=0}^{\infty}$ induced by Algorithm 3 converges to the set $\mathsf{NE}(\tilde{\epsilon})$ with probability 1, if $\sum_{i \in I} \theta_i < L\epsilon$, where ϵ is such that $\epsilon + h_\epsilon = \tilde{\epsilon}$. Following the same steps from the proof of Theorem 4.2.1, it is sufficient to characterize the convergent set of the dynamical system (4.10) in order to study the asymptotic behavior of the policy updates in Algorithm 3.

From the proof of Lemma 4.2.4, we know that any absolutely continuous trajectory of (4.10) converges to the set Π_ϵ^*, if $\sum_{i \in I} \theta_i < L\epsilon$. The proof concludes by noting that $\Pi_\epsilon^* \subseteq \mathsf{NE}(\epsilon + h_\epsilon) = \mathsf{NE}(\tilde{\epsilon})$, due to Assumption 4.2.3.

C.3 Auxiliary Lemma

Lemma C.3.1 ([238]). *Consider a Markov potential game $\mathcal{G}$, with potential function Φ. Then, for every $s \in S, i \in I, a_i \in A_i, \pi \in \Pi$,*

$$\frac{\partial V_i(s,\pi)}{\partial \pi_i(s,a_i)} = \frac{\partial \Phi(s,\pi)}{\partial \pi_i(s,a_i)}. \tag{C.16}$$

Proof. To prove this result, we first show that for every $i \in I$, there exists a function $U_i : S \times \Pi_{-i} \to \mathbb{R}$ such that

$$V_i(s,\pi) = \Phi(s,\pi) + U_i(s,\pi_{-i}), \quad \forall\ s \in S, \pi \in \Pi. \tag{C.17}$$

Fix arbitrary $i \in I, \pi_i, \pi_i', \pi_i'' \in \Pi_i$, and $\pi_{-i} \in \Pi_{-i}$. By Definition 2.3.1, it holds that, for every $s \in S$,

$$\begin{aligned}
\Phi(s,\pi_i,\pi_{-i}) - \Phi(s,\pi_i',\pi_{-i}) &= V_i(s,\pi_i,\pi_{-i}) - V_i(s,\pi_i',\pi_{-i}) \\
\Phi(s,\pi_i,\pi_{-i}) - \Phi(s,\pi_i'',\pi_{-i}) &= V_i(s,\pi_i,\pi_{-i}) - V_i(s,\pi_i'',\pi_{-i}).
\end{aligned}$$

By re-arranging the terms in the above equation, we have

$$\begin{aligned}
V_i(s,\pi_i,\pi_{-i}) - \Phi(s,\pi_i,\pi_{-i}) &= V_i(s,\pi_i',\pi_{-i}) - \Phi(s,\pi_i',\pi_{-i}) \\
V_i(s,\pi_i,\pi_{-i}) - \Phi(s,\pi_i,\pi_{-i}) &= V_i(s,\pi_i'',\pi_{-i}) - \Phi(s,\pi_i'',\pi_{-i}).
\end{aligned} \tag{C.18}$$

Thus, equating the RHS in the above equation, we obtain that

$$V_i(s,\pi_i',\pi_{-i}) - \Phi(s,\pi_i',\pi_{-i}) = V_i(s,\pi_i'',\pi_{-i}) - \Phi(s,\pi_i'',\pi_{-i}).$$

Since π_i'', π_i' are arbitrary, we know that for every $i \in I, s \in S, \pi_{-i} \in \Pi_{-i}$, $V_i(s,\pi_i,\pi_{-i}) - \Phi(s,\pi_i,\pi_{-i})$ does not depend on π_i. Thus, using (C.18), we conclude that (C.17) holds.

Equation (C.16) follows from taking the derivative with respect to $\pi_i(s,a_i)$ on both sides of (C.17) and noting that $U_i(s,\pi_{-i})$ does not depend on π_i. □

Lemma C.3.2 (Characterization of Nash equilibrium)**.** *A policy $\pi^* \in \Pi$ is a Nash equilibrium of $\mathcal{G}$ if and only if $\pi^*(s) = \mathrm{br}_i(s; \pi^*)$ for all $i \in I$ and all $s \in S$.*

Proof. We prove the claim in two parts – first, we show that any policy π^* such that $\pi^*(s) = \mathrm{br}_i(s; \pi^*)$ is a Nash equilibrium of $\mathcal{G}$. Next, we show the converse.

First, we provide an important characterization of the one-step optimal deviation which is crucial for the following proof

$$\begin{aligned}
\mathrm{br}_i(s; \pi_i^*, \pi_{-i}^{*,(\theta)}) &= \arg\max_{\hat{\pi}_i \in \gamma(A_i)} \hat{\pi}_i^\top Q_i(s; \pi_i^*, \pi_{-i}^*) \\
&= \arg\max_{\hat{\pi} \in \gamma(A_i)} \Bigg(u_i(s, \hat{\pi}_i, \pi_{-i}^*) && \text{(C.19)} \\
&\qquad + \gamma \sum_{s'} P(s'|s, \hat{\pi}_i, \pi_{-i}^*) V_i(s', \pi_i^*, \pi_{-i}^*) \Bigg). && \text{(C.20)}
\end{aligned}$$

First, we prove that $\tilde{\pi}^*$ is a Nash equilibrium of $\mathcal{G}$, we need to show that for every $i \in I, s \in S, \pi_i' \in \Pi_i$,

$$V_i(s, \tilde{\pi}_i^*, \pi_{-i}^*) \geqslant V_i(s, \pi_i', \pi_{-i}^*). \tag{C.21}$$

Before proving (C.21), we first show that for any integer $K \geqslant 1$, any $s \in S$, any $i \in I$, and any $\pi_i' \in \Pi_i$,

$$\begin{aligned}
&V_i(s, \pi_i^*, \pi_{-i}^*) \\
&\geqslant \mathbb{E}\left[\sum_{k=0}^{K-1} \gamma^k u_i(s^k, \pi_i', \pi_{-i}^*) + \gamma^K V_i(s^K, \tilde{\pi}_i^*, \pi_{-i}^*) \right], && \text{(C.22)}
\end{aligned}$$

where $s^0 = s, a_i^k \sim \pi_i'(s^k), a_{-i}^k \sim \pi_{-i}^*(s^k), s^k \sim P(\cdot|s^{k-1}, a^{k-1})$.

Consider $K = 1$, for any $s \in S$, any $i \in I$, any $\pi_i' \in \Pi_i$ we have,

$$\begin{aligned}
&V_i(s, \tilde{\pi}_i^*, \pi_{-i}^*) \\
&= u_i(s, \pi_i^*, \pi_{-i}^*) + \gamma \sum_{s'} P(s'|s, \tilde{\pi}_i^*, \pi_{-i}^*) V(s', \tilde{\pi}_i^*, \pi_{-i}^*) && \text{(C.23)} \\
&\geqslant u_i(s, \pi_i', \pi_{-i}^*) + \gamma \sum_{s'} P(s'|s, \pi_i', \tilde{\pi}_{-i}^*) V(s', \tilde{\pi}_i^*, \pi_{-i}^*) \\
&= \mathbb{E}\left[u_i(s^0, \pi_i', \pi_{-i}^{*,(\theta)}) + \gamma V_i(s^1, \tilde{\pi}_i^*, \pi_{-i}^{*,(\theta)}) \right],
\end{aligned}$$

where, again, $s^0 = s$, $a_i^0 \sim \pi_i'(s^0)$, $a_{-i}^0 \sim \pi_{-i}^*(s^0)$, $s^1 \sim P(\cdot|s^0, a^0)$ and the inequality follows from (C.19) as $\pi_i^*(s) \in \mathsf{br}_i(s; \pi^*)$ for every $i \in I, s \in S$.

Next, suppose that (C.22) holds for some integer K, we consider $K+1$:

$$\begin{aligned}
&V_i(s,\pi_i^*,\pi_{-i}^*)\\
\underset{(a)}{\geqslant}&\mathbb{E}\left[\sum_{k=0}^{K-1}\gamma^k u_i(s^k,\pi_i',\pi_{-i}^*)+\gamma^K V_i(s^K,\tilde{\pi}_i^*,\pi_{-i}^*)\right]\\
\underset{(b)}{=}&\mathbb{E}\Bigg[\sum_{k=0}^{K-1}\gamma^k u_i(s^k,\pi_i',\pi_{-i}^*)+\gamma^K\Bigg(u_i(s^K,\pi_i^*,\pi_{-i}^*)\\
&\qquad+\gamma\sum_{s'}P(s'|s^K,\pi_i^*,\pi_{-i}^*)V_i(s',\tilde{\pi}_i^*,\pi_{-i}^*)\Bigg)\Bigg]\\
\underset{(c)}{\geqslant}&\mathbb{E}\Bigg[\sum_{k=0}^{K-1}\gamma^k u_i(s^k,\pi_i',\pi_{-i}^*)+\gamma^K\Bigg(u_i(s^K,\pi_i',\pi_{-i}^*)\\
&\qquad+\gamma\sum_{s'}P(s'|s^K,\pi_i',\pi_{-i}^*)V_i(s',\tilde{\pi}_i^*,\pi_{-i}^*)\Bigg)\Bigg]\\
\underset{(d)}{=}&\mathbb{E}\left[\sum_{k=0}^{K}\gamma^k u_i(s^k,\pi_i',\pi_{-i}^*)+\gamma^{K+1}V_i(s^{K+1},\tilde{\pi}_i^*,\pi_{-i}^*)\right],
\end{aligned}$$

where (a) is by induction hypothesis, (b) is due to (C.23), (c) is due to (C.19) and (d) is by rearrangement of terms. Thus, by mathematical induction, we have established that (C.22) holds for all K. Let $K\to\infty$ in (C.22), we have

$$V_i(s,\pi_i^*,\pi_{-i}^*)\geqslant\mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k u_i(s^k,\pi_i',\pi_{-i}^*)\right]=V_i(s,\pi_i',\pi_{-i}^*),$$

for every $s\in S, i\in I,\pi_i'\in\Pi$. Thus, we have proved (C.21), i.e. $\tilde{\pi}^*$ is a Nash equilibrium of game $\mathcal{G}$.

Next, we show that any Nash equilibrium π^* of $\mathcal{G}$ satisfies that $\tilde{\pi}_i^*(s)\in\mathrm{br}_i(s;\tilde{\pi}^*)$ for every $i\in I, s\in S$. We prove this by contradiction. Suppose there exists a player $i\in I$, and set of states $\bar{S}\subset S$ such that for every $\bar{s}\in\bar{S}$ it holds that $\tilde{\pi}_i^*(\bar{s})\notin\mathrm{br}_i(\bar{s};\tilde{\pi}^*)$. Let π' be a policy such that for all $s\in S, i\in I$, $\pi_i'(s)\in\mathrm{br}_i(s;\tilde{\pi}^*)$. Without loss of generality we assume $|\bar{S}|=1$.

We claim that for any integer $K\geqslant 1$, any $s\in S$, $i\in I$ it holds that

$$\begin{aligned}
&V_i(s,\pi_i^*,\pi_{-i}^*)\\
&\leqslant\mathbb{E}\left[\sum_{k=0}^{K-1}\gamma^k u_i(s^k,\pi_i',\pi_{-i}^*)+\gamma^K V_i(s^K,\tilde{\pi}_i^*,\pi_{-i}^*)\right],
\end{aligned}\tag{C.24}$$

where $s^0=s, a_i^k\sim\pi_i'(s^k), a_{-i}^k\sim\pi_{-i}^*(s^k), s^k\sim P(\cdot|s^{k-1},a^{k-1})$ and the inequality is strict for $s=\bar{s}$.

Consider $K = 1$, for any $s \in S$, any $i \in I$, we have

$$\begin{aligned}
&V_i(s, \pi_i^*, \pi_{-i}^*) \\
&= u_i(s, \tilde{\pi}_i^*, \pi_{-i}^*) + \gamma \sum_{s'} P(s'|s, \tilde{\pi}_i^*, \pi_{-i}^*) V_i(s', \tilde{\pi}_i^*, \pi_{-i}^*) \\
&\leqslant u_i(s, \pi_i', \tilde{\pi}_{-i}^*) + \gamma \sum_{s'} P(s'|s, \pi_i', \tilde{\pi}_{-i}^*) V_i(s', \tilde{\pi}^*) \\
&= \mathbb{E}\left[u_i(s^0, \pi_i', \pi_{-i}^*) + \gamma V_i(s^1, \tilde{\pi}^*)\right],
\end{aligned}$$

where, again, $s^0 = s$, $a_i^0 \sim \pi_i'(s^0)$, $a_{-i}^0 \sim \pi_{-i}^*(s^0)$, $s^1 \sim P(\cdot|s^0, a^0)$ and the inequality follows from (C.19). Note that inequality is strict for $s^0 = \bar{s}$.

Next, suppose (C.24) holds for some integer K, we consider $K + 1$:

$$\begin{aligned}
V_i(s, \pi_i^*, \pi_{-i}^*) &\underset{(a)}{\leqslant} \mathbb{E}\left[\sum_{k=0}^{K-1} \gamma^k u_i(s^k, \pi_i', \pi_{-i}^*) + \gamma^K V_i(s^K, \tilde{\pi}^*)\right] \\
=&\mathbb{E}\Bigg[\sum_{k=0}^{K-1} \gamma^k u_i(s^k, \pi_i', \pi_{-i}^*) + \gamma^K \Bigg(u_i(s^K, \pi_i^*, \pi_{-i}^*) \\
&\qquad + \gamma \sum_{s'} P(s'|s^K, \pi_i^*, \pi_{-i}^*) V_i(s', \tilde{\pi}^*)\Bigg)\Bigg] \\
\underset{(b)}{\leqslant}&\mathbb{E}\Bigg[\sum_{k=0}^{K-1} \gamma^k u_i(s^k, \pi_i', \pi_{-i}^*) + \gamma^K \Bigg(u_i(s^K, \pi_i', \pi_{-i}^*) \\
&\qquad + \gamma \sum_{s'} P(s'|s^K, \pi_i', \pi_{-i}^*) V_i(s', \tilde{\pi}^*)\Bigg)\Bigg] \\
\underset{(c)}{=}&\mathbb{E}\left[\sum_{k=0}^{K} \gamma^k u_i(s^k, \pi_i', \pi_{-i}^*) + \gamma^{K+1} V_i(s^{K+1}, \tilde{\pi}^*)\right],
\end{aligned}$$

where (a) is by induction hypothesis, (b) is due to (C.19) and (c) is by rearrangement of terms. Thus, by mathematical induction, we have established that (C.24) holds for all K. Let $K \to \infty$ in (C.24), we have

$$V_i(s, \pi_i^*, \pi_{-i}^*) \leqslant \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k u_i(s^k, \pi_i', \pi_{-i}^*)\right] = V_i(s, \pi_i', \pi_{-i}^*),$$

for every $s \in S, i \in I, \pi_i' \in \Pi$. Furthermore,

$$V_i(\bar{s}, \pi_i^*, \pi_{-i}^*) < \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k u_i(s^k, \pi_i', \pi_{-i}^*)\right] = V_i(\bar{s}, \pi_i', \pi_{-i}^*),$$

This contradicts the fact that π_i^* is a Nash equilibrium of game $\mathcal{G}$. □

Appendix D

Appendix for Chapter 5

D.1 Technical Results for the Proof of Theorem 5.3.1

We now present a technical result used in proof of Theorem 5.3.2.

Lemma D.1.1. *For any Markov game G, an associated MNPF Φ with closeness parameter κ, and $\mu \in \Delta(S)$, the mapping $\pi \mapsto \Phi(\mu, \pi)$ is $L-$Lipschitz continuous, with*

$$L = \left(\kappa \sqrt{N} + \frac{r_{\max}}{(1-\gamma)^2} \sqrt{N|S||A|} \right).$$

Proof. To show the desired Lipschitz bound, it is sufficient to show that

$$|\Phi(\mu, \pi) - \Phi(\mu, \pi')| \leqslant L \|\pi - \pi'\|, \quad \forall\ \pi, \pi' \in \Pi.$$

For the remaining proof, consider two arbitrary policies $\pi = (\pi_1, \pi_2, ..., \pi_N)$, and $\pi' = (\pi'_1, \pi'_2, ..., \pi'_N)$. For any $i \in \{0, 1, .., N\}$, define a joint policy

$$\pi^{(i)} = (\pi_1, \pi_2, \cdots, \pi_i, \pi'_{i+1}, \cdots, \pi'_N).$$

Naturally, $\pi^{(0)} = \pi'$ and $\pi^{(N)} = \pi$.

Note that

$$|\Phi(\mu, \pi) - \Phi(\mu, \pi')| = |\Phi(\mu, \pi^{(N)}) - \Phi(\mu, \pi^{(0)})| \leqslant \sum_{i=0}^{N-1} |\Phi(\mu, \pi^{(i+1)}) - \Phi(\mu, \pi^{(i)})|.$$

Since $\pi^{(i+1)}$ and $\pi^{(i)}$ only differ in the policy of player $(i+1)$, using Definition 5.2.1, we

obtain

$$|\Phi(s,\pi)-\Phi(s,\pi')| \leqslant \kappa \sum_{i=0}^{N-1} \|\pi^{(i+1)}-\pi^{(i)}\| + \sum_{i=0}^{N-1} |V_{i+1}(s,\pi^{(i+1)}) - V_{i+1}(s,\pi^{(i)})|$$
$$\leqslant \kappa \sum_{i=0}^{N-1} \|\pi_{i+1}-\pi'_{i+1}\|$$
$$+ \sum_{i=0}^{N-1} \left| \sum_{s'\in S, a'_{i+1}\in A_{i+1}} \frac{\partial V_{i+1}(s,\tilde{\pi})}{\partial \tilde{\pi}_{i+1}(s',a'_{i+1})}\Big|_{\tilde{\pi}=\zeta^{(i)}} (\pi_{i+1}(s',a'_{i+1}) - \pi'_{i+1}(s',a'_{i+1}))\right|, \tag{D.1}$$

where $\zeta^{(i)} = (\pi_1, \pi_2, \cdots, \pi_i, \xi'_{i+1}, \pi'_{i+2}..., \pi'_N)$ and $\xi'_{i+1} = \pi_{i+1} + t(\pi_{i+1} - \pi'_{i+1})$ for some $t \in [0,1]$.

Using (D.1), along with Lemma 4.2.3-(a), we obtain

$$|\Phi(\mu,\pi)-\Phi(\mu,\pi')|$$
$$\leqslant \kappa \sum_{i=0}^{N-1} \|\pi_{i+1}-\pi'_{i+1}\|$$
$$+ \frac{1}{(1-\gamma)} \sum_{i=0}^{N-1} \left| \sum_{s'\in S, a'_{i+1}\in A_{i+1}} d_\mu^{\zeta^{(i)}}(s') Q_{i+1}(s',a'_{i+1},\zeta^{(i)})(\pi_{i+1}(s',a'_{i+1}) - \pi'_{i+1}(s',a'_{i+1}))\right|,$$
$$\leqslant \kappa \sum_{i=0}^{N-1} \|\pi_{i+1}-\pi'_{i+1}\|$$
$$+ \frac{1}{(1-\gamma)} \sum_{i=0}^{N-1} \sum_{s'\in S, a'_{i+1}\in A_{i+1}} \left| d_\mu^{\zeta^{(i)}}(s') Q_{i+1}(s',a'_{i+1},\zeta^{(i)})(\pi_{i+1}(s',a'_{i+1}) - \pi'_{i+1}(s',a'_{i+1}))\right|.$$

Additionally, using the fact that $|d_\mu^{(\zeta^{(i)})}(s')| \in [0,1]$, we obtain

$$|\Phi(\mu,\pi)-\Phi(\mu,\pi')| \leqslant \kappa \sum_{i=0}^{N-1} \|\pi_{i+1}-\pi'_{i+1}\|$$
$$+ \frac{1}{(1-\gamma)} \sum_{i=0}^{N-1} \sum_{s'\in S, a'_{i+1}\in A_{i+1}} \left| Q_{i+1}(s',a'_{i+1},\zeta^{(i)})(\pi_{i+1}(s',a'_{i+1}) - \pi'_{i+1}(s',a'_{i+1}))\right|$$
$$\leqslant \kappa \sum_{i=0}^{N-1} \|\pi_{i+1}-\pi'_{i+1}\|$$
$$+ \frac{1}{(1-\gamma)} \sum_{i=0}^{N-1} \max_{s',a'_{i+1},\zeta^{(i)}} Q_{i+1}(s',a'_{i+1},\zeta^{(i)}) \sum_{s',a'_{i+1}\in A_{i+1}} \left| (\pi_{i+1}(s',a'_{i+1}) - \pi'_{i+1}(s',a'_{i+1}))\right|.$$

Furthermore, we note that

$$\max_{s',a'_{i+1},\zeta^{(i)}} |Q_{i+1}(s',a'_{i+1},\zeta^{(i)})| \leqslant \frac{u_{\max}}{(1-\gamma)}.$$

where $u_{\max} = \max_{i,s,a} u_i(s,a)$. Therefore, we obtain,

$$\begin{aligned}
&|\Phi(\mu,\pi) - \Phi(\mu,\pi')| \\
&\leqslant \kappa \sum_{i=0}^{N-1} \|\pi_{i+1} - \pi'_{i+1}\| \\
&\quad + \frac{u_{\max}}{(1-\gamma)^2} \sum_{i=0}^{N-1} \sum_{s'\in S, a'_{i+1}\in A_{i+1}} \left| (\pi_{i+1}(s',a'_{i+1}) - \pi'_{i+1}(s',a'_{i+1})) \right|.
\end{aligned}$$

Finally, by using Cauchy-Schwarz inequality, we conclude that

$$\begin{aligned}
&|\Phi(s,\pi) - \Phi(s,\pi')| \\
&\leqslant \left(\kappa\sqrt{N} + \frac{u_{\max}}{(1-\gamma)^2} \sqrt{|S||A|N} \right) \|\pi - \pi'\|.
\end{aligned}$$

□

Lemma D.1.2. *For any $\delta \geqslant 0$, the function $\Gamma(\delta)$ (defined in (5.21)) exists, is upper semi-continuous and weakly increasing.*

Proof. First, we show that $\Gamma(\cdot)$ exists for every $\delta \geqslant 0$ and is upper semicontinous. This follows directly from the fact that $\mathsf{NE}(\delta)$ is a non-empty and compact for any non-negative δ, and the mapping $\pi \mapsto \min_{k\in[K]} \|\pi - \pi^{*k}\|$ is continuous. Furthermore, the upper semicontinuity follows directly from Berge's maximum theorem. Finally, we show that $\Gamma(\cdot)$ is a weakly increasing function. This is due to the fact that $\mathsf{NE}(\delta) \subseteq \mathsf{NE}(\delta')$ for any $\delta' > \delta \geqslant 0$. □

Lemma D.1.3 (Lemma 4.2.3-(c))**.** *For any $i \in I, \pi_i \in \Pi_i, \pi_{-i} \in \Pi_{-i}$ it holds that*

$$\max_{s\in S} |V_i(s,\pi_i,\pi_{-i}) - V_i(s,\pi_i,\pi^\theta_{-i})| \leqslant \frac{2\theta|I|}{(1-\gamma)^2} r_{\max},$$

where $r_{\max} := \max_{i,s,a} |r_i(s,a)|$, $\pi^\theta_{-i}(s) := (1-\theta)\pi_{-i}(s) + \theta\pi^\circ$ and $\pi^\circ := (1/|A_{-i}|)\mathbb{1}_{A_{-i}}$.

Lemma D.1.4 (Lemma 4.2.3-(d))**.** *For any $i \in I, \pi_i \in \Pi_i, \pi_{-i} \in \Pi_{-i}$, it holds that*

$$\max_{s,a_i} |Q_i(s,a_i;\pi_i,\pi_{-i}) - Q_i(s,a_i;\pi_i,\pi^\theta_{-i})| \leqslant \frac{2\theta|I|}{(1-\gamma)^2} r_{\max},$$

where $r_{\max} := \max_{i,s,a} |r_i(s,a)|$, $\pi^\theta_{-i}(s) := (1-\theta)\pi_{-i}(s) + \theta\pi^\circ$ and $\pi^\circ := (1/|A_{-i}|)\mathbb{1}_{A_{-i}}$.

Lemma D.1.5 (Lemma 4.2.3-(b))**.** *For any policy* $\pi = (\pi_i, \pi_{-i}), \pi' = (\pi'_i, \pi_{-i}) \in \Pi$ *and any* $\mu \in \Delta(S)$*,*

$$V_i(\mu, \pi) - V_i(\mu, \pi') = (1/(1-\gamma)) \cdot \sum_{s'} d^{\pi}_{\mu}(s') \Gamma_i(s', \pi_i; \pi'),$$

where $\Gamma_i(s, a_i; \pi) \coloneqq Q_i(s, a_i; \pi) - V_i(s, \pi)$.

Appendix E

Appendix for Chapter 6

In Section E.1, we review the adaptive adversarial algorithms proposed in [72] and specialize them to our setting to derive the regret bounds presented in Chapter 6. In Section E.2, we provide the proofs of the lemmas stated in Section 6.3. In Section E.3, we present the proof of the main theorem stated in Section 6.3. In Section E.4, we present the main technical lemmas used in the proof of our main result (Theorem 6.3.1). In Section E.5, we provide the Thompson sampling-based variant of Algorithm 6 and present the analogous results to those in Section 6.3.

Before, discussing the content of this chapter, we present the table of notations (Table E.0.1) that will be useful in navigating through the notations.

E.1 Adaptive Adversarial Algorithms

In this work, we deploy the optimistic mirror descent-based adversarial bandit module. We adapt algorithms from [72], which build upon and improve the algorithm originally proposed in [425]. In this section, we recap the results from [72]. For completeness, we restate the problem formulation and algorithm. Toward the end, we specialize their results to the setting of this chapter and state a useful result that characterizes the regret of such algorithms—within the bandit structure described in Section 6.2—in terms of the number of matchings and collisions.

Problem formulation from [72]

In this section, we review the algorithm described in [72], which improves upon the one introduced in [425]. Consider a multi-armed bandit problem that unfolds over τ time steps with $A \leqslant \tau$ fixed actions. In each round t, the algorithm selects one arm $i(t) \in [A]$, while simultaneously, an adversary chooses the loss vector $\ell(t) = (\ell_i(t))_{i\in[A]} \in [-1, 1]^A$. The adversary may be adaptive, meaning it can base its choices on the algorithm's past actions. The goal of the algorithm is to minimize the regret, defined as the gap between the total

Notation	Description
A	Set of agents
F	Set of firms/arms
M	Union of agents and firms
$u_a(f)$	Utility for agent a when matched with firm f
$u_f(a)$	Utility for firm f when matched with agent a
$f_a(t)$	Firm chosen by agent a at time t
f_a^*	Stable match of agent a
$\overline{\mathbb{F}}_a$	Set of super-optimal firms for agent a
$\underline{\mathbb{F}}_a$	Set of sub-optimal firms for agent a
K	Number of markets formed by decomposition as stated in Remark 6.1.2
$\mathcal{A}_i$	Agents forming fixed pairs after $i-1$ rounds of elimination (Remark 6.1.2)
$\mathcal{F}_i$	Firms forming fixed pairs after $i-1$ rounds of elimination (Remark 6.1.2)
$r_{a,f}$	Noisy reward that agent a receives on getting matched with firm f
$\mathbb{A}_f$	Set of agents that pull firm f
$M_{a,f}(T)$	Number of times agent a has successfully matched with firm f till time T
$C_{a,f}(T)$	Number of times agent a has collided on firm f till time T
$p_{a,f}(t)$	Probability that agent a will pull firm f at time t
$P_{a,f}(t)$	An indicator if agent a has pulled arm f at time t
$Y_a(t)$	An indicator if agent a got successfully matched at time t
$\hat{\mu}_{a,f}(t)$	Empirical mean of utility derived by agent a on matching with f
$\mathsf{UCB}_{a,f}(t)$	UCB estimate of reward from firm f to agent a at time t
$\mathcal{T}_{a,f}(t)$	Thompson Sampling index of reward from firm f to agent a at time t
$E_{a,f}^{(\mathsf{r})}(t)$	An indicator if agent a pulled firm f at time t
$E_{a,f}^{(\mathsf{c})}(t)$	An indicator if all the firms with higher index than f got pruned at time t
$\tau_{a,f}(T)$	Time steps during which $E_{a,f}^{(\mathsf{c})}(t)=1$
$\Delta_{a,f}$	$u_a(f_a^*)-u_a(f)$
$H_{a,f}(t)$	Event when some other more preferred agent than a has requested firm f at time t

Table E.0.1: Table of notations

accumulated loss and the loss of the best fixed arm in hindsight:

$$\text{Regret}^{(\mathsf{adv})}(\tau) = \max_{i^\star \in [A]} \mathbb{E}\left[\sum_{t=1}^{\tau} \ell_{i(t)}(t) - \sum_{t=1}^{\tau} \ell_{i^\star}(t)\right].$$

The algorithm is based on the optimistic mirror descent framework. At any time t, the algorithm samples an arm $i(t) \in [A]$ according to a probability distribution $p(t) \in \Delta([A])$. The algorithm only observes the loss associated with the chosen action and not those of the other actions. Therefore, upon receiving the loss $\ell_{i(t)}(t)$, the algorithm constructs an unbiased estimator of the losses for all actions. The estimator is

$$\hat{L}_i(t) = \frac{\ell_i(t) - L(t-1)}{2p_i(t)} \mathbb{1}\left(i(t) = i\right) + \frac{1 + L(t-1)}{2}, \quad \forall\ i$$

The unbiased loss estimate $\hat{L}(t)$ is used to update the an auxiliary probability distribution $x(t+1) \in \Delta([A])$ through an optimistic mirror descend update with learning rate η. The optimistic mirror descend update is constructed from the Bregman divergence[1] associated with a log-barrier regularizer $\mathbb{R}^A \ni x \mapsto \psi(x) = \frac{1}{\eta}\sum_{i=1}^{A} \ln \frac{1}{x_i}$ as follows

$$x(t+1) = \operatorname*{arg\,min}_{z \in \Delta([A])} \left\langle z, \hat{L}(t) \right\rangle + D_\psi(z, x(t)).$$

The distribution $x(t+1)$ is used to update the arm sampling distribution $p(t+1)$ after mixing a small bias towards most recently picked arm as follows

$$p(t+1) = (1 - \lambda(t+1))x(t+1) + \lambda(t+1)\mathbf{e}_{i(t)},$$

where $\mathbf{e}_{i^t} \in \mathbb{R}^A$ is an element of standard basis in $\mathbb{R}^A$ with $i(t)$ element as 1 and all others as zero and

$$\lambda(t+1) = \frac{\lambda(1 - L(t))}{2 + \lambda(1 - L(t))},$$

for some $\lambda > 0$.

Against the preceding backdrop, we restate Theorem 2 from [72] below:

Theorem E.1.1. *Algorithm 15 with $\eta \leqslant \frac{1}{50}$, $\lambda = 8\eta$ ensures that*

$$\textit{Regret}^{(\mathsf{adv})}(\tau) = \mathcal{O}\left(\frac{A\ln(T)}{\eta}\right) + 8\eta\mathbb{E}\left[V(T)\right],$$

where $V(T) \coloneqq \sum_{t=2}^{T} |\ell_{i(t-1)}(t) - \ell_{i(t-1)}(t-1)|$ is commonly referred as "path-length".

Remark E.1.1. *Note that Theorem 2 in [72] requires[2] $\eta \leqslant 1/162$ and $\lambda = 8\eta$. But in fact the proof goes through for $\eta \leqslant 1/50$. This is because in [72] for the proof of Theorem 2, they directly lift [425, Theorem 7] where $\eta \leqslant 1/162$ which is not tuned completely.*

[1] Bregman divergence between two point x, y with respect to a convex regularizer ψ is given as $D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y\rangle$.

[2] Moreover, it is an algebraic exercise to establish that $\eta < \frac{1}{24}$ and $\lambda = \frac{1 - 12\eta - c\cdot\sqrt{1-24\eta}}{24}$ also works for some $c \in (0, 1)$. But we don't go in this direction to retain simplicity of algorithmic description.

Algorithm 15 Optimistic Mirror Descent for Adversarial Bandits

1: **Parameters:** $\eta, \lambda \in (0,1)$
2: **Initialize:** $p(1),\ x(1) \sim \mathsf{Unif}([A]),\ \psi(x) = \frac{1}{\eta}\sum_{i=1}^{A} \ln \frac{1}{x_i}$
3: **for** $t = 1$ to τ **do**
4: Play $i(t) \sim p(t)$ and observe $L(t) = \ell_{i(t)}(t)$
5: **for** each $i \in [A]$ **do**
6: $\hat{L}_i(t) = \frac{\ell_i(t) - L(t-1)}{2p_i(t)} \cdot \mathbf{1}_{\{i(t)=i\}} + \frac{1+L(t-1)}{2}$
7: **end for**
8: Update $x(t+1) = \arg\min_{z \in \Delta([A])} \langle z, \hat{L}(t)\rangle + D_\psi(z, x(t))$
9: $\lambda(t+1) = \frac{\lambda(1-L(t))}{2+\lambda(1-L(t))}$
10: $p(t+1) = (1-\lambda(t+1))x(t+1) + \lambda(t+1)\mathbf{e}_{i(t)}$
11: **end for**

Adaptive Adversarial Module

In this section we describe AB_Subroutine in Algorithm 6 which is based on the algorithm presented in Section E.1.

For any $(a,f) \in A \times F$, the adversarial bandit module associated with (a,f) (as described in Algorithm 7) is a specialized version of Algorithm 15 for the case with two actions: *request the firm f* or *prune the firm f*. In this setup, the loss incurred by pruning firm f is always 0, while the loss from requesting firm f depends on whether agent a was successfully matched with it or experienced a collision. In this special case of two actions, the optimistic mirror descent update (line 4 in Algorithm 15) admits a closed-form expression (see Lemma E.1.2). Note that the adversarial bandit module associated with any agent-firm pair (a,f) is only active during rounds $t \in \tau_{a,f}(T) \subset [T]$.

Lemma E.1.1. *Given a scalar $\eta \leqslant \frac{1}{50}$, for any agent-firm pair $(a,f) \in A \times F$, the regret of the adversarial bandit algorithm is bounded as*

$$\begin{aligned}&\mathbb{E}[\mathit{Regret}_{a,f}^{(\mathsf{adv})}(\tau_{a,f}(T))] \\ &\leqslant \mathcal{O}\left(\frac{\log(T)}{\eta}\right) + 32\eta\mathbb{E}\left[\min\left\{M_{a,f}^{\star}(T), C_{a,f}^{\star}(T), M_{a,f}(T) + C_{a,f}(T)\right\}\right],\end{aligned}$$

where $M_{a,f}^{\star}(T) = \sum_{t=1}^{T} \mathbb{1}\left(H_{a,f}^{\mathsf{c}}(t)\right)$ and $C_{a,f}^{\star}(T) = \sum_{t=1}^{T} \mathbb{1}\left(H_{a,f}(t)\right)$.

Proof. To prove this lemma, it suffices to bound the path length $V_{a,f}(T)$ in Theorem E.1.1. We claim that the path length satisfies

$$V_{a,f}(T) \leqslant \min\left\{C_{a,f}^{\star}(T), M_{a,f}^{\star}(T)\right\}.$$

Recall that $\tau_{a,f}(T) = \{t \in T : E_{a,f}^{(\mathsf{c})}(t) = 1\}$. For the remainder of the proof, for any $t \in \tau_{a,f}(T)$, by $t-1$ we mean $\max\{\mathfrak{t} < t : \mathfrak{t} \in \tau_{a,f}(T)\}$.

For any $t \in \tau_{a,f}(T)$, let $\ell_{a,f}^{(\text{prune})}(t)$ denote the loss due to pruning at time t, and let $\ell_{a,f}^{(\text{pull})}(t)$ denote the loss due to pulling at time t. By design of loss function, we know that, for any $t \in \tau_{a,f}(T)$,

$$\ell_{a,f}^{(\text{prune})}(t) = 0, \quad \text{and} \quad \ell_{a,f}^{(\text{pull})}(t) = 1 - 2Y_a(t).$$

Furthermore, note that

$$\begin{aligned}
V_{a,f}(T) &\leqslant \sum_{t \in \tau_{a,f}(T)} |\ell_{a,f}^{(pull)}(t) - \ell_{a,f}^{(pull)}(t-1)| \\
&\underset{(a)}{\leqslant} 2 \sum_{t \in \tau_{a,f}(T)} \mathbb{1}\left(H_{a,f}(t-1), H_{a,f}^{\mathsf{c}}(t)\right) + \mathbb{1}\left(H_{a,f}^{\mathsf{c}}(t-1), H_{a,f}(t)\right) \\
&\leqslant 4 \min\left\{ \sum_{t=1}^{T} \mathbb{1}\left(H_{a,f}^{\mathsf{c}}(t)\right), \sum_{t=1}^{T} \mathbb{1}\left(H_{a,f}(t)\right) \right\} \\
&= 4 \min\left\{ M_{a,f}^{\star}(T), C_{a,f}^{\star}(T) \right\},
\end{aligned}$$

where the factor of 2 in is by the fact that a path length change in going from matching to potential collision or collision to potential matching is 2. The remaining inequalities follow from algebra.

Furthermore, we have

$$\begin{aligned}
V_{a,f}(T) &= \sum_{t \in \tau_{a,f}(T)} \mathbb{1}\left(P_{a,f}(t) = 1, P_{a,f}(t-1) = 1\right) |\ell_{a,f}^{(pull)}(t) - \ell_{a,f}^{(pull)}(t-1)| \\
&\qquad + \sum_{t \in \tau_{a,f}(T)} \mathbb{1}\left(P_{a,f}(t) = 0, P_{a,f}(t-1) = 1\right) |\ell_{a,f}^{(pull)}(t) - \ell_{a,f}^{(pull)}(t-1)| \\
&\leqslant \sum_{t \in \tau_{a,f}(T)} \mathbb{1}\left(P_{a,f}(t) = 1, P_{a,f}(t-1) = 1\right) |\ell_{a,f}^{(pull)}(t) - \ell_{a,f}^{(pull)}(t-1)| \\
&\qquad + 2 \sum_{t \in \tau_{a,f}(T)} \mathbb{1}\left(P_{a,f}(t) = 0, P_{a,f}(t-1) = 1\right) \\
&= \sum_{t \in \tau_{a,f}(T)} \mathbb{1}\left(P_{a,f}(t) = 1, P_{a,f}(t-1) = 1\right) |\ell_{a,f}^{(pull)}(t) - \ell_{a,f}^{(pull)}(t-1)| \\
&\qquad + 2 \sum_{t \in \tau_{a,f}(T)} \mathbb{1}\left(P_{a,f}(t) = 0, P_{a,f}(t-1) = 1\right) \\
&= 2 \sum_{t \in \tau_{a,f}(T)} \mathbb{1}\left(P_{a,f}(t) = 1, P_{a,f}(t-1) = 1, Y_a(t) = 0, Y_a(t-1) = 1\right) \\
&\qquad + 2 \sum_{t \in \tau_{a,f}(T)} \mathbb{1}\left(P_{a,f}(t) = 1, P_{a,f}(t-1) = 1, Y_a(t) = 1, Y_a(t-1) = 0\right) \\
&\qquad + 2 \sum_{t \in \tau_{a,f}(T)} \mathbb{1}\left(P_{a,f}(t) = 0, P_{a,f}(t-1) = 1\right)
\end{aligned}$$

$$\leqslant 2\left(\sum_{t\in\tau_{a,f}(T)} \mathbb{1}\left(P_{a,f}(t)=1, Y_a(t)=0\right) + \mathbb{1}\left(P_{a,f}(t-1)=1, Y_a(t-1)=1\right)\right)$$
$$+ 2\sum_{t\in\tau_{a,f}(T)} \mathbb{1}\left(P_{a,f}(t)=0, P_{a,f}(t-1)=1\right)$$
$$\leqslant 4\left(M_{a,f}(T)+C_{a,f}(T)\right).$$

□

Technical Lemma

Lemma E.1.2. *For any $L\in\mathbb{R}^2$ and $X\in\Delta(\mathbb{R}^2)$, the update*

$$X_+ = \operatorname*{arg\,min}_{Z\in\Delta(\mathbb{R}^2)} \langle Z, L\rangle + D_\psi(Z,X)$$

admits a closed-form solution given by $X_+=[x_+, 1-x_+]$, where

$$x_+ = \frac{2+\xi-\sqrt{4+\xi^2}}{2\xi} \tag{E.1}$$

and $\xi=\eta(L_1-L_2)+\frac{1}{X_1}-\frac{1}{X_2}$. For better interpretability, we provide a plot of the update in Equation (E.1) *in Figure E.1.*

Proof. For any $X, Z\in\Delta(\mathbb{R}^2)$, we represent $X=[x,1-x]$ and $Z=[z,1-z]$ for $x,z\in[0,1]$. Under this notation we can write

$$D_\psi(Z,X) = \frac{1}{\eta}\left(\log\left(\frac{x}{z}\right)+\log\left(\frac{1-x}{1-z}\right)+\frac{z-x}{x}+\frac{x-z}{1-x}\right).$$

Thus the optimization problem becomes

$$\begin{aligned} x_+ &= \operatorname*{arg\,min}_{z\in[0,1]} \langle z, L\rangle + D_\psi(z,X) \\ &= \operatorname*{arg\,min}_{z\in[0,1]} zL_1+(1-z)L_2+\frac{1}{\eta}\left(\log\left(\frac{x}{z}\right)+\log\left(\frac{1-x}{1-z}\right)+\frac{z-x}{x}+\frac{x-z}{1-x}\right) \\ &= \operatorname*{arg\,min}_{z\in[0,1]} zL_1+(1-z)L_2+\frac{1}{\eta}\left(-\log(z)-\log(1-z)+\frac{z}{x}-\frac{z}{1-x}\right). \end{aligned}$$

Let $f(z)=zL_1+(1-z)L_2+\frac{1}{\eta}\left(-\log(z)-\log(1-z)+\frac{z}{x}-\frac{z}{1-x}\right)$. Note that $f(0)=+\infty$, and $f(1)=+\infty$ so the minimizer of $f(z)$ lies stricly inside $[0,1]$. Therefore, $\nabla f(x_+)=0$. We compute

$$\begin{aligned} \nabla f(z) &= L_1-L_2+\frac{1}{\eta(1-z)}-\frac{1}{\eta z}+\frac{1}{\eta x}-\frac{1}{\eta(1-x)} \\ &= L_1-L_2+\frac{2z-1}{\eta z(1-z)}+\frac{1}{\eta x}-\frac{1}{\eta(1-x)}. \end{aligned}$$

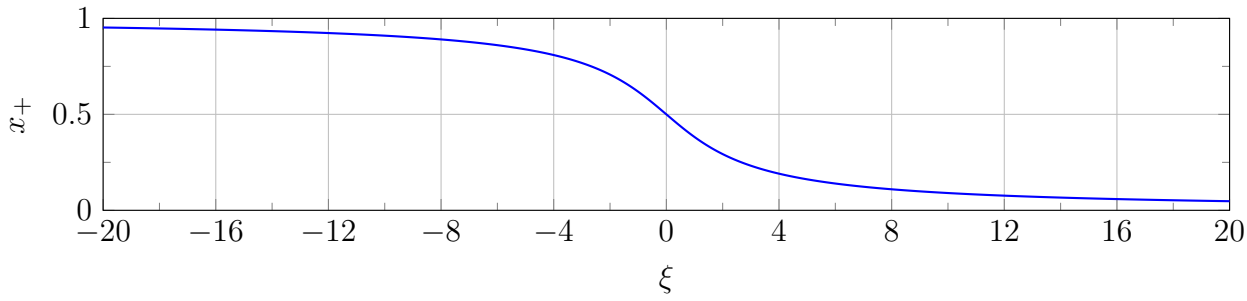


Figure E.1: Update function of pulling probability based on line 10 in Algorithm 7

Imposing the condition $\nabla f(x_+) = 0$ implies that

$$\xi x_+^2 - (2+\xi)x_+ + 1 = 0,$$

where $\xi = \eta(L_1 - L_2) + \frac{1}{x} - \frac{1}{1-x}$. Thus there are two possibilities

$$x_+ = \frac{2+\xi+\sqrt{4+\xi^2}}{2\xi}, \quad \text{or} \quad x_+ = \frac{2+\xi-\sqrt{4+\xi^2}}{2\xi},$$

However, the first possibility implies that $x_+ > 1$, thus the only solution which lies in $(0,1)$ is the latter. This completes the proof. □

E.2 Proofs of main Lemmas

We introduce the following notation for every $a \in A, f \in F$

$$H_{a,f}(t) = \mathbb{1}\left(\exists a' \in A : f_{a'}(t) = f, u_f(a') > u_f(a)\right),$$

which characterizes an event some agent more preferred than a by firm f has requested firm f. We now present the proofs of Lemmas in Chapter 6 in the following subsections.

Proof of Lemma 6.3.2

Proof of Lemma 6.3.2 follows directly from the following Lemma.

Lemma E.2.1. *The event that agent a chooses a firm $f \in F$ at time $t \in [T]$ satisfies*

$$\begin{aligned}&\{Y_a(t) = 1, f_a(t) = f\}\\ &\subset \left\{Y_a(t) = 1, \mathsf{UCB}_{a,f_a^*}(t) \leqslant \mathsf{UCB}_{a,f}(t)\right\} \bigcup \left\{E_{a,f}^{(r)}(t) = 1, E_{a,f_a^*}^{(r)}(t) = 0\right\}. \end{aligned} \tag{E.2}$$

Proof. For any agent a fix some f. Recall that $f_a(t) = f$ implies that agent a has chosen to pull arm f. Based on design of Algorithm 6 there are two possibilities: either all the firms with higher UCB than firm f got pruned and the firm f was requested; or all of the firms in F got pruned and the firm f got selected as it was having highest UCB. Thus,

$$
\begin{aligned}
\{f_a(t) = f\} &= \left\{E^{(\mathsf{r})}_{a,f}(t) = 1\right\} \bigcup \left\{E^{(\mathsf{r})}_{a,f}(t) = 0 \ \forall\ f \in F, \mathsf{UCB}_{a,f} \geqslant \mathsf{UCB}_{a,f'} \ \forall\ f' \in F\right\} \\
&\underset{(i)}{=} \left\{E^{(\mathsf{r})}_{a,f}(t) = 1, \mathsf{UCB}_{a,f^*_a}(t) \geqslant \mathsf{UCB}_{a,f}(t)\right\} \bigcup \left\{E^{(\mathsf{r})}_{a,f}(t) = 1, \mathsf{UCB}_{a,f^*_a}(t) \leqslant \mathsf{UCB}_{a,f}(t)\right\} \\
&\qquad \bigcup \left\{E^{(\mathsf{r})}_{a,f}(t) = 0 \ \forall\ f \in F, \mathsf{UCB}_{a,f} \geqslant \mathsf{UCB}_{a,f'} \ \forall\ f' \in F\right\} \\
&\underset{(ii)}{\subset} \left\{E^{(\mathsf{r})}_{a,f}(t) = 1, \mathsf{UCB}_{a,f^*_a}(t) \geqslant \mathsf{UCB}_{a,f}(t)\right\} \bigcup \left\{E^{(\mathsf{r})}_{a,f}(t) = 1, \mathsf{UCB}_{a,f^*_a}(t) \leqslant \mathsf{UCB}_{a,f}(t)\right\} \\
&\qquad \bigcup \left\{\mathsf{UCB}_{a,f^*_a}(t) \leqslant \mathsf{UCB}_{a,f}(t)\right\} \\
&\underset{(iii)}{\subset} \left\{E^{(\mathsf{r})}_{a,f}(t) = 1, E^{(\mathsf{r})}_{a,f^*_a}(t) = 0, \mathsf{UCB}_{a,f^*_a}(t) \geqslant \mathsf{UCB}_{a,f}(t)\right\} \\
&\qquad \bigcup \left\{E^{(\mathsf{r})}_{a,f}(t) = 1, \mathsf{UCB}_{a,f^*_a}(t) \leqslant \mathsf{UCB}_{a,f}(t)\right\} \bigcup \left\{\mathsf{UCB}_{a,f^*_a}(t) \leqslant \mathsf{UCB}_{a,f}(t)\right\} \\
&\underset{(iv)}{\subset} \left\{E^{(\mathsf{r})}_{a,f}(t) = 1, E^{(\mathsf{r})}_{a,f^*_a}(t) = 0, \mathsf{UCB}_{a,f^*_a}(t) \geqslant \mathsf{UCB}_{a,f}(t)\right\} \bigcup \left\{\mathsf{UCB}_{a,f^*_a}(t) \leqslant \mathsf{UCB}_{a,f}(t)\right\} \\
&\underset{(v)}{\subset} \left\{E^{(\mathsf{r})}_{a,f}(t) = 1, E^{(\mathsf{r})}_{a,f^*_a}(t) = 0\right\} \bigcup \left\{\mathsf{UCB}_{a,f^*_a}(t) \leqslant \mathsf{UCB}_{a,f}(t)\right\},
\end{aligned}
$$

where in (i) we introduced two complementary events $\{\mathsf{UCB}_{a,f^*_a}(t) \geqslant \mathsf{UCB}_{a,f}(t)\}$ and $\{\mathsf{UCB}_{a,f^*_a}(t) \leqslant \mathsf{UCB}_{a,f}(t)\}$. Note that (ii) holds due to the fact that

$$\{\mathsf{UCB}_{a,f_a(t)} \geqslant \mathsf{UCB}_{a,f} \ \forall\ f \in F\}$$

implies $\{\mathsf{UCB}_{a,f_a(t)} \geqslant \mathsf{UCB}_{a,f^*_a}\}$. Furthermore, (iii) holds due to the fact that a firm with lower UCB will be pulled only if all the firms with higher UCB are pruned. Finally, $(iv), (v)$ holds by dropping appropriate events.

The result follows by noting that

$$
\begin{aligned}
&\mathbb{1}\left(Y_a(t) = 1, f_a(t) = f\right) \\
&\subset \left(\left\{E^{(\mathsf{r})}_{a,f}(t) = 1, E^{(\mathsf{r})}_{a,f^*_a}(t) = 0\right\} \bigcup \left\{\mathsf{UCB}_{a,f^*_a}(t) \leqslant \mathsf{UCB}_{a,f}(t)\right\}\right) \bigcap \mathbb{1}\left(Y_a(t) = 1\right) \\
&\subset \left\{Y_a(t) = 1, \mathsf{UCB}_{a,f^*_a}(t) \leqslant \mathsf{UCB}_{a,f}(t)\right\} \bigcup \left\{E^{(\mathsf{r})}_{a,f}(t) = 1, E^{(\mathsf{r})}_{a,f^*_a}(t) = 0\right\}.
\end{aligned}
$$

□

Remark E.2.1. *The results in Lemma E.2.1 holds even if we replace UCB subroutine in Algorithm 6 with any other index based stochastic bandit subroutine, e.g. Thompson sampling.*

Proof of Lemma 6.3.1

We present the proof of each result **(L1)-(L5)** in Lemma 6.3.1 individually in the following subsubsections. Before that we recall an important notation as follows:

$$H_{a,f}(t) = \mathbb{1}\left(\exists\ a' \in A : f_{a'}(t) = f, u_f(a') \geqslant u_f(a)\right) \tag{E.3}$$

Proof of (L1) in Lemma 6.3.1

From (6.2), we get

$$\begin{aligned}\sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i} R_a &\leqslant \bar{\Delta}\sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\sum_{f\in\underline{\mathbb{F}}_a}\mathbb{E}[M_{a,f}(T)] + u\sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\sum_{f\in F\setminus\{f_a^*\}}\mathbb{E}[C_{a,f}(T)] \\ &\quad + \bar{u}\sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\mathbb{E}[C_{a,f_a^*}(T)] \\ &\leqslant \bar{C}\Bigg(\sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\sum_{f\in\underline{\mathbb{F}}_a}\mathbb{E}[M_{a,f}(T)] + \sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\sum_{f\in F\setminus\{f_a^*\}}\mathbb{E}[C_{a,f}(T)] \\ &\quad + \sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\mathbb{E}[\sum_{t=1}^{T} H_{a,f_a^*}(t)]\Bigg),\end{aligned}$$

where $\bar{\Delta} = \max_{a,f}\Delta_a(f)$ and $\bar{u} = \max_a u_a(f_a^*)$. This completes the proof.

Proof of (L2) in Lemma 6.3.1

Proof of **(L2)** in Lemma 6.3.1 follows immediately from the following more general result.

Lemma E.2.2. *For any agent $a \in A$ using Algorithm 6 the expected number of matches with any set $\tilde{F} \subseteq \underline{\mathbb{F}}_a$ can be bounded as*

$$\mathbb{E}[M_{a,\tilde{F}}(T)] \leqslant \mathcal{O}\left(|\tilde{F}|\left(\log(T) + \frac{\log(T)}{\Delta^2}\right) + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f_a^*}(t)\right)\right]\right),$$

where $\Delta = \min_{a,f}\Delta_a(f)$.

Proof. Recall that we say an agent a matches with firm f at time t if $Y_a(t) = 1$ and $f_a(t) = f$. Therefore, the total number of matchings between a and f up to time T is given by

$$M_{a,f}(T) = \sum_{t=1}^{T}\mathbb{1}\left(Y_a(t) = 1,\ f_a(t) = f\right).$$

Hence, from Lemma 6.3.2, the following holds for every $f \in \tilde{F}$:

$$
\begin{aligned}
M_{a,\tilde{F}}(T) &= \sum_{f\in\tilde{F}} \sum_{t=1}^{T} \mathbb{1}\left(Y_a(t)=1, f_a(t)=f\right) \\
&\leqslant \sum_{f\in\tilde{F}} \sum_{t=1}^{T} \Bigg(\mathbb{1}\left(Y_a(t)=1, f_a(t)=f, \mathsf{UCB}_{a,f}(t) \geqslant \mathsf{UCB}_{a,f_a^*}(t)\right) \\
&\qquad + \mathbb{1}\left(E^{(\mathsf{r})}_{a,f}(t)=1, E^{(\mathsf{r})}_{a,f_a^*}=0\right)\Bigg) \\
&\leqslant \sum_{f\in\tilde{F}} \sum_{t=1}^{T} \mathbb{1}\left(Y_a(t)=1, f_a(t)=f, \mathsf{UCB}_{a,f}(t) \geqslant \mathsf{UCB}_{a,f_a^*}(t)\right) \\
&\qquad + \sum_{t=1}^{T} \sum_{f\in\tilde{F}} \mathbb{1}\left(E^{(\mathsf{r})}_{a,f}(t)=1, E^{(\mathsf{r})}_{a,f_a^*}=0\right) \\
&\leqslant \sum_{f\in\tilde{F}} \underbrace{\sum_{t=1}^{T} \mathbb{1}\left(Y_a(t)=1, f_a(t)=f, \mathsf{UCB}_{a,f}(t) \geqslant \mathsf{UCB}_{a,f_a^*}(t)\right)}_{\text{Term A}} + \underbrace{\sum_{t=1}^{T} \mathbb{1}\left(E^{(\mathsf{r})}_{a,f_a^*}=0\right)}_{\text{Term B}}.
\end{aligned}
$$

For any fixed firm $f \in \tilde{F}$, we now bound Term A. For that purpose, define an event

$$
\mathcal{Z}_{a,f}(t) \coloneqq \left\{\mathsf{UCB}_{a,f}(t) \geqslant u_a(f_a^*) - \epsilon\right\} = \left\{\hat{\mu}_{a,f}(t-1) + \sqrt{\frac{2\log(B_a(t))}{M_{a,f}(t-1)}} \geqslant u_a(f_a^*) - \epsilon\right\},
$$

where $B_a(t) \coloneqq 1 + \bar{M}_a(t)\log^2\left(\bar{M}_a(t)\right) \leqslant 1 + t\log^2(t) =: \bar{B}(t)$[3].

Using this notation, we have

$$
\begin{aligned}
\text{Term A} = &\underbrace{\sum_{t=1}^{T} \mathbb{1}(Y_a(t)=1, f_a(t)=f, \mathsf{UCB}_{a,f}(t) \geqslant \mathsf{UCB}_{a,f_a^*}(t), \mathcal{Z}_{a,f}(t))}_{\text{Term C}} \\
&+ \underbrace{\sum_{t=1}^{T} \mathbb{1}(Y_a(t)=1, f_a(t)=f, \mathsf{UCB}_{a,f}(t) \geqslant \mathsf{UCB}_{a,f_a^*}(t), \mathcal{Z}^{\mathsf{c}}_{a,f}(t))}_{\text{Term D}}.
\end{aligned}
$$

[3]The inequality holds due to the fact that $\bar{M}_a(t) \leqslant t$ and monotonicity of the mapping $x \mapsto 1 + x\log^2(x)$.

We shall first bound $\mathbb{E}[\text{Term C}]$ below:

$$
\begin{aligned}
\text{Term C} &= \sum_{t=1}^{T} \mathbb{1}(Y_a(t)=1, f_a(t)=f, \mathsf{UCB}_{a,f}(t) \geqslant \mathsf{UCB}_{a,f_a^*}(t), \mathcal{Z}_{a,f}(t)) \\
&\leqslant \sum_{t=1}^{T} \mathbb{1}(Y_a(t)=1, f_a(t)=f, \mathcal{Z}_{a,f}(t)) \\
&= \sum_{t=1}^{T} \mathbb{1}\left(Y_a(t)=1, f_a(t)=f, \hat{\mu}_{a,f}(t-1) + \sqrt{\frac{2\log(B_a(t))}{M_{a,f}(t-1)}} \geqslant u_a(f_a^*) - \epsilon\right) \\
&\leqslant \sum_{t=1}^{T} \mathbb{1}\left(Y_a(t)=1, f_a(t)=f, \hat{\mu}_{a,f}(t-1) + \sqrt{\frac{2\log(B_a(T))}{M_{a,f}(t-1)}} \geqslant u_a(f_a^*) - \epsilon\right) \\
&= \sum_{t=1}^{T} \sum_{s=0}^{t-1} \mathbb{1}\left(Y_a(t)=1, f_a(t)=f, \hat{\mu}_{a,f}^{(s)} + \sqrt{\frac{2\log(B_a(T))}{s}} \geqslant u_a(f_a^*) - \epsilon, M_{a,f}(t-1)=s\right) \\
&\leqslant \sum_{s=0}^{T-1} \sum_{t=s+1}^{T} \\
&\quad \mathbb{1}\left(f_a(t)=f, \hat{\mu}_{a,f}^{(s)} + \sqrt{\frac{2\log(B_a(T))}{s}} \geqslant u_a(f_a^*) - \epsilon, M_{a,f}(t-1)=s, M_{a,f}(t)=s+1\right) \\
&\leqslant \sum_{s=0}^{T-1} \mathbb{1}\left(\hat{\mu}_{a,f}^{(s)} + \sqrt{\frac{2\log(B_a(T))}{s}} \geqslant u_a(f_a^*) - \epsilon\right) \\
&\leqslant \sum_{s=0}^{T-1} \mathbb{1}\left(\hat{\mu}_{a,f}^{(s)} - u_a(f) + \sqrt{\frac{2\log(\bar{B}(T))}{s}} \geqslant \underbrace{u_a(f_a^*) - u_a(f)}_{\Delta_a(f)} - \epsilon\right),
\end{aligned}
$$

where $\mu_{a,f}^{(s)}$ is defined to be the empirical utility that agent a obtains on s independent successful pulls of arm f. Using Lemma E.4.1 to further bound $\mathbb{E}[\text{Term C}]$ we get

$$
\mathbb{E}[\text{Term C}] \leqslant 1 + \frac{2}{(\Delta_a(f)-\epsilon)^2}\left(\log(\bar{B}(T) + \sqrt{\pi\log(\bar{B}(T))} + 1)\right).
$$

Next, we bound $\mathbb{E}[\text{Term D}]$ below:

$$
\begin{aligned}
&\mathbb{E}[\text{Term D}] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(Y_a(t)=1, f_a(t)=f, \mathsf{UCB}_{a,f}(t) \geqslant \mathsf{UCB}_{a,f_a^*}(t), \mathsf{UCB}_{a,f}(t) \leqslant u_a(f_a^*)-\epsilon\right] \\
&\leqslant \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(Y_a(t)=1, \hat{\mu}_{a,f_a^*}(t-1)+\sqrt{\frac{2\log(B_a(t))}{M_{a,f_a^*}(t-1)}} \leqslant u_a(f_a^*)-\epsilon\right)\right] \\
&\leqslant \sum_{t=1}^{T}\sum_{s=0}^{T-1} \mathsf{Pr}\left(\hat{\mu}_{a,f_a^*}^{(s)}+\sqrt{\frac{2\log(\bar{B}(t))}{s}} \leqslant u_a(f_a^*)-\epsilon\right) \\
&\leqslant \sum_{t=1}^{T}\sum_{s=0}^{T-1} \exp\left(-\frac{s\left(\sqrt{\frac{2\log(\bar{B}(t))}{s}}+\epsilon\right)^2}{2}\right) \\
&\leqslant \sum_{t=1}^{T} \frac{1}{\bar{B}(t)} \sum_{s=1}^{T} \exp\left(-\frac{s\epsilon^2}{2}\right) \\
&\leqslant \frac{\epsilon^2}{2}\sum_{t=0}^{T-1}\frac{1}{\bar{B}(t)},
\end{aligned}
$$

which can further be bounded as $\mathbb{E}[\text{Term D}] \leqslant \frac{5}{\epsilon^2}$ in [228, Exercise 8.1]. For simplicity, we choose $\epsilon = \Delta_a(f)/2$ which ensures that $\mathbb{E}[\text{Term A}] \leqslant \mathcal{O}\left(\frac{\log(T)}{(\Delta_a(f))^2}\right)$.

Now let's turn our attention to Term B which characterizes the number of times agent a has pruned the stable match. Using Lemma E.4.3, we have

$$
\mathbb{E}[\text{Term B}] \leqslant \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(H_{a,f_a^*}(t)\right)\right] + \log(T)\right).
$$

Thus the Term A is bounded by number of there can be potential collisions at the stable firm. This concludes the proof of this lemma. □

Proof of (L3) in Lemma 6.3.1

In this part, we prove a result which is more general than **(L3)** in Lemma 6.3.1.

Lemma E.2.3. *Expected number of collisions faced by agent a on the set of firms $F^\dagger \subseteq F\backslash\{f_a^*\}$*

$$\begin{aligned}
&\sum_{f\in F^\dagger} \mathbb{E}[C_{a,f}(T)] \\
&\leqslant \mathcal{O}\left(|F^\dagger|\log(T) + \mathbb{E}[M_{a,\underline{F}_a^\dagger}(T)] + \mathbb{E}[M_{a,\bar{F}_a^\dagger}(T)] + \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\left(H_{a,f_a^*}(t)\right)\right]\right), \qquad \text{(E.4)}
\end{aligned}$$

where $\underline{F}_a^\dagger = \underline{\mathbb{F}}_a \cap F^\dagger$ and $\bar{F}_a^\dagger = \overline{\mathbb{F}}_a \cap F^\dagger$. Additionally,

$$\mathbb{E}\left[C_{a,f_a^*}(T)\right] \leqslant \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\left(H_{a,f_a^*}(t)\right)\right]. \qquad \text{(E.5)}$$

Proof. To compute the number of collisions, we compute the following for $a \in A$ and $f \in F\backslash\{f_a^*\}$

$$\begin{aligned}
&\sum_{f\in F^\dagger} C_{a,f}(T) = \sum_{f\in F^\dagger}\sum_{t=1}^T \mathbb{1}\left(f_a(t) = f, H_{a,f}(t)\right) \\
&= \sum_{f\in F^\dagger}\sum_{t=1}^T \mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t) = 1, E_{a,f}^{(\mathsf{c})}(t) = 1, H_{a,f}(t)\right) \\
&\quad + \sum_{f\in F^\dagger}\sum_{t=1}^T \mathbb{1}\left(E_{a,f'}^{(\mathsf{r})}(t) = 0 \ \forall\ f' \in F, f_a(t) = f, H_{a,f}(t)\right) \\
&\leqslant \sum_{f\in F^\dagger}\sum_{t=1}^T \mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t) = 1, E_{a,f}^{(\mathsf{c})}(t) = 1, H_{a,f}(t)\right) + \sum_{f\in F^\dagger}\sum_{t=1}^T \mathbb{1}\left(E_{a,f_a^*}^{(\mathsf{r})}(t) = 0, f_a(t) = f\right), \\
&\leqslant \sum_{f\in F^\dagger}\sum_{t=1}^T \mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t) = 1, E_{a,f}^{(\mathsf{c})}(t) = 1, H_{a,f}(t)\right) + \sum_{t=1}^T \mathbb{1}\left(E_{a,f_a^*}^{(\mathsf{r})}(t) = 0\right),
\end{aligned}$$

where the first inequality holds because $\{E_{a,f'}^{(\mathsf{r})}(t) = 0 \ \forall\ f' \in F\}$ implies that $\{E_{a,f_a^*}^{(\mathsf{r})}(t) = 0\}$. Using (E.10) we have: for all $a \in A$, $f \in F$ and $\varpi \in (0, 32\eta) \subset (0,1)$,

$$\begin{aligned}
&\sum_{f\in F^\dagger} \mathbb{E}[C_{a,f}(T)] \\
&\leqslant \sum_{f\in F^\dagger}\left((1+\varpi)\mathbb{E}[M_{a,f}(T)] + \mathcal{O}(\log(T)) + \varpi\mathbb{E}[C_{a,f}(T)] + \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\left(E_{a,f_a^*}^{(\mathsf{r})} = 0\right)\right]\right) \\
&\leqslant \mathcal{O}\left(|F^\dagger|\log(T) + \sum_{f\in F^\dagger}\mathbb{E}[M_{a,f}(T)]\right) + \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\left(H_{a,f_a^*}(t)\right)\right] + \varpi\sum_{f\in F^\dagger}\mathbb{E}[C_{a,f}(T)],
\end{aligned}$$

where the last inequality is due to Lemma E.4.3. In summary,

$$\sum_{f\in F^\dagger}\mathbb{E}[C_{a,f}(T)] \leqslant \mathcal{O}\left(|F|\mathcal{O}(\log(T)) + \sum_{f\in F^\dagger}(\mathbb{E}[M_{a,f}(T)])\right) + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f_a^*}(t)\right)\right]$$
$$\leqslant \mathcal{O}\left(|F^\dagger|\log(T) + \mathbb{E}[M_{a,\underline{F}_a^\dagger}(T)] + \mathbb{E}[M_{a,\bar{F}_a^\dagger}(T)] + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f_a^*}(t)\right)\right]\right).$$

This completes the proof of (E.4). We now prove (E.5). We note that

$$\mathbb{E}\left[C_{a,f_a^*}(T)\right] = \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(f_a(t)=f, H_{a,f_a^*}(t)\right)\right] \leqslant \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f_a^*}(t)\right)\right].$$

This completes the proof. □

Proof of (L4) in Lemma 6.3.1

We restate **(L4)** from Lemma 6.3.1 below:

Lemma E.2.4. *For any $i \in [K]$ we have*

$$\sum_{j=1}^{i}\sum_{a\in\mathcal{A}_j}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f_a^*}(t)\right)\right] = \mathcal{O}\left(C_i|F|\left(\sum_{j=1}^{i}|\mathcal{A}_j|\right)\log(T)\left(1+\frac{1}{\Delta^2}\right)\right),$$

where C_i is a constant dependent on market M_i such that $C_1 < C_2 < ... < C_K$.

Proof. For any $k \in [K]$, define

$$S_k = \sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\mathbb{E}[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f_a^*}(t)\right)],$$

and

$$Z(T,\Delta) = |F|\log(T)\left(1+\frac{1}{\Delta^2}\right).$$

Define $f(\theta;\ell) = \sum_{j=1}^{\ell}\theta^j$, $f(\theta;0) = 1$ and $g(\theta;\ell) = \sum_{j=0}^{\ell-1}\theta^j$. Moreover, let

$$\mathcal{H}_i = \sum_{a\in\mathcal{A}_i}\mathbb{E}[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f_a^*}(t)\right)].$$

Consequently, $S_k = \sum_{i=1}^{k}\mathcal{H}_i$. We claim that

$$S_K \leqslant S_{K-\ell} + f(\theta;\ell)\mathcal{H}_{K-\ell} + \sum_{p=1}^{\ell} g(\theta;p) \sum_{a\in\mathcal{A}_{K-p+1}} \sum_{a'\in\cup_{j=1}^{K-\ell-1}\mathcal{A}_j} \mathbb{E}\left[M_{a',f_a^*}(T)\right]$$
$$+ Z(T,\Delta)\sum_{r=1}^{\ell} f(\theta;r)|\mathcal{A}_{K-r}|. \tag{E.6}$$

We prove this via induction. We first show that this holds for $\ell = 1$. Indeed note that

$$\begin{aligned}
S_K &= S_{K-1} + \mathcal{H}_K = S_{K-1} + \sum_{a\in\mathcal{A}_K} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\Big(H_{a,f_a^*}(t)\Big)\right] \\
&\underset{(a)}{\leqslant} S_{K-1} + \sum_{a\in\mathcal{A}_K} \sum_{a'\in\cup_{j=1}^{K-2}\mathcal{A}_j} \mathbb{E}\left[M_{a',f_a^*}(T)\right] + \sum_{a\in\mathcal{A}_K} \sum_{a'\in\mathcal{A}_{K-1}} \mathbb{E}\left[M_{a',f_a^*}(T)\right] \\
&\underset{(b)}{=} S_{K-1} + \sum_{a\in\mathcal{A}_K} \sum_{a'\in\cup_{j=1}^{K-2}\mathcal{A}_j} \mathbb{E}\left[M_{a',f_a^*}(T)\right] + \sum_{a'\in\mathcal{A}_{K-1}} \sum_{f\in\mathcal{F}_K} \mathbb{E}\left[M_{a',f}(T)\right] \\
&\underset{(c)}{\leqslant} S_{K-1} + \sum_{a\in\mathcal{A}_K} \sum_{a'\in\cup_{j=1}^{K-2}\mathcal{A}_j} \mathbb{E}\left[M_{a',f_a^*}(T)\right] + \sum_{a'\in\mathcal{A}_{K-1}} \mathbb{E}\left[M_{a',\underline{\mathbb{F}}_{a'}}(T)\right] \\
&\underset{(d)}{\leqslant} S_{K-1} + \theta \sum_{a'\in\mathcal{A}_{K-1}} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\Big(H_{a',f_{a'}^*}(t)\Big)\right] + \sum_{a\in\mathcal{A}_K} \sum_{a'\in\cup_{j=1}^{K-2}\mathcal{A}_j} \mathbb{E}\left[M_{a',f_a^*}(T)\right] + \\
&\qquad \theta|\mathcal{A}_{K-1}|Z(T,\Delta) \\
&= S_{K-1} + \theta\mathcal{H}_{K-1} + \sum_{a\in\mathcal{A}_K} \sum_{a'\in\cup_{j=1}^{K-2}\mathcal{A}_j} \mathbb{E}\left[M_{a',f_a^*}(T)\right] + \theta|\mathcal{A}_{K-1}|Z(T,\Delta),
\end{aligned}$$

where the (a) holds due to $\alpha-$reducible structure which says that any agent in $\mathcal{A}_K$ will only get collided at stable arm if some agent from $\cup_{j=1}^{k-1}\mathcal{A}_j$ has also requested the stable firm. Next, (b) holds due to the fact that for any agent $a \in \mathcal{A}_k$, the corresponding stable match $f_a^* \in \mathcal{F}_k$(see Remark 6.1.2). Next, (c) follows because for agents in $\mathcal{A}_{K-1}$, the set of suboptimal firms is super set of F_K. This is again a property of $\alpha-$reducible structure. Finally (d) follows from **(L2)** in Lemma 6.3.1 where θ is the corresponding constant from big-oh notation.

Suppose the bound in (E.6) holds for $\ell = L$ for some integer $\ell \in \{2, 3, ..., K\}$. Then we

show it also holds for $\ell+1$. That is,

$$
\begin{aligned}
S_K &\underset{(a)}{\leqslant} S_{K-\ell} + f(\theta;\ell)\mathcal{H}_{K-\ell} + \sum_{p=1}^{\ell} g(\theta;p) \sum_{a\in\mathcal{A}_{K-p+1}} \sum_{a'\in\cup_{j=1}^{K-\ell-1}\mathcal{A}_j} \mathbb{E}\left[M_{a',f_a^*}(T)\right] \\
&\quad + Z(T,\Delta)\sum_{r=1}^{\ell} f(\theta;r)|\mathcal{A}_{K-r}| \\
&\underset{(b)}{=} S_{K-\ell-1} + g(\theta;\ell+1)\mathcal{H}_{K-\ell} + \sum_{p=1}^{\ell} g(\theta;p) \sum_{a\in\mathcal{A}_{K-p+1}} \sum_{a'\in\cup_{j=1}^{K-\ell-1}\mathcal{A}_j} \mathbb{E}\left[M_{a',f_a^*}(T)\right] \\
&\quad + Z(T,\Delta)\sum_{r=1}^{\ell} f(\theta;r)|\mathcal{A}_{K-r}|
\end{aligned}
$$

$$
\begin{aligned}
&\underset{(c)}{\leqslant} S_{K-\ell-1} + g(\theta;\ell+1)\left(\mathcal{H}_{K-\ell} + \sum_{p=1}^{\ell}\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{a'\in\mathcal{A}_{K-\ell-1}}\mathbb{E}\left[M_{a',f_a^*}(T)\right]\right)\\
&\qquad + \sum_{p=1}^{\ell} g(\theta;p)\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{a'\in\cup_{j=1}^{K-\ell-2}\mathcal{A}_j}\mathbb{E}\left[M_{a',f_a^*}(T)\right] + Z(T,\Delta)\sum_{r=1}^{\ell} f(\theta;r)|\mathcal{A}_{K-r}|\\
&\underset{(d)}{\leqslant} S_{K-\ell-1}\\
&\quad + g(\theta;\ell+1)\left(\sum_{p=1}^{K-\ell-1}\sum_{a'\in\mathcal{A}_p}\sum_{a\in\mathcal{A}_{K-\ell}}\mathbb{E}[M_{a',f_a^*}] + \sum_{p=1}^{\ell}\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{a'\in\mathcal{A}_{K-\ell-1}}\mathbb{E}\left[M_{a',f_a^*}(T)\right]\right)\\
&\qquad + \sum_{p=1}^{\ell} g(\theta;p)\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{a'\in\cup_{j=1}^{K-\ell-2}\mathcal{A}_j}\mathbb{E}\left[M_{a',f_a^*}(T)\right] + Z(T,\Delta)\sum_{r=1}^{\ell} f(\theta;r)|\mathcal{A}_{K-r}|\\
&\underset{(e)}{=} S_{K-\ell-1}\\
&\quad + g(\theta;\ell+1)\left(\sum_{p=1}^{K-\ell-2}\sum_{a'\in\mathcal{A}_p}\sum_{a\in\mathcal{A}_{K-\ell}}\mathbb{E}[M_{a',f_a^*}] + \sum_{p=1}^{\ell+1}\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{a'\in\mathcal{A}_{K-\ell-1}}\mathbb{E}\left[M_{a',f_a^*}(T)\right]\right)\\
&\qquad + \sum_{p=1}^{\ell} g(\theta;p)\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{a'\in\cup_{j=1}^{K-\ell-2}\mathcal{A}_j}\mathbb{E}\left[M_{a',f_a^*}(T)\right] + Z(T,\Delta)\sum_{r=1}^{\ell} f(\theta;r)|\mathcal{A}_{K-r}|\\
&\underset{(f)}{\leqslant} S_{K-\ell-1} + g(\theta;\ell+1)\left(\sum_{p=1}^{K-\ell-2}\sum_{a'\in\mathcal{A}_p}\sum_{a\in\mathcal{A}_{K-\ell}}\mathbb{E}[M_{a',f_a^*}] + \sum_{a'\in\mathcal{A}_{K-\ell-1}}\mathbb{E}\left[M_{a',\underline{\mathbb{F}}_{a'}}(T)\right]\right)\\
&\qquad + \sum_{p=1}^{\ell} g(\theta;p)\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{a'\in\cup_{j=1}^{K-\ell-2}\mathcal{A}_j}\mathbb{E}\left[M_{a',f_a^*}(T)\right] + Z(T,\Delta)\sum_{r=1}^{\ell} f(\theta;r)|\mathcal{A}_{K-r}|\\
&\underset{(g)}{=} S_{K-\ell-1} + g(\theta;\ell+1)\left(\sum_{a'\in\mathcal{A}_{K-\ell-1}}\mathbb{E}\left[M_{a',\underline{\mathbb{F}}_{a'}}(T)\right]\right)\\
&\qquad + \sum_{p=1}^{\ell+1} g(\theta;p)\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{a'\in\cup_{j=1}^{K-\ell-2}\mathcal{A}_j}\mathbb{E}\left[M_{a',f_a^*}(T)\right] + Z(T,\Delta)\sum_{r=1}^{\ell} f(\theta;r)|\mathcal{A}_{K-r}|\\
&\underset{(h)}{\leqslant} S_{K-\ell-1} + g(\theta;\ell+1)\left(\theta|F|Z(T,\Delta)|\mathcal{A}_{K-\ell-1}| + \theta\mathcal{H}_{K-\ell-1}\right)\\
&\qquad + \sum_{p=1}^{\ell+1} g(\theta;p)\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{a'\in\cup_{j=1}^{K-\ell-2}\mathcal{A}_j}\mathbb{E}\left[M_{a',f_a^*}(T)\right] + Z(T,\Delta)\sum_{r=1}^{\ell} f(\theta;r)|\mathcal{A}_{K-r}|\\
&\underset{(i)}{=} S_{K-\ell-1} + f(\theta;\ell+1)\mathcal{H}_{K-\ell-1} + \sum_{p=1}^{\ell+1} g(\theta;p)\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{a'\in\cup_{j=1}^{K-\ell-2}\mathcal{A}_j}\mathbb{E}\left[M_{a',f_a^*}(T)\right]\\
&\qquad + Z(T,\Delta)\sum_{r=1}^{\ell+1} f(\theta;r)|\mathcal{A}_{K-r}|.
\end{aligned}
$$

Here, (a) holds by the induction hypothesis; (b) holds by the definition of S_k, and of the functions $f(\theta;\ell)$ and $g(\theta;\ell)$; (c) follows by rearranging terms and noting that $g(\theta;\cdot)$ is increasing. Next, (d) holds by α-reducibility and the definition of $\mathcal{H}_k$ (using the same analysis as in the base case of the induction). Then, (e) follows by splitting the terms. Next, (f) holds by the definition of α-reducibility. Then, (g) follows by combining similar terms. Next, (h) holds by **(L2)** in Lemma 6.3.1. Finally, (i) follows from combining similar terms.

Thus we conclude that induction claim (E.6) holds true. We know that $S_1 = 0$ therefore from (E.6) we obtain

$$S_k \leqslant Z(T,\Delta)\sum_{r=1}^{K-1} f(\theta;r)|\mathcal{A}_{K-r}| \leqslant \left(\sum_{j=1}^{K-1}|\mathcal{A}_j|\right) K\theta^{K-1}Z(T,\Delta). \tag{E.7}$$

The term $C_k = k\theta^{k-1}$ in the statement. This completes the proof.

□

Proof of (L5) in Lemma 6.3.1

So only thing to bound is matching with superoptimal firms.

Lemma E.2.5. *For any $k \in [K]$ we have*

$$\sum_{j=1}^{k}\sum_{a\in\mathcal{A}_j}\sum_{f\in\overline{\mathbb{F}}_a}\mathbb{E}[M_{a,f}(T)] \leqslant \mathcal{O}\left(C_i\left(\sum_{j=1}^{k-1}|\mathcal{A}_j|\right)|F|\log(T)\left(1+\frac{1}{\Delta^2}\right)\right),$$

where C_i is a constant dependent on market M_i such that $C_1 < C_2 < ... < C_K$.

Proof. For any $k \in [K]$, define

$$\tilde{S}_k = \sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\mathbb{E}[M_{a,\overline{\mathbb{F}}_a}(T)],$$

and

$$Z(T,\Delta) = |F|\log(T)\left(1+1/\Delta^2\right).$$

Define $f(\theta;\ell) = \sum_{j=1}^{\ell}\theta^j$, $f(\theta;0) = 1$ and $g(\theta;\ell) = \sum_{j=0}^{\ell-1}\theta^j$. Let

$$\mathcal{H}_i = \sum_{a\in\mathcal{A}_i}\mathbb{E}[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f_a^*}(t)\right)]$$

and

$$\mathbb{M}_i = \sum_{a\in\mathcal{A}_i}\mathbb{E}[M_{a,\overline{\mathbb{F}}_a}(T)].$$

Then, $\tilde{S}_k = \sum_{i=1}^{k} \mathbb{M}_i$. We claim that

$$\tilde{S}_k \leqslant \mathcal{O}\left(\tilde{\theta}^{k-1}\left(\sum_{j=1}^{k-1}|\mathcal{A}_j|\right)|F|Z(T,\Delta)\right), \tag{E.8}$$

where $\tilde{\theta}$ is a constant greater than 1. Note that the bound holds for $k = 1$ as there is not super-optimal firms for those agents. Let (E.8) holds till some integer $K-1$ then we show that it holds for K as well. We claim that

$$\begin{aligned}\tilde{S}_K \leqslant \tilde{S}_{K-\ell} + f(\tilde{\theta};\ell)\mathbb{M}_{K-\ell} + \sum_{p=1}^{\ell} g(\tilde{\theta};p) \sum_{a\in\mathcal{A}_{K-p+1}} \sum_{f\in\cup_{j\leqslant K-\ell-1}\mathcal{F}_j} \mathbb{E}\left[M_{a,f}\right] + \sum_{p=1}^{\ell} f(\tilde{\theta},p)\mathcal{H}_{K-p} \\ + Z(T,\Delta)\sum_{p=1}^{\ell} f(\tilde{\theta},p)|\mathcal{A}_{K-p}|. \end{aligned} \tag{E.9}$$

We prove (E.8) by induction. First, consider the case $\ell = 1$

$$\begin{aligned}
\tilde{S}_K &= \sum_{i=1}^{K}\sum_{a\in\mathcal{A}_i}\mathbb{E}[M_{a,\bar{\mathbb{F}}_a}(T)] \\
&\underset{(a)}{=} \tilde{S}_{K-1} + \sum_{a\in\mathcal{A}_K}\mathbb{E}[M_{a,\bar{\mathbb{F}}_a}(T)] \\
&\underset{(b)}{\leqslant} \tilde{S}_{K-1} + \sum_{a\in\mathcal{A}_K}\sum_{f\in\cup_{j\leqslant K-2}\mathcal{F}_j}\mathbb{E}[M_{a,f}(T)] + \sum_{a\in\mathcal{A}_K}\sum_{f\in\mathcal{F}_{K-1}}\mathbb{E}[M_{a,f}(T)] \\
&\underset{(c)}{=} \tilde{S}_{K-1} + \sum_{a\in\mathcal{A}_K}\sum_{f\in\cup_{j\leqslant K-2}\mathcal{F}_j}\mathbb{E}[M_{a,f}(T)] + \sum_{a'\in\mathcal{A}_{K-1}}\sum_{a\in\mathcal{A}_K}\mathbb{E}[M_{a,f^*_{a'}}(T)] \\
&\underset{(d)}{\leqslant} \tilde{S}_{K-1} + \tilde{\theta}\sum_{a'\in\mathcal{A}_{K-1}}\mathbb{E}[M_{a',\bar{\mathbb{F}}_{a'}}(T)] + \sum_{a\in\mathcal{A}_K}\sum_{a'\in\cup_{j\leqslant K-2}\mathcal{A}_j}\mathbb{E}[M_{a,f^*_{a'}}(T)] \\
&\qquad + \sum_{a'\in\mathcal{A}_{K-1}}\tilde{\theta}\Big(H_{a',f^*_{a'}} + Z(T,\Delta)\Big) \\
&\underset{(e)}{=} \tilde{S}_{K-1} + \tilde{\theta}\mathbb{M}_{K-1} + \sum_{a\in\mathcal{A}_K}\sum_{f\in\cup_{j\leqslant K-2}\mathcal{F}_j}\mathbb{E}[M_{a,f}(T)] + \tilde{\theta}\mathcal{H}_{K-1} + Z(T,\Delta)\tilde{\theta}|\mathcal{A}_{K-1}|.
\end{aligned}$$

Here, (a) holds by definition. (b) holds by using the α-reducible structure, which ensures that the set of superoptimal firms for any agent lies in markets preceding it. Next, (c) holds by the property of α-reducible markets, which guarantees that for any firm $f \in F_{K-1}$, there exists an agent $a' \in \mathcal{A}_{K-1}$ such that $f = f^*_{a'}$. Then, (d) follows from Lemma E.4.4, and (e) holds by rearranging the terms. Next, we show that if equation (E.8) holds for some ℓ, then

it also holds for $\ell+1$. That is,

$$
\begin{aligned}
\tilde{S}_K &\underset{(a)}{\leqslant} \tilde{S}_{K-\ell} + f(\tilde{\theta};\ell)\mathbb{M}_{K-\ell} + \sum_{p=1}^{\ell} g(\tilde{\theta};p) \sum_{a\in\mathcal{A}_{K-p+1}} \sum_{f\in\cup_{j\leqslant K-\ell-1}\mathcal{F}_j} \mathbb{E}\left[M_{a,f}\right] + \sum_{p=1}^{\ell} f(\tilde{\theta},p)\mathcal{H}_{K-p} \\
&\quad + Z(T,\Delta)\sum_{p=1}^{\ell} f(\tilde{\theta},p)|\mathcal{A}_{K-p}| \\
&\underset{(b)}{=} \tilde{S}_{K-\ell-1} \\
&\quad + g(\tilde{\theta};\ell+1)\mathbb{M}_{K-\ell} + \sum_{p=1}^{\ell} g(\tilde{\theta};p) \sum_{a\in\mathcal{A}_{K-p+1}} \sum_{f\in\cup_{j\leqslant K-\ell-1}\mathcal{F}_j} \mathbb{E}\left[M_{a,f}\right] + \sum_{p=1}^{\ell} f(\tilde{\theta},p)\mathcal{H}_{K-p} \\
&\quad + Z(T,\Delta)\sum_{p=1}^{\ell} f(\tilde{\theta},p)|\mathcal{A}_{K-p}| \\
&\underset{(c)}{=} \tilde{S}_{K-\ell-1} + g(\tilde{\theta};\ell+1)\left(\sum_{a\in\mathcal{A}_{K-\ell}} \sum_{f\in\cup_{j\leqslant K-\ell-2}\mathcal{F}_j} \mathbb{E}\left[M_{a,f}\right] + \sum_{a\in\mathcal{A}_{K-\ell}} \sum_{f\in F_{K-\ell-1}} \mathbb{E}\left[M_{a,f}(T)\right]\right) \\
&\quad + \sum_{p=1}^{\ell} g(\tilde{\theta};p) \sum_{a\in\mathcal{A}_{K-p+1}} \sum_{f\in\cup_{j\leqslant K-\ell-1}\mathcal{F}_j} \mathbb{E}\left[M_{a,f}\right] + \sum_{p=1}^{\ell} f(\tilde{\theta},p)\mathcal{H}_{K-p} \\
&\quad + Z(T,\Delta)\sum_{p=1}^{\ell} f(\tilde{\theta},p)|\mathcal{A}_{K-p}| \\
&\underset{(d)}{\leqslant} \tilde{S}_{K-\ell-1} + g(\tilde{\theta};\ell+1)\left(\sum_{p=1}^{\ell+1} \sum_{a\in\mathcal{A}_{K-p+1}} \sum_{f\in F_{K-\ell-1}} \mathbb{E}\left[M_{a,f}(T)\right]\right) \\
&\quad + \sum_{p=1}^{\ell+1} g(\tilde{\theta};p) \sum_{a\in\mathcal{A}_{K-p+1}} \sum_{f\in\cup_{j\leqslant K-\ell-2}\mathcal{F}_j} \mathbb{E}\left[M_{a,f}\right] + \sum_{p=1}^{\ell} f(\tilde{\theta},p)\mathcal{H}_{K-p} \\
&\quad + Z(T,\Delta)\sum_{p=1}^{\ell} f(\tilde{\theta},p)|\mathcal{A}_{K-p}| \\
&\underset{(e)}{=} \tilde{S}_{K-\ell-1} + g(\tilde{\theta};\ell+1)\left(\sum_{a'\in\mathcal{A}_{K-\ell-1}} \sum_{p=1}^{\ell+1} \sum_{a\in\mathcal{A}_{K-p+1}} \mathbb{E}\left[M_{a,f^*_{a'}}(T)\right]\right) \\
&\quad + \sum_{p=1}^{\ell+1} g(\tilde{\theta};p) \sum_{a\in\mathcal{A}_{K-p+1}} \sum_{f\in\cup_{j\leqslant K-\ell-2}\mathcal{F}_j} \mathbb{E}\left[M_{a,f}\right] + \sum_{p=1}^{\ell} f(\tilde{\theta},p)\mathcal{H}_{K-p} \\
&\quad + Z(T,\Delta)\sum_{p=1}^{\ell} f(\tilde{\theta},p)|\mathcal{A}_{K-p}|
\end{aligned}
$$

$$
\begin{aligned}
&\underset{(f)}{\leqslant} \tilde{S}_{K-\ell-1} + g(\tilde{\theta};\ell+1)\left(\tilde{\theta}\mathcal{H}_{K-\ell-1} + \tilde{\theta}\mathbb{M}_{K-\ell-1} + \tilde{\theta}Z(T,\Delta)|\mathcal{A}_{K-\ell-1}|\right) \\
&\qquad + \sum_{p=1}^{\ell+1} g(\tilde{\theta};p) \sum_{a\in\mathcal{A}_{K-p+1}} \sum_{f\in\cup_{j\leqslant K-\ell-2}\mathcal{F}_j} \mathbb{E}\left[M_{a,f}\right] + \sum_{p=1}^{\ell} f(\tilde{\theta},p)\mathcal{H}_{K-p} \\
&\qquad + Z(T,\Delta)\sum_{p=1}^{\ell} f(\tilde{\theta},p)|\mathcal{A}_{K-p}| \\
&\underset{(g)}{=} \tilde{S}_{K-\ell-1} + f(\tilde{\theta};\ell+1)\mathbb{M}_{K-\ell-1} + \\
&\qquad + \sum_{p=1}^{\ell+1} g(\tilde{\theta};p) \sum_{a\in\mathcal{A}_{K-p+1}} \sum_{f\in\cup_{j\leqslant K-\ell-2}\mathcal{F}_j} \mathbb{E}\left[M_{a,f}\right] + \sum_{p=1}^{\ell+1} f(\tilde{\theta},p)\mathcal{H}_{K-p} \\
&\qquad + Z(T,\Delta)\sum_{p=1}^{\ell+1} f(\tilde{\theta},p)|\mathcal{A}_{K-p}|.
\end{aligned}
$$

Here, (a) follows from the induction hypothesis; (b) is by decomposing $\tilde{S}_{K-\ell}$; (c) follows from the definition of $\mathbb{M}_{K-\ell}$; (d) is by rearranging terms and using the fact that $g(\tilde{\theta},\cdot)$ is increasing. Next, (e) follows from rearranging terms and using the fact that for any $f \in \mathcal{F}_k$ (for some k), there exists $a' \in \mathcal{A}_k$ such that $f = f^*_{a'}$. Then, (f) follows from Lemma E.4.4, and (g) is obtained by combining similar terms. This concludes the induction proof.

We know that $\tilde{S}_1 = \mathbb{M}_1 = 0$ because of $\alpha-$reducible structure which ensures that these firms do not have superoptimal firms. Thus in (E.8) if take $\ell = K-1$ then we get

$$
\begin{aligned}
\tilde{S}_K &\leqslant \sum_{p=1}^{K-1} f(\tilde{\theta},p)\mathcal{H}_{K-p} + Z(T,\Delta)\sum_{p=1}^{K-1} f(\tilde{\theta},p)|\mathcal{A}_{K-p}| \\
&\leqslant \sum_{p=1}^{K-1}\sum_{j=1}^{p} \tilde{\theta}^j\mathcal{H}_{K-p} + Z(T,\Delta)\sum_{p=1}^{K-1} f(\tilde{\theta},p)|\mathcal{A}_{K-p}| \\
&\leqslant \sum_{j=1}^{K-1} \tilde{\theta}^j \sum_{p=j}^{K-1} \mathcal{H}_{K-p} + Z(T,\Delta)\sum_{p=1}^{K-1} f(\tilde{\theta},p)|\mathcal{A}_{K-p}| \\
&\underset{(a)}{=} \sum_{j=1}^{K-1} \tilde{\theta}^j S_{K-j} + Z(T,\Delta)\left(\sum_{j=1}^{K-1}|\mathcal{A}_j|\right) K\tilde{\theta}^{K-1} \\
&\underset{(b)}{\leqslant} Z(T,\Delta)\left(\sum_{j=1}^{K-1}|\mathcal{A}_j|\right)\sum_{j=1}^{K-1}\tilde{\theta}^j(K-j)\theta^{K-j-1} + Z(T,\Delta)\left(\sum_{j=1}^{K-1}|\mathcal{A}_j|\right)K\tilde{\theta}^{K-1},
\end{aligned}
$$

where S_{K-j} in (a) is from proof of **(L4)** in Lemma 6.3.1 and (b) is by (E.7). Define $\tilde{C}_k = k\tilde{\theta}^{k-1} + \sum_{j=1}^{k-1}\tilde{\theta}^j(k-j)\theta^{k-j-1}$. Thus we see that

$$
\tilde{S}_K \leqslant |F|\log(T)\left(1+\frac{1}{\Delta^2}\right)\left(\sum_{j=1}^{K-1}|\mathcal{A}_j|\right)\tilde{C}_K.
$$

□

E.3 Proof of Theorem 6.3.1

We now look at the joint regret for any $k \in [K]$. Before that, we define $Z(T,\Delta) = |F|\log(T)\left(1+\frac{1}{\Delta^2}\right)$. Note that

$$\begin{aligned}
\sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i} R_a &\underset{(a)}{=} \mathcal{O}\Bigg(\sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\mathbb{E}[M_{a,\underline{\mathbb{F}}_a}(T)] + \sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\sum_{f\in F\setminus\{f_a^*\}}\mathbb{E}[C_{a,f}(T)] \\
&\quad + \sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\mathbb{E}[\sum_{t=1}^{T} H_{a,f_a^*}(t)]\Bigg) \\
&\underset{(b)}{=} \mathcal{O}\left(\sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\mathbb{E}[M_{a,\underline{\mathbb{F}}_a}(T)] + \sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\mathbb{E}[M_{a,\overline{\mathbb{F}}_a}(T)] + \sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\mathbb{E}[\sum_{t=1}^{T} H_{a,f_a^*}(t)]\right) \\
&\quad + \mathcal{O}\left(|F|\sum_{i=1}^{k}|\mathcal{A}_i|\log(T)\right) \\
&\underset{(c)}{=} \mathcal{O}\left(\sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\mathbb{E}[M_{a,\overline{\mathbb{F}}_a}(T)] + \sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\mathbb{E}[\sum_{t=1}^{T} H_{a,f_a^*}(t)]\right) + \mathcal{O}(\sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}|\underline{\mathbb{F}}_a|Z(T,\Delta)) \\
&\quad + \mathcal{O}\left(|F|\sum_{i=1}^{k}|\mathcal{A}_i|\log(T)\right) \\
&\underset{(d)}{=} \mathcal{O}(\tilde{C}_k\left(\sum_{p=1}^{k}|A_p|\right)Z(T,\Delta)) + \mathcal{O}(\left(\sum_{p=1}^{k}|A_p|\right)C_k Z(T,\Delta)) + \mathcal{O}(\sum_{p=1}^{k}\sum_{a\in\mathcal{A}_p}|\underline{\mathbb{F}}_a|Z(T,\Delta)) \\
&\quad + \mathcal{O}\left(|F|\sum_{p=1}^{k}|\mathcal{A}_p|\log(T)\right) \\
&\underset{(e)}{=} \mathcal{O}\left((C_k+\tilde{C}_k)|F|\left(\sum_{p=1}^{k}|A_p|\right)\right)\log(T)\left(1+\frac{1}{\Delta^2}\right),
\end{aligned}$$

where (a) holds due to **(L1)** in Lemma 6.3.1, (b) holds due to **(L3)** in Lemma 6.3.1, (c) is due to **(L2)** in Lemma 6.3.1. Next, (d) is due to **(L4)-(L5)** in Lemma 6.3.1. Finally, (e) follows by combining terms.

E.4 Technical Lemmas

In this section we present some technical lemmas which are helpful in the proofs in next section.

Lemma E.4.1. *(Lemma 8.2,[228]) Let $X_1, X_2, \ldots, X_T$ be a sequence of independent 1-subgaussian random variable, and $\hat{\mu}^{(t)} := \frac{1}{t}\sum_{s=1}^{t} X_s, \epsilon > 0, a > 0$ and*

$$\kappa := \sum_{t=1}^{n} \mathbb{1}\left(\hat{\mu}_t + \sqrt{\frac{2a}{t}} \geqslant \epsilon\right), \quad \kappa' := u + \sum_{t=\lceil u \rceil}^{T} \mathbb{1}\left(\hat{\mu}_t + \sqrt{\frac{2a}{t}} \geqslant \epsilon\right),$$

where $u = \frac{2a}{\epsilon^2}$. Then,

$$\mathbb{E}[\kappa] \leqslant \mathbb{E}[\kappa'] \leqslant 1 + \frac{2}{\epsilon^2}(a + \sqrt{\pi a} + 1).$$

Lemma E.4.2. *Suppose we use the AB subroutine Algorithm 7 with $\eta \leqslant 1/50$ then the following two inequalities hold:*

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(E_{a,f}^{(r)}(t) = 1, E_{a,f}^{(c)}(t) = 1, H_{a,f}(t)\right)\right] \\ &\qquad \leqslant (1+\varpi)\mathbb{E}[M_{a,f}(T)] + \mathcal{O}(\log(T)) + \varpi\mathbb{E}[C_{a,f}(T)], \end{aligned} \tag{E.10}$$

where $0 < \varpi \leqslant 32\eta < 1$ and

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(E_{a,f}^{(r)}(t) = 0, E_{a,f}^{(c)}(t) = 1, H_{a,f}^{\mathsf{c}}(t)\right)\right] \\ &\qquad \leqslant \mathcal{O}\left(\log(T) + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(H_{a,f}(t)\right)\right] + \mathbb{E}[C_{a,f}^{\star}(T)]\right). \end{aligned} \tag{E.11}$$

Proof. To simplify the presentation of proof, let's define

$$\begin{aligned} &L_{a,f}^{(\mathsf{adv})}(T) \\ &:= \sum_{t=1}^{T}\left(\mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t) = 1, E_{a,f}^{(\mathsf{c})}(t) = 1, H_{a,f}(t)\right) - \mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t) = 1, E_{a,f}^{(\mathsf{c})}(t) = 1, H_{a,f}^{\mathsf{c}}(t)\right)\right) \end{aligned}$$

The regret bound for adversarial bandit algorithm from Lemma E.1.1 under $\eta \leqslant 1/50$ implies

$$\begin{aligned} &\mathbb{E}\left[L_{a,f}^{(\mathsf{adv})}(T)\right] \\ &\leqslant \mathcal{O}(\log(T)) + \varpi\mathbb{E}\left[\min\left\{M_{a,f}^{\star}(T), C_{a,f}^{\star}(T), M_{a,f}(T) + C_{a,f}(T)\right\}\right] \\ &\mathbb{E}\left[L_{a,f}^{(\mathsf{adv})}(T) - \ell_{a,f}(T)\right] \\ &\leqslant \mathcal{O}(\log(T)) + \varpi\mathbb{E}\left[\min\left\{M_{a,f}^{\star}(T), C_{a,f}^{\star}(T), M_{a,f}(T) + C_{a,f}(T)\right\}\right], \end{aligned} \tag{E.12}$$

where $\varpi \leqslant 32\eta$ and

$$\ell_{a,f}(T) = \sum_{t=1}^{T}\left(\mathbb{1}\left(E_{a,f}^{(\mathsf{c})}(t) = 1, H_{a,f}(t)\right) - \mathbb{1}\left(E_{a,f}^{(\mathsf{c})}(t) = 1, H_{a,f}^{\mathsf{c}}(t)\right)\right),$$

which denotes the total loss received by the adversarial bandit subroutine associated with (a, f) in time T *if* it never take pruning action. Therefore, in (E.12) LHS in first inequality is the regret associated with always pruning. While LHS in second inequality is the regret associated with never pruning.

In the following proof we shall analyze each of the equations in (E.12) separately.

1. The first inequality in (E.12) implies

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t)=1, E_{a,f}^{(\mathsf{c})}(t)=1, H_{a,f}(t)\right)-\mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t)=1, E_{a,f}^{(\mathsf{c})}(t)=1, H_{a,f}^{\mathsf{c}}(t)\right)\right)\right] \leqslant \mathcal{O}(\log(T))+\varpi\left(\mathbb{E}[M_{a,f}(T)+C_{a,f}(T)]\right).$$

This in turn leads to

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T}\left(\mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t)=1, E_{a,f}^{(\mathsf{c})}(t)=1, H_{a,f}(t)\right)\right)\right]\\ &\leqslant \mathbb{E}\left[\mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t)=1, E_{a,f}^{(\mathsf{c})}(t)=1, H_{a,f}^{\mathsf{c}}(t)\right)\right]+\mathcal{O}(\log(T))\\ &+\frac{1}{2}\left(\mathbb{E}[M_{a,f}(T)+C_{a,f}(T)]\right)\\ &\leqslant (1+\varpi)\,\mathbb{E}[M_{a,f}(T)]+\mathcal{O}(\log(T))+\varpi\mathbb{E}[C_{a,f}(T)].\end{aligned}$$

2. Using the definition of $\ell_{a,f}(T)$ in the second inequality in (E.12) we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(-\mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t)=0, E_{a,f}^{(\mathsf{c})}(t)=1, H_{a,f}(t)\right)+\mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t)=0, E_{a,f}^{(\mathsf{c})}(t)=1, H_{a,f}^{\mathsf{c}}(t)\right)\right)\right] \leqslant \mathcal{O}(\log(T)+\mathbb{E}[\min\{M_{a,f}^{\star}(T), C_{a,f}^{\star}(T)\}]),$$

which implies

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t)=0, E_{a,f}^{(\mathsf{c})}(t)=1, H_{a,f}^{\mathsf{c}}(t)\right)\right]\\ &\leqslant \mathcal{O}\Bigg(\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t)=0, E_{a,f}^{(\mathsf{c})}(t)=1, H_{a,f}(t)\right)\right]+\mathcal{O}(\log(T))\\ &\qquad+\mathbb{E}[\min\{M_{a,f}^{\star}(T), C_{a,f}^{\star}(T)\}]\Bigg)\\ &\leqslant \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f}(t)\right)\right]+\log(T)+\mathbb{E}[\min\{M_{a,f}^{\star}(T), C_{a,f}^{\star}(T)\}]\right).\end{aligned}$$

This concludes the proof. □

Lemma E.4.3 (Pruning stable match)**.** *For any $a \in A$,*

$$\underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(E^{(r)}_{a,f^*_a}(t)=0, E^{(c)}_{a,f^*_a}(t)=1\right)\right]}_{\mathbb{E}[\textit{Term I}]} \leqslant \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(H_{a,f^*_a}(t)\right)\right] + \log(T)\right)$$

Proof. We note that

$$\begin{aligned}
\mathbb{E}[\text{Term I}] &\leqslant \mathbb{E}\Bigg[\sum_{t=1}^{T} \mathbb{1}\left(E^{(\mathsf{r})}_{a,f^*_a}(t)=0, E^{(\mathsf{c})}_{a,f^*_a}(t)=1, H_{a,f^*_a}(t)\right) \\
&\qquad + \sum_{t=1}^{T} \mathbb{1}\left(E^{(\mathsf{r})}_{a,f^*_a}(t)=0, E^{(\mathsf{c})}_{a,f^*_a}(t)=1, H^{\mathsf{c}}_{a,f^*_a}(t)\right)\Bigg] \\
&\leqslant \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(H_{a,f^*_a}(t)\right)\right] + \mathcal{O}(\log(T)) + \mathbb{E}[C^{\star}_{a,f^*_a}(T)]\right) \\
&\leqslant \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(H_{a,f^*_a}(t)\right)\right] + \mathcal{O}(\log(T))\right)
\end{aligned}$$

where the first inequality is due to (E.11) and the last inequality holds due to Lemma E.2.3. □

Lemma E.4.4. *For any $a \in A$ and $a' \in A\backslash\{a\}$ we have*

$$\sum_{a' \in A} \mathbb{E}[M_{a',f^*_a}(T)] \leqslant \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(H_{a,f^*_a}(t)\right)\right] + |F| Z(T,\Delta) + \mathbb{E}[M_{a,\overline{\mathbb{F}}_a}(T)]\right)$$

Proof. For any agent $a \in A$ we know that at every time step it either gets matched with some firm or gets collided. This implies

$$\sum_{f' \in F} \mathbb{E}[C_{a,f'}(T)] + \sum_{f' \in F\backslash\{f^*_a\}} \mathbb{E}[M_{a,f'}(T)] + \mathbb{E}[M_{a,f^*_a}(T)] = T. \tag{E.13}$$

Furthermore, in T steps the firm f^*_a can get matched with some agents or remain unmatched. This implies

$$\sum_{a' \in A\backslash\{a\}} \mathbb{E}[M_{a',f^*_a}(T)] + \mathbb{E}[M_{a,f^*_a}(T)] \leqslant T. \tag{E.14}$$

Combining (E.13), (E.14) and Lemma E.2.3 we see that

$$\begin{aligned}
&\sum_{a' \in A} \mathbb{E}[M_{a',f^*_a}(T)] \leqslant \sum_{f' \in F} \mathbb{E}[C_{a,f'}(T)] + \sum_{f' \in F\backslash\{f^*_a\}} \mathbb{E}[M_{a,f'}(T)] \\
&\leqslant \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(H_{a,f^*_a}(t)\right)\right] + |F|\log(T)\right) + \mathcal{O}\left(\mathbb{E}[M_{a,\underline{\mathbb{F}}_a}(T)] + \mathbb{E}[M_{a,\overline{\mathbb{F}}_a}(T)]\right).
\end{aligned}$$

Note that from Lemma E.2.2 we have

$$\sum_{a' \in A} \mathbb{E}[M_{a',f_a^*}(T)] \leqslant \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(H_{a,f_a^*}(t)\right)\right] + |F|\log(T) + |\underline{\mathbb{F}}_a| Z(T,\Delta) + \mathbb{E}[M_{a,\bar{\mathbb{F}}_a}(T)]\right)$$
$$\leqslant \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(H_{a,f_a^*}(t)\right)\right] + |F| Z(T,\Delta) + \mathbb{E}[M_{a,\bar{\mathbb{F}}_a}(T)]\right).$$

This completes the proof.

□

E.5 Thompson Sampling based Decentralized Matching Algorithm

Algorithmic Description

In this section we present a variant of Algorithm 6 but with Thompson sampling based stochastic bandit subroutine. For simplicity, we consider the scenario where the noise in (6.1) is sampled from a normal distribution. To compute the Thompson sampling index each agent a maintains an empirical average of utility generated from any firm f till time t which is $\hat{\mu}_{a,f}(t-1)$. At time step t any agent $a \in A$ will maintain an index of every firm $f \in F$ by sampling it from a normal distribution with mean $\hat{\mu}_{a,f}(t-1)$ and variance $\frac{1}{\sum_{f \in F} M_{a,f}}$ (refer line 3 in Algorithm 16).

Bounds for Algorithm 16

We first present the regret bound for Algorithm 16.

Theorem E.5.1. *Suppose every agent $a \in A$ uses Algorithm 16. Then for any $i \in [K]$:*

$$\sum_{j=1}^{i} \sum_{a \in \mathcal{A}_j} \mathbb{E}[\mathcal{R}_a(T)] = \mathcal{O}\left(C_i |F||A| \left(\frac{1}{\Delta^2}\log\left(\frac{1}{\Delta}\right) + \frac{\log(T)}{\Delta^2} + \log(T)\right)\right),$$

where $\Delta = \min_{a,f} \Delta_{a,f}$ and C_i is a constant dependent on market M_i and $C_1 < C_2 < ... < C_K$.

The only difference between proof of Theorem 6.3.1 and Theorem E.5.1 is the bound on expected number of matchings with suboptimal firms (refer **(L2)** in Lemma 6.3.1). We now present the analogue of **(L2)** of Lemma 6.3.1 below.

Algorithm 16 Thompson Sampling based Decentralized Matching Algorithm (TS-DMA)

Initialize: $\mu_{a,f} \leftarrow 0,\ N_{a,f} \leftarrow 0,\ \pi_{a,f} \leftarrow 0.5,\ q_{a,f} \leftarrow 0.5,\ \ell_{a,f} \leftarrow 0 \quad \forall a \in \mathcal{A},\ f \in \mathcal{F}$
for $t = 1$ to T **do**
  for each $f \in \mathcal{F}$ **do**
    Sample $\theta_{a,f} \sim \mathcal{N}\left(\mu_{a,f}, \frac{1}{\sum_{f' \in \mathcal{F}} N_{a,f'}}\right)$
  end for
  Let $\theta_a \leftarrow$ descending sort of $\{\theta_{a,f}\}_{f \in \mathcal{F}}$
  $i \leftarrow 1$
  while $i \leqslant n$ **do**
    $f \leftarrow \theta_a^{[i]}$
    Sample $z_{a,f} \sim \text{Bernoulli}(\pi_{a,f})$
    if $z_{a,f} = 0$ **then**
      $(q_{a,f}, \pi_{a,f}, \ell_{a,f}) \leftarrow \texttt{PullModule}(z_{a,f}, q_{a,f}, \pi_{a,f}, \ell_{a,f}, \text{matched}_a)$
    else if $z_{a,f} = 1$ **then**
      Query firm f and receive $(r_a, \text{matched}_a)$
      $\mu_{a,f} \leftarrow \text{matched}_a \cdot \frac{\mu_{a,f} \cdot N_{a,f} + r_a}{N_{a,f}+1} + (1 - \text{matched}_a) \cdot \mu_{a,f}$
      $N_{a,f} \leftarrow N_{a,f} + \text{matched}_a$
      $(q_{a,f}, \pi_{a,f}, \ell_{a,f}) \leftarrow \texttt{PullModule}(z_{a,f}, q_{a,f}, \pi_{a,f}, \ell_{a,f}, \text{matched}_a)$
      break
    end if
    $i \leftarrow i + 1$
  end while
  if $i = |\mathcal{F}| + 1$ **then**
    Query firm $\theta_a^{[1]}$ and receive $(r_a, \text{matched}_a)$
    $\mu_{a,f} \leftarrow \text{matched}_a \cdot \frac{\mu_{a,f} \cdot N_{a,f} + r_a}{N_{a,f}+1} + (1 - \text{matched}_a) \cdot \mu_{a,f}$
    $N_{a,f} \leftarrow N_{a,f} + \text{matched}_a$
  end if
end for

Lemma E.5.1. *For any $i \in [K]$, the expected matches with suboptimal firm satisfies*

$$\sum_{j=1}^{i} \sum_{a \in \mathcal{A}_j} \mathbb{E}[M_{a,\underline{\mathbb{F}}_a}(T)]$$
$$= \mathcal{O}\left(\sum_{j=1}^{i} \sum_{a \in \mathcal{A}_j} \left(|\underline{\mathbb{F}}_a| \left(\frac{1}{\Delta^2} \log\left(\frac{1}{\Delta}\right) + \frac{\log(T)}{\Delta^2} + \log(T) \right) + \mathbb{E}\left[\sum_{t=1}^{T} H_{a,f_a^*}(t) \right] \right) \right),$$

where $\Delta = \min_{a,f} \Delta_a(f)$

Proof. Note that we call an agent a matches with firm f at time t if $Y_a(t) = 1$ and $f_a(t) = f$. Therefore the total number of matchings between a and f till time T is $M_{a,f}(T) = \sum_{t=1}^{T} \mathbb{1}\left(Y_a(t) = 1, f_a(t) = f\right)$. Therefore from Lemma E.2.1 and Remark E.2.1 the following holds for every $f \in \underline{\mathbb{F}}_a$:

$$\begin{aligned}
M_{a,\underline{\mathbb{F}}_a}(T) &= \sum_{f \in \underline{\mathbb{F}}_a} \sum_{t=1}^{T} \mathbb{1}\left(Y_a(t) = 1, f_a(t) = f\right) \\
&\leqslant \sum_{f \in \underline{\mathbb{F}}_a} \sum_{t=1}^{T} \left(\mathbb{1}\left(Y_a(t) = 1, f_a(t) = f, \mathcal{T}_{a,f}(t) \geqslant \mathcal{T}_{a,f_a^*}(t)\right) + \mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t) = 1, E_{a,f_a^*}^{(\mathsf{r})} = 0\right) \right) \\
&\leqslant \sum_{f \in \underline{\mathbb{F}}_a} \sum_{t=1}^{T} \mathbb{1}\left(Y_a(t) = 1, f_a(t) = f, \mathcal{T}_{a,f}(t) \geqslant \mathcal{T}_{a,f_a^*}(t)\right) \\
&\qquad + \sum_{t=1}^{T} \sum_{f \in \underline{\mathbb{F}}_a} \mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t) = 1, E_{a,f_a^*}^{(\mathsf{r})} = 0\right) \\
&\leqslant \sum_{f \in \underline{\mathbb{F}}_a} \underbrace{\sum_{t=1}^{T} \mathbb{1}\left(Y_a(t) = 1, f_a(t) = f, \mathcal{T}_{a,f}(t) \geqslant \mathcal{T}_{a,f_a^*}(t)\right)}_{\text{Term A}} + \underbrace{\sum_{t=1}^{T} \mathbb{1}\left(E_{a,f_a^*}^{(\mathsf{r})} = 0\right)}_{\text{Term B}}.
\end{aligned}$$

Let's first analyze Term A. Define $\mathcal{F}_{t-1} = \{\{f_a(\tau), Y_a(\tau), r_a(\tau)\}_{\tau=1}^{t-1}\}_{a \in A}$. We first observe that

$$\begin{aligned}
&\mathbb{1}\left(Y_a(t) = 1, E_{a,f}^{(\mathsf{r})}(t) = 1, E_{a,f}^{(\mathsf{c})}(t) = 1, \mathcal{T}_{a,f_a^*} \leqslant \mathcal{T}_{a,f}(t)\right) \\
&\quad = \underbrace{\mathbb{1}\left(Y_a(t) = 1, E_{a,f}^{(\mathsf{r})}(t) = 1, E_{a,f}^{(\mathsf{c})}(t) = 1, \mathcal{T}_{a,f_a^*} \leqslant \mathcal{T}_{a,f}(t), \mathcal{T}_{a,f}(t) < \hat{\mu}_{a,f_a^*} - \epsilon\right)}_{\text{Term C}} \\
&\quad + \underbrace{\mathbb{1}\left(Y_a(t) = 1, E_{a,f}^{(\mathsf{r})}(t) = 1, E_{a,f}^{(\mathsf{c})}(t) = 1, \mathcal{T}_{a,f_a^*} \leqslant \mathcal{T}_{a,f}(t), \mathcal{T}_{a,f}(t) \geqslant \hat{\mu}_{a,f_a^*} - \epsilon\right)}_{\text{Term D}}.
\end{aligned} \tag{E.15}$$

We first provide a bound on Term C. Prior to that let's define some notations. Let's define $G_{a,f}^{(s)}(\epsilon) = 1 - F_{a,f}^{(s)}(\hat{\mu}_{a,f_a^*} - \epsilon)$. Furthermore, conditioned on the event that atleast

one arm is pulled, for any agent a let's define $\mathcal{P}_a(t)$ to be the set of arms that are pruned before one is chosen to be played at time t. Moreover let $\tilde{A}^{\mathsf{select}}_{a,f}(t)$ be a random variable such that $\tilde{A}^{\mathsf{select}}_{a,f}(t) = 1$ iff f is the firm with maximum index value in all of the non-pruned arms at time t. That is, $\tilde{A}^{\mathsf{select}}_{a,f}(t) = \mathbb{1}\left(f \in \arg\max_{f' \in F \setminus \{\mathcal{P}(t) \cup \{f^*_a\}\}} \mathcal{T}_{a,f'}(t)\right)$. Using this the following holds:

$$\begin{aligned}
&\mathbb{E}[\text{Term C}] = \mathbb{E}[\mathbb{E}[\text{Term C}|\mathcal{F}_{t-1}]] \\
&= \mathbb{E}[\mathsf{Pr}\left(Y_a(t) = 1, E^{(\mathsf{r})}_{a,f}(t) = 1, E^{(\mathsf{c})}_{a,f}(t) = 1, \mathcal{T}_{a,f^*_a} \leqslant \mathcal{T}_{a,f}(t), \mathcal{T}_{a,f}(t) < \hat{\mu}_{a,f^*_a} - \epsilon|\mathcal{F}_{t-1}\right)] \\
&\leqslant \mathbb{E}\left[\mathsf{Pr}\left(\mathcal{T}_{a,f^*_a} < \hat{\mu}_{a,f^*_a} - \epsilon|\mathcal{F}_{t-1}\right)\mathsf{Pr}\left(Y_a(t) = 1, \tilde{A}^{\mathsf{select}}_{a,f}(t) = 1, \mathcal{T}_{a,f}(t) < \hat{\mu}_{a,f^*_a} - \epsilon|\mathcal{F}_{t-1}\right)\right].
\end{aligned} \tag{E.16}$$

Moreover, note that

$$\begin{aligned}
&\mathsf{Pr}\left(Y_a(t) = 1, E^{(\mathsf{c})}_{a,f^*_a}(t) = 1, \mathcal{T}_{a,f}(t)(t) < \hat{\mu}_{a,f^*_a} - \epsilon|\mathcal{F}_{t-1}\right) \\
&\geqslant \mathsf{Pr}\left(Y_a(t) = 1, \tilde{A}^{\mathsf{select}}_{a,f}(t) = 1, \mathcal{T}_{a,f}(t)(t) < \hat{\mu}_{a,f^*_a} - \epsilon, \mathcal{T}_{a,f^*_a}(t) > \hat{\mu}_{a,f^*} - \epsilon|\mathcal{F}_{t-1}\right) \\
&= \mathsf{Pr}\left(\mathcal{T}_{a,f^*_a}(t) > \hat{\mu}_{a,f^*_a}(t-1) - \epsilon|\mathcal{F}_{t-1}\right) &\text{(E.17)} \\
&\quad \cdot \mathsf{Pr}\left(Y_a(t) = 1, \tilde{A}^{\mathsf{select}}_{a,f}(t) = 1, \mathcal{T}_{a,f}(t)(t) < \hat{\mu}_{a,f^*_a} - \epsilon|\mathcal{F}_{t-1}\right). &\text{(E.18)}
\end{aligned}$$

Using (E.17) in (E.16), we obtain the following

$$\begin{aligned}
\mathbb{E}[\text{Term C}] = \mathbb{E}\Bigg[&\frac{\mathsf{Pr}\left(\mathcal{T}_{a,f^*_a} < \hat{\mu}_{a,f^*_a} - \epsilon|\mathcal{F}_{t-1}\right)}{\mathsf{Pr}\left(\mathcal{T}_{a,f^*_a}(t) > \hat{\mu}_{a,f^*_a}(t-1) - \epsilon|\mathcal{F}_{t-1}\right)} \cdot \\
&\mathsf{Pr}\left(Y_a(t) = 1, E^{(\mathsf{c})}_{a,f^*_a}(t) = 1, \mathcal{T}_{a,f}(t)(t) < \hat{\mu}_{a,f^*_a} - \epsilon|\mathcal{F}_{t-1}\right)\Bigg] \\
= \mathbb{E}&\left[\frac{1 - G^{(M_{a,f^*_a}(t-1))}_{a,f^*_a}(\epsilon)}{G^{(M_{a,f^*_a}(t-1))}_{a,f^*_a}(\epsilon)}\mathsf{Pr}\left(Y_a(t) = 1, E^{(\mathsf{c})}_{a,f^*_a}(t) = 1, \mathcal{T}_{a,f}(t)(t) < \hat{\mu}_{a,f^*_a} - \epsilon|\mathcal{F}_{t-1}\right)\right] \\
\leqslant \mathbb{E}&\left[\frac{1 - G^{(M_{a,f^*_a}(t-1))}_{a,f^*_a}(\epsilon)}{G^{(M_{a,f^*_a}(t-1))}_{a,f^*_a}(\epsilon)}\mathsf{Pr}\left(Y_a(t) = 1, E^{(\mathsf{c})}_{a,f^*_a}(t) = 1|\mathcal{F}_{t-1}\right)\right].
\end{aligned}$$

Further, evaluating the expectation of Term C we have:

$$\begin{aligned}
&\mathbb{E}[\text{Term C}] = \sum_{t=1}^{T} \mathbb{E}\left[\frac{1-G_{a,f_a^*}^{(M_{a,f_a^*}(t-1))}(\epsilon)}{G_{a,f_a^*}^{(M_{a,f_a^*}(t-1))}(\epsilon)} \mathbb{1}\left(E_{a,f_a^*}^{(\mathsf{c})}(t)=1, E_{a,f_a^*}^{(\mathsf{r})}(t)=1, Y_a(t)=1\right)\right] \\
&= \sum_{t=1}^{T}\sum_{s=1}^{t} \mathbb{E}\left[\frac{1-G_{a,f_a^*}^{(s)}(\epsilon)}{G_{a,f_a^*}^{(s)}(\epsilon)} \mathbb{1}\left(E_{a,f_a^*}^{(\mathsf{c})}(t)=1, E_{a,f_a^*}^{(\mathsf{r})}(t)=1, Y_a(t)=1, M_{a,f_a^*}(t-1)=s\right)\right] \\
&\leqslant \mathbb{E}\left[\sum_{s=1}^{T} \frac{1-G_{a,f_a^*}^{(s)}(\epsilon)}{G_{a,f_a^*}^{(s)}(\epsilon)} \sum_{t=s+1}^{T} \mathbb{1}\left(M_{a,f}(t-1)=s, M_{a,f}(t)=s+1\right)\right] \\
&\leqslant \sum_{s=0}^{\infty} \frac{1-G_{a,f_a^*}^{(s)}(\epsilon)}{G_{a,f_a^*}^{(s)}(\epsilon)} \leqslant \frac{1}{\epsilon^2}\log(\frac{1}{\epsilon}),
\end{aligned}$$

where the last inequality is due to [228]. Now let's look at Term D. Let's set of time indices when $\mathcal{J}_{a,f} = \{t : G_{a,f}^{(M_{a,f}(t-1))}(\epsilon) > 1/T\}$.

$$\begin{aligned}
&\mathbb{E}[\text{Term D}] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\left(Y_a(t)=1, E_{a,f}^{(\mathsf{r})}(t)=1, E_{a,f}^{(\mathsf{c})}(t)=1, \mathcal{T}_{a,f_a^*} \leqslant \mathcal{T}_{a,f}(t), \mathcal{T}_{a,f}(t) \geqslant \hat{\mu}_{a,f_a^*} - \epsilon\right)\right] \\
&\leqslant \underbrace{\sum_{t\in\mathcal{J}_{a,f}} \mathbb{E}\left[\mathbb{1}\left(Y_a(t)=1, E_{a,f}^{(\mathsf{r})}(t)=1\right)\right]}_{\text{Term E}} + \underbrace{\sum_{t\notin\mathcal{J}_{a,f}} \mathbb{E}\left[\mathbb{1}\left(\mathcal{T}_{a,f}(t) \geqslant \hat{\mu}_{a,f_a^*} - \epsilon\right)\right]}_{\text{Term F}}.
\end{aligned}$$

Let's first analyze the Term E above. Note that

$$\begin{aligned}
&\sum_{t\in\mathcal{J}_{a,f}} \mathbb{1}\left(Y_a(t)=1, E_{a,f}^{(\mathsf{r})}(t)=1\right) \\
&\leqslant \sum_{t=1}^{T}\sum_{s=1}^{t-1} \mathbb{1}\left(Y_a(t)=1, E_{a,f}^{(\mathsf{r})}(t)=1, G_{a,f}^{s}(\epsilon) > \frac{1}{T}, M_{a,f}(t-1)=s, M_{a,f}(t)=s+1\right) \\
&= \sum_{s=0}^{T-1} \mathbb{1}\left(G_{a,f}^{(s)}(\epsilon) > \frac{1}{T}\right) \sum_{t=s+1}^{T} \mathbb{1}\left(M_{a,f}(t-1)=s, M_{a,f}(t)=s+1\right) \\
&= \sum_{s=0}^{T-1} \mathbb{1}\left(G_{a,f}^{(s)}(\epsilon) > \frac{1}{T}\right) \leqslant \mathcal{O}\left(\frac{\log(T)}{(\Delta_{a,f}-\epsilon)^2} + \log(T)\right),
\end{aligned}$$

where the last property is a property of concentration of normal distribution and is standard in frequentist Thompson sampling analysis. For reader's reference we point to the book

[228]. Next, we bound Term F below:

$$\begin{aligned}
&\sum_{t\notin\mathcal{J}_{a,f}} \mathbb{E}\left[\mathbb{1}\left(\mathcal{T}_{a,f}(t) \geqslant \hat{\mu}_{a,f_a^*} - \epsilon\right)\right] = \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\left(\mathcal{T}_{a,f}(t) \geqslant \hat{\mu}_{a,f_a^*} - \epsilon, G_{a,f}^{(M_{a,f}(t-1))}(\epsilon) \leqslant \frac{1}{T}\right)\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left(\mathcal{T}_{a,f}(t) \geqslant \hat{\mu}_{a,f_a^*} - \epsilon, G_{a,f}^{(M_{a,f}(t-1))}(\epsilon) \leqslant \frac{1}{T}\right)\right] |\mathcal{F}_{t-1}\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[G_{a,f}^{(M_{a,f}(t-1))}(\epsilon)\mathbb{1}\left(G_{a,f}^{(M_{a,f}(t-1))}(\epsilon) < \frac{1}{T}\right)\right] \\
&\leqslant 1.
\end{aligned}$$

Combining the bounds on Term C, Term E and Term F and choosing $\epsilon = \frac{\Delta}{2}$ we have

$$\begin{aligned}
\sum_{f\in\underline{\mathbb{F}}_a} \mathbb{E}[M_{a,f}(T)] &\leqslant |\underline{\mathbb{F}}_a|\mathcal{O}\left(\frac{1}{\Delta^2}\log\left(\frac{1}{\Delta}\right) + \frac{\log(T)}{\Delta^2} + \log(T)\right) \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(E_{a,f_a^*}^{(\mathsf{c})}(t) = 1, E_{a,f_a^*}^{(\mathsf{r})}(t) = 0\right)\right] \\
&\leqslant |\underline{\mathbb{F}}_a|\mathcal{O}\left(\frac{1}{\Delta^2}\log\left(\frac{1}{\Delta}\right) + \frac{\log(T)}{\Delta^2} + \log(T)\right) + \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(H_{a,f_a^*}(t)\right)\right]\right),
\end{aligned}$$

where the second inequality is due to Lemma E.4.3. This concludes the proof. □

Appendix F

Appendix for Chapter 7

F.1 Proof of Main Results

Note that proof of Theorem 7.2.1, follows immediately from Lemma 7.2.1 as stated in Section 7.2. Therefore, in this section we provide the proof of Lemma 7.2.1.

Before presenting the proof, we introduce some notations which are crucial for the subsequent exposition. Let Γ_ℓ be the number of time steps in ℓth interval when there is a change in the underlying preference of any player. For $k \in [\Gamma_\ell]$, let t_k^ℓ be the kth time step when there is change in preference of some player. Let $t_0^\ell = t_s^\ell$. Before presenting the proof we introduce few definitions about matching market which are crucial in the proof.

Definition F.1.1 (Set of all Matchings Blocked by $(p_j, a_k, a_{k'})$)**.** *For a fixed preference ordering, the set of all matching in which player p_j is matched to $a_{k'}$ and (p_j, a_k) is a blocking pair is denoted by $B_{j,k,k'}$.*

Definition F.1.2 (Cover of a Set of Matchings)**.** *Let Q denote be set of triplets $(p_j, a_k, a_{k'})$. For a fixed preference ordering of players and set S of matchings, we say Q is a cover of set S, or $Q \in \mathcal{C}(S)$, if $\bigcup_{(p_j,a_k,a_{k'})\in Q} B_{j,k,k'} \supseteq S$.*

Lemma F.1.1 (Restatement of Lemma 7.2.1)**.** *Suppose Assumption 7.1.1 holds. Then, under the RCB algorithm (Algorithm 8), the pessimal regret for a player i between $(\ell-1)$th restart and ℓth restart is given by*

$$R_\ell^i \leqslant \mathcal{O}(KL_\ell H) + \mathcal{O}\left(K\left(1 + \frac{\log(H)}{\Delta^2}\right)\right).$$

Proof. We note that the regret in any interval ℓ can be decomposed as follows:

$$
\begin{aligned}
R_\ell^i &= \mathbb{E}\left[\sum_{t=t_s^\ell}^{t_e^\ell} \left(\mu_{i,t}(\underline{\mathbf{m}}_t(i)) - \mu_{i,t}(m_t(i))\right)\right] \\
&= \mathbb{E}\left[\sum_{t=t_s^\ell}^{t_e^\ell} \sum_{k=1}^{K} \mathbb{I}(m_t(i)=k)\left(\mu_{i,t}(\underline{\mathbf{m}}_t(i)) - \mu_{i,t}(k)\right)\right] \\
&\leqslant \mathbb{E}\left[\sum_{t=t_s^\ell}^{t_e^\ell} \sum_{k=1}^{K} \mathbb{I}(m_t(i)=k)\left|\mu_{i,t}(\underline{\mathbf{m}}_t(i)) - \mu_{i,t}(k)\right|\right] \\
&\leqslant \mathbb{E}\left[2\bar{\mu}\sum_{k=1}^{K}\sum_{t=t_s^\ell}^{t_e^\ell} \mathbb{I}(m_t(i)=k, m_t \text{is unstable})\right] \\
&\leqslant \mathbb{E}\left[2\bar{\mu}\sum_{k=1}^{K}\sum_{q=0}^{\Gamma_\ell-1}\sum_{t=t_q^\ell}^{t_{q+1}^\ell-1} \mathbb{I}(m_t(i)=k, m_t \text{is unstable})\right] \\
&\leqslant \mathbb{E}\Bigg[2\bar{\mu}\Bigg(\underbrace{\sum_{k=1}^{K}\sum_{t=t_0^\ell}^{t_1^\ell-1} \mathbb{I}(m_t(i)=k, m_t \text{is unstable})}_{\text{Term A}} \\
&\qquad + \underbrace{\sum_{k=1}^{K}\sum_{q=1}^{\Gamma_\ell-1}\sum_{t=t_q^\ell}^{t_{q+1}^\ell-1} \mathbb{I}(m_t(i)=k, m_t \text{is unstable})}_{\text{Term B}}\Bigg)\Bigg].
\end{aligned}
$$

First we bound Term A. We note that this term can be bounded similar to analysis in [256] for the setting where the preferences are stationary. This is because for $t \in [t_0^\ell, t_1^\ell - 1]$ the preferences do not change and at $t = t_0^\ell$ the UCB are reintitialized by players as per Algorithm 8. We provide the detailed proof here for the sake of completeness.

For any interval ℓ, player p_i and arm a_k, let $\mathcal{M}_{ik}^\ell$ denote the unstable matchings where player p_i is matched to arm a_k based on the preferences before the first change in the interval ℓ. Furthermore, for any matching m let $T_m([\tilde{t}, \tilde{t}'])$ denote the number of time steps when

$m_t = m$ for $t \in [\tilde{t}, \tilde{t}']$. Note that for any k and $Q^\ell \in \mathcal{C}(\mathcal{M}_{ik}^\ell)$

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=t_0^\ell}^{t_1^\ell - 1} \mathbb{I}(m_t(i) = k, m_t \text{is unstable})\right] &\leqslant \mathbb{E}\left[\sum_{m \in \mathcal{M}_{ik}^\ell} T_m([t_0^\ell, t_1^\ell - 1])\right] \\
&\leqslant \min_{Q^\ell \in \mathcal{C}(\mathcal{M}_{ik}^\ell)} \mathbb{E}\left[\sum_{(p_j, a_s, a_{s'}) \in Q^\ell} \sum_{m \in B_{j,s,s'}} T_m([t_0^\ell, t_1^\ell - 1])\right] \\
&\leqslant \mathbb{E}\left[\sum_{(p_j, a_s, a_{s'}) \in \tilde{Q}^\ell} \sum_{m \in B_{j,s,s'}} T_m([t_0^\ell, t_1^\ell - 1])\right] \\
&\overset{(a)}{\leqslant} \mathbb{E}\left[\sum_{(p_j, a_s, a_{s'}) \in \tilde{Q}^\ell} \left(5 + \frac{6\log(H)}{\Delta_{j,s,s'}^2}\right)\right] \\
&\overset{(b)}{\leqslant} \mathbb{E}\left[\sum_{(p_j, a_s, a_{s'}) \in \tilde{Q}^\ell} \left(5 + \frac{6\log(H)}{\Delta^2}\right)\right] \\
&\leqslant \bar{C}\left(5 + \frac{6\log(H)}{\Delta^2}\right),
\end{aligned}$$

where

$$\tilde{Q}^\ell \in \arg\min_{Q^\ell \in \mathcal{C}(\mathcal{M}_{ik}^\ell)} \mathbb{E}\left[\sum_{(p_j, a_s, a_{s'}) \in Q^\ell} \sum_{m \in B_{j,s,s'}} T_m([t_0^\ell, t_1^\ell - 1])\right],$$

and $\bar{C} = \max_\ell |\tilde{Q}^\ell|$. Here, inequality (a) follows due to the property of UCB estimates (refer [256, Proof of Theorem 3]) and (b) follows from Assumption 7.1.1-(ii). Therefore, we have

$$\mathbb{E}[\text{Term A}] \leqslant \bar{C}K\left(5 + \frac{6\log(H)}{\Delta^2}\right).$$

We next bound Term B. Note that by definition of L_ℓ we have

$$\begin{aligned}
\text{Term B} &= \sum_{k=1}^{K} \sum_{q=1}^{\Gamma_\ell - 1} \sum_{t=t_q^\ell}^{t_{q+1}^\ell - 1} \mathbb{I}(m_t(i) = k, m_t \text{is unstable}) \\
&= \sum_{k=1}^{K} \sum_{t=t_1^\ell}^{t_e^\ell} \mathbb{I}(m_t(i) = k, m_t \text{is unstable}) \leqslant K L_\ell H.
\end{aligned}$$

Thus, we have

$$R_i^\ell \leqslant 2\bar{\mu}\bar{C}K\left(5 + \frac{6\log(H)}{\Delta^2}\right) + 2\bar{\mu}KL_\ell H.$$

□

Appendix G

Appendix for Chapter 8

G.1 Properties of Depth and Height

In the main text, we recursively defined some dynamical quantities, such as the time evolution of the traffic flows $w \in \mathbb{R}^{|A|}$ and the latency-to-go $z \in \mathbb{R}^{|A|}$, in a component-wise fashion, either from the origin of the Condensed DAG G towards the destination, or from the destination to the origin. To facilitate these recursive definitions, we require the following characterizations regarding the depths and heights of arcs in a Condensed DAG G.

Depth

First, we define the concept of depth of a directed acyclic graph (DAG), which will be crucial for the remaining exposition.

Definition G.1.1 (**Depth of a DAG**). *Given a DAG $G = (I, A)$ describing a single-origin single-destination traffic network, the depth of G, denoted $\ell(G)$, is defined by:*

$$\ell(G) := \max_{a \in A} \ell_a.$$

In this work, we consider only acyclic routes in traffic networks with finitely many edges, so we have $\ell(G) < \infty$. Moreover, the case $\ell(G) = 1$ corresponds to a parallel link network, for which the results of the following proposition have already been analyzed in [270]. Therefore, we assume below that $\ell(G) \geqslant 2$.

Proposition G.1.1. *Given a Condensed DAG $G = (I, A)$ with the route set $\boldsymbol{R}$:*

1. *For any $a \in A$, we have $\ell_a = 1$ if and only if $i_a = o$. Similarly, if $\ell_a = \ell(G)$, then $j_a = d$.*
2. *For any fixed $r \in \boldsymbol{R}$, and any $a, a' \in r$ with $\ell_{a,r} < \ell_{a',r}$, we have $\ell_a < \ell_{a'}$ i.e., arcs along a route have strictly increasing depth from the origin to the destination.*

3. *Fix any $a \in A$, and any $r \in \boldsymbol{R}$ containing a such that $\ell_{a,r} = \ell_a$. Then, for any $a' \in \boldsymbol{R}$ preceding a in r, we have $\ell_{a',r} = \ell_{a'}$.*

4. *For each depth $k \in [\ell(G)] := \{1, \cdots, \ell(G)\}$, there exists some $a \in A$ such that $\ell_a = k$.*

Proof.

1. If $\ell_a \neq 1$, then $\ell_a \geqslant 2$, so there exists at least one route $r \in \mathbf{R}$ containing $a \in A$ such that $\ell_{a,r} \geqslant 2$. Thus, $i_a \neq o$ (otherwise the first $\ell_{a,r} - 1$ arcs of r would form a cycle). Conversely, if $i_a \neq o$, then no route $r \in \mathbf{R}$ contains $a \in A$ as its first arc, i.e., $\ell_{a,r} \geqslant 2$ for each $r \in \mathbf{R}$ containing a. Thus, $\ell_a = \max_{r \in \mathbf{R}: a \in r} \ell_{a,r} \geqslant 2$; in particular, $\ell_a \neq 1$. This establishes that $\ell_a = 1$ if and only if $i_a = o$.

   Now, suppose by contradiction that there exists some $a \in A$ such that $\ell_a = \ell(G)$ but $j_a \neq d$. Fix any $r \in \mathbf{R}$ such that $a \in r$ and $\ell_{a,r} = \ell_a$. Then a cannot be at the end of $\mathbf{R}$, since by definition, routes must end at d. Let $a' \in r$ be the arc immediately after a in r. Then $\ell_{a'} \geqslant \ell_{a',r} = \ell_{a,r} + 1 = \ell(G) + 1$, a contradiction to the definition of $\ell(G)$.

2. Fix $r \in \mathbf{R}$, $a, a' \in r$ such that $\ell_{a,r} < \ell_{a',r}$. If $\ell_a = 1$, then $\ell_{a'} \geqslant \ell_{a',r} > \ell_{a,r} = 1 = \ell_a$, and we are done. Suppose $\ell_a \geqslant 2$. By definition of ℓ_a, there exists some route r_2 such that $\ell_{a,r_2} = \ell_a$. Construct a new route $r_3 \in \mathbf{R}$ by replacing the first $\ell_{a,r}$ arcs of r with the first ℓ_{a,r_2} arcs of r_2. Then $\ell_{a'} \geqslant \ell_{a',r_3} = \ell_{a',r} - \ell_{a,r} + \ell_{a,r_2} > \ell_{a,r_2} = \ell_a$.

3. Fix any $a \in A$, and any $r \in \mathbf{R}$ containing a such that $\ell_{a,r} = \ell_a$. Suppose by contradiction that there exists some $a' \in \mathbf{R}$, preceding a in r, for which $\ell_{a'} \geqslant \ell_{a',r} + 1$. Then, by applying the second part of this lemma along the $(\ell_{a,r} - \ell_{a',r})$ arcs of $\mathbf{R}$ from a' to a, we find that $\ell_a \geqslant \ell_{a'} + (\ell_{a,r} - \ell_{a',r}) \geqslant \ell_{a,r} + 1 = \ell_a + 1$, a contradiction.

4. Fix any arc $a \in A$ with $\ell_a = \ell(G)$. Then there exists some $r \in \mathbf{R}$ containing a such that $\ell_{a,r} = \ell_a = \ell(G)$. It follows from the third part of this proposition that, for each $k \in [\ell(G)]$, the k-th arc in $\mathbf{R}$ is of depth k.

□

Height

Next, we define the concept of height of a directed acyclic graph (DAG), which will be crucial for the remaining exposition.

Definition G.1.2 (**Height of a DAG**). *Given a DAG $G = (I, A)$ describing a single-origin single-destination traffic network, the height of G, denoted $m(G)$, is defined by:*

$$m(G) := \max_{a \in A} m_a.$$

Since the traffic network under study is finite, and we consider only acyclic routes, we have $m(G) < \infty$. Moreover, the case $m(G) = 1$ corresponds to a parallel link network, for which the results of the following proposition have already been extensively analyzed in [270]. We will henceforth assume that $m(G) \geqslant 2$.

Proposition G.1.2. *Given an Condensed DAG $G = (I, A)$ with the route set $\boldsymbol{R}$:*

1. *For any $a \in A$, we have $m_a = 1$ if and only if $j_a = d$. Similarly, if $m_a = m(G)$, then $i_a = o$.*
2. *For any fixed $r \in \boldsymbol{R}$, and any $a, a' \in r$ with $m_{a,r} < m_{a',r}$, we have $m_a < m_{a'}$ i.e., arcs along a route from the origin to the destination have strictly decreasing depth.*
3. *Fix any $a \in A$, and any $r \in \boldsymbol{R}$ containing a such that $m_{a,r} = m_a$. Then, for any $a' \in \boldsymbol{R}$ following a in r, we have $m_{a',r} = m_{a'}$.*
4. *For each height $k \in [m(G)] := \{1, \cdots, m(G)\}$, there exists an arc $a \in A$ such that $m_a = k$.*

The proof of Proposition G.1.2 parallels that of Proposition G.1.1, and is omitted for brevity.

G.2 Proofs of Results in Section 8.2

Proof of Lemma 8.2.1 Here, we establish Lemma 8.2.1, restated as follow: The map $F : \mathcal{W} \to \mathbb{R}$, as given below, is strictly convex.

$$F(w) := \sum_{[a] \in A_O} \int_0^{w_{[a]}} \ell_{[a]}(u)\, du + \frac{1}{\beta} \sum_{i \neq d} \left[\sum_{a \in A_i^+} w_a \ln w_a - \left(\sum_{a \in A_i^+} w_a \right) \ln \left(\sum_{a \in A_i^+} w_a \right) \right].$$

For convenience, we define $f_{[a]} : \mathcal{W} \to \mathbb{R}$, $\chi_i : \mathbb{R}^{|A_i^+|} \to \mathbb{R}$, $F : \mathcal{W} \to \mathbb{R}$ for each $[a] \in A_O$, $i \in I \backslash \{d\}$ by:

$$f_{[a]}(w) := \int_0^{w_{[a]}} \ell_{[a]}(u)\, du, \qquad \forall\, [a] \in A_O,$$

$$\chi_i(w_{A_i^+}) := \sum_{a \in A_i^+} w_a \ln w_a - \left(\sum_{a \in A_i^+} w_a \right) \ln \left(\sum_{a \in A_i^+} w_a \right), \qquad \forall\, i \neq I \backslash \{d\},$$

where $w_{A_i^+} \in \mathbb{R}^{|A_i^+|}$ denotes the components of w corresponding to arcs in A_i^+. Then:

$$F(w) = \sum_{[a] \in A_0} f_{[a]}(w) + \frac{1}{\beta} \sum_{i \in I \backslash \{d\}} \chi_i^{\beta}(w).$$

Also, for convenience, define:

$$\mathcal{W}_s := \left\{ w \in \mathbb{R}^{|A|} : \sum_{a \in A_i^+} w_a = \sum_{a \in A_i^-} w_a, \forall\, i \neq o, d, \sum_{a \in A_o^+} w_a = 0. \right\}. \tag{G.1}$$

Essentially, $\mathcal{W}_s$ is the tangent space of the linear manifold with boundary $\mathcal{W}$. Note that, using the notation described in Chapter 8, we can rewrite (G.1) as:

$$\mathcal{W}_s = \left\{ e_{A_i^-} - e_{A_i^+} : i \neq o, d \right\}^{\perp} \cap \left\{ e_{A_o^+} \right\}^{\perp}.$$

We can now establish the strict convexity of F.

We first establish the convexity of F. It suffices to show that $f_{[a]}$ and χ_i are convex for each $[a] \in A_O$, $i \in I \backslash \{d\}$. Note that each $f_{[a]}$ is convex since it is the composition of a convex function ($g(w) = \sum_{a \in A_0} \int_0^{w_a} s_a(u)du$) with a linear function ($w_{[a]} := \sum_{a' \in [a]} w_{a'}$). We show below that χ_i is convex, for each $i \in I \backslash \{d\}$.

Fix $i \in I \backslash \{d\}$. For any $a, a' \in A_i^+$ and each $w \in \mathcal{W}$:

$$\frac{\partial^2 \chi_i}{\partial w_a \partial w_{a'}}(w) = \frac{1}{w_a} \mathbf{1}\{a' = a\} - \frac{1}{\sum_{\bar{a} \in A_i^+} w_{\bar{a}}}.$$

Thus, for any $y \in \mathbb{R}^{|A_i^+|}$:

$$\begin{aligned}
& y^\top \nabla_w^2 \chi_i(w) y \\
&= \sum_{a, a' \in A_i^+} y_a y_{a'} \frac{\partial^2 \chi_i}{\partial w_a \partial w_{a'}}(w) \\
&= \sum_{a \in A_i^+} \frac{y_a^2}{w_a} - \frac{1}{\sum_{\bar{a} \in A_i^+} w_{\bar{a}}} \cdot \sum_{a, a' \in A_i^+} y_a y_{a'} \\
&= \frac{1}{\sum_{\bar{a} \in A_i^+} w_{\bar{a}}} \left(\sum_{\bar{a} \in A_i^+} w_{\bar{a}} \cdot \sum_{a \in A_i^+} \frac{y_a^2}{w_a} - \left(\sum_{a' \in A_i^+} y_{a'} \right)^2 \right) \\
&= \frac{1}{\sum_{\bar{a} \in A_i^+} w_{\bar{a}}} \left(\sum_{\bar{a} \in A_i^+} \left(\sqrt{w_{\bar{a}}} \right)^2 \cdot \sum_{a \in A_i^+} \left(\frac{y_a}{\sqrt{w_a}} \right)^2 - \left(\sum_{a' \in A_i^+} \sqrt{w_{a'}} \cdot \frac{y_{a'}}{\sqrt{w_{a'}}} \right)^2 \right) \\
&\geqslant 0,
\end{aligned} \tag{G.2}$$

where the final inequality follows from the Cauchy-Schwarz inequality. Cauchy-Schwarz also implies that equality holds in (G.2) if and only if the vectors $\left(\sqrt{w_a}\right)_{a \in A_i^+} \in \mathbb{R}^{|A_i^+|}$ and $\left(y_a / \sqrt{w_a}\right)_{a \in A_i^+} \in \mathbb{R}^{|A_i^+|}$ are parallel, i.e., if $(y_a)_{a \in A_i^+}$ and $(w_a)_{a \in A_i^+}$ are scalar multiples of each other. This shows that χ_i is convex, and $dim(N(\nabla_w^2 \chi_i)) = 1$.

Second, suppose by contradiction that F is not strictly convex on $\mathcal{W}$. Then there exists some $\bar{w} \in \mathcal{W}$, $z \in \mathcal{W}_s \backslash \{0\}$ such that:

$$z^\top \nabla_w^2 F(\bar{w}) z = 0.$$

Since $\nabla_w^2 F(\bar{w})$ is symmetric positive semidefinite, this is equivalent to stating that z is in $N(\nabla_w^2 F(\bar{w}))$, the null space of $\nabla_w^2 F(\bar{w})$. Let A_z denote the set of arc indices for which z has a nonzero component, i.e.:

$$A_z := \{a' \in A : z_{a'} \neq 0\}.$$

Since z is not the zero vector, A_z is non-empty. Since there are a discrete and finite number of levels of G, there exists some $a \in A_z$ such that $\ell_a \leqslant \ell_{a'}$ for all $a' \in A_y$, i.e., $\ell_a = \min\{\ell_{a'} : a' \in A_y\}$. Without loss of generality, we consider the case $z_a > 0$ (if not, then replace z with $-z$, which would also be a nonzero vector in $N(\nabla_w^2 F(\bar{w}))$). We claim that $w_a \neq 0$, and that for all $a' \in A_{i_a}^+$:

$$z_{a'} = z_a \cdot \frac{w_{a'}}{w_a} \geqslant 0.$$

To see this, note that otherwise, the vectors $(z_a)_{a \in A_i^+} \in \mathbb{R}^{|A_i^+|}$ and $(w_a)_{a \in A_i^+}$ are not parallel, and so equality cannot be obtained in (G.2), i.e.,:

$$z^\top \nabla_w^2 \chi_i(\bar{w}) z > 0,$$

where, with a slight abuse of notation, we have defined $\chi_i(w) = \chi_i(A_i^+)$. As a result:

$$\begin{aligned}
& z^\top \nabla_w^2 F(\bar{w}) z \\
= & \sum_{[a] \in A} z^\top \nabla_w^2 f_{[a]}(\bar{w}) z + \frac{1}{\beta} \sum_{i' \neq d} z^\top \nabla_w^2 \chi_{i'}(\bar{w}) z \\
\geqslant & \frac{1}{\beta} z^\top \nabla_w^2 \chi_i(\bar{w}) z \\
> & 0,
\end{aligned}$$

a contradiction. Thus, $z_a > 0$, and $z_{a'} \geqslant 0$ for each $a' \in A_{i_a}^+$, so:

$$\sum_{a' \in A_{i_a}^+} z_{a'} > 0.$$

If $\ell_a = 1$, i.e., $i_a = o$, we arrive at a contradiction, since the fact that $z \in \mathcal{W}_s$ implies $\sum_{a' \in A_{i_a}^+} z_{a'} = 0$. If $\ell_a > 1$, we also arrive at a contradiction, since the fact that $z \in \mathcal{W}_s$ implies:

$$\sum_{\hat{a} \in A_{i_a}^-} z_{\hat{a}} = \sum_{a' \in A_{i_a}^+} z_{a'} > 0,$$

so there exists at least one $\ell_{\hat{a}} \in A^-_{i_a}$ with $z_{\hat{a}} > 0$. Then, by definition of $a \in A$, we have $\ell_a \leqslant \ell_{\hat{a}}$; this contradicts Proposition G.1.1, Part 2, which implies that since $\hat{a} \in A^-_{i_a}$, there exists at least one arc containing $\hat{a}$ immediately before $a \in A$, and thus $\ell_{\hat{a}} \leqslant \ell_a - 1$. These contradictions complete the proof of the strict convexity of F on $\mathcal{W}$.

Proof of Theorem 8.2.1 We present the proof of Theorem 8.2.1, restated as follows: The Condensed DAG Equilibrium $\bar{w}^\beta \in \mathcal{W}$ exists, is unique, and is the unique optimal solution to the following convex optimization problem:

$$\min_{w \in \mathcal{W}} \sum_{[a] \in A_0} \int_0^{w_{[a]}} \ell_{[a]}(u)\, dz + \frac{1}{\beta} \sum_{i \neq d} \left[\sum_{a \in A_i^+} w_a \ln w_a - \left(\sum_{a \in A_i^+} w_a \right) \ln \left(\sum_{a \in A_i^+} w_a \right) \right].$$

Proof. (**Proof of Theorem 8.2.1**) The following proof parallels that of Baillon, Cominetti [26, Theorem 2]. Recall that N denotes the set of nodes of the corresponding DAG. The Lagrangian $\mathcal{L} : W \times \mathbb{R}^{|N|-1} \in \mathbb{R}^{|A|} \to \mathbb{R}$ corresponding to the above optimization problem is:

$$\mathcal{L}(w, \mu, \lambda) := \sum_{[a] \in A_0} \int_0^{w_{[a]}} \ell_{[a]}(u)\, dz + \frac{1}{\beta} \sum_{i \neq d} \left[\sum_{a \in A_i^+} w_a \ln w_a - \left(\sum_{a \in A_i^+} w_a \right) \ln \left(\sum_{a \in A_i^+} w_a \right) \right]$$
$$+ \sum_{i \neq d} \mu_i \left(g_i + \sum_{a' \in A_i^-} w_{a'} - \sum_{a' \in A_i^+} w_{a'} \right) + \sum_{a \in A} \lambda_a w_a,$$

with $g_i = g_o \cdot \mathbf{1}\{i = o\}$, where $\mathbf{1}\{\cdot\}$ is the indicator function that returns 1 if the input argument is true, and 0 otherwise. At optimum $(w^\star, \mu^\star) \in \mathcal{W} \times \mathbb{R}^{|N|-1}$, the KKT conditions give, for each $a \in A$:

$$0 = \frac{\partial \mathcal{L}}{\partial w_a}(w^\star, \mu^\star) = \ell_{[a]}(w^\star_{[a]}) + \frac{1}{\beta} \ln \left(\frac{w^\star_a}{\sum_{a' \in A^+_{i_a}} w^\star_{a'}} \right) + \mu^\star_{j_a} - \mu^\star_{i_a} + \lambda_a,$$
$$0 = \lambda_a w_a, \quad \forall\, a \in A.$$

We claim that $(\hat{w}, \hat{\mu}) \in \mathcal{W} \times \mathbb{R}^{|N|-1}$, as given by the Condensed DAG equilibrium definition: For each $a \in A$, $i \in N$:

$$\hat{w}_a = \left(g_{i_a} + \sum_{a' \in A^-_{i_a}} \hat{w}_{a'} \right) \cdot \frac{\exp(-\beta z_a(\hat{w}))}{\sum_{a' \in A^+_{i_a}} \exp(-\beta z_{a'}(\hat{w}))}, \quad a \in A,$$
$$\hat{\mu}_i = \varphi_i(z(\hat{w})) = -\frac{1}{\beta} \ln \left(\sum_{a' \in A_i^+} e^{-\beta z_{a'}(\hat{w})} \right), \quad i \in N,$$
$$\hat{\lambda}_a = 0, \quad \forall\, a \in A,$$

satisfies the KKT conditions stated above. Indeed, we have $\hat{w}_a \geqslant 0$ for each $a \in A$, and:

$$\begin{aligned}
&\frac{\partial \mathcal{L}}{\partial w_a}(\hat{w}, \hat{\mu}, \hat{\lambda}) = \ell_{[a]}(\hat{w}_{[a]}) + \frac{1}{\beta} \ln \left(\frac{\hat{w}_a}{\sum_{a' \in A_{ia}^+} \hat{w}_{a'}} \right) + \hat{\mu}_{j_a} - \hat{\mu}_{i_a} + \sum_{a \in A} \lambda_a \\
=& \ell_{[a]}(\hat{w}_{[a]}) + \frac{1}{\beta} \ln \left(\frac{\exp(-\beta z_a(w))}{\sum_{a' \in A_{ia}^+} \exp(-\beta z_{a'}(w))} \right) + \varphi_{j_a}(z) - \varphi_{i_a}(z) \\
=& \ell_{[a]}(\hat{w}_{[a]}) - z_a(w) + \varphi_{i_a}(z) + \varphi_{j_a}(z) - \varphi_{i_a}(z) \\
=& \ell_{[a]}(\hat{w}_{[a]}) + \varphi_{j_a}(z) - z_a(w) \\
=& 0,
\end{aligned}$$

where the final equality follows from the definition of $(z_a)_{a \in A}$. □

G.3 Proofs for Section 8.3

Proof of Lemma 8.3.1 We present the proof of Lemma 8.3.1, stated formally as follows: Suppose $w(0) \in \mathcal{W}$, and:

$$K_i > \frac{g_o}{C_w} \max\{K_{i_{\hat{a}}} : \hat{a} \in A_i^-\}$$

for each $i \in I \backslash \{d\}$, with C_w given by Lemma 8.3.2. Then, the continuous-time dynamical system (G.6) for the traffic flow $w(t)$ globally asymptotically converges to the corresponding Condensed DAG Equilibrium $\bar{w}^\beta \in \mathcal{W}$.

Proof. (**Proof of Lemma 8.3.1**) We recursively write the continuous-time evolution of the arc flows $w(\cdot)$ as follows, from (8.13) and (8.14). Recall that for any fixed $w \in \mathcal{W}$, at each

non-destination node $i \in I\backslash\{d\}$, we have $\sum_{a' \in A_{ia}^+} w_{a'} = \sum_{\hat{a} \in A_{ia}^-} w_{\hat{a}}$. Thus, for each $a \notin A_o^+$:

$$\begin{aligned}
&\dot{w}_a(t) = \dot{\xi}_a(t) \cdot \left(g_{i_a} + \sum_{\hat{a} \in A_{ia}^-} w_{\hat{a}}(t) \right) + \xi_a(t) \cdot \sum_{\hat{a} \in A_{ia}^-} \dot{w}_a(t) \\
&= K_{i_a} \left(-\xi_a(t) + \frac{\exp(-\beta z_a(w(t)))}{\sum_{a' \in A_{ia}^+} \exp(-\beta z_{a'}(w(t)))} \right) \cdot \left(g_{i_a} + \sum_{\hat{a} \in A_{ia}^-} w_{\hat{a}}(t) \right) + \xi_a(t) \cdot \sum_{\hat{a} \in A_{ia}^-} \dot{w}_{\hat{a}}(t) \\
&= -K_{i_a} w_a(t) + K_{i_a} \cdot \left(g_{i_a} + \sum_{\hat{a} \in A_{ia}^-} w_{\hat{a}}(t) \right) \cdot \frac{\exp(-\beta z_a(w(t)))}{\sum_{a' \in A_{ia}^+} \exp(-\beta z_{a'}(w(t)))} \\
&\qquad + \frac{w_a(t)}{\sum_{a' \in A_{ia}^+} w_{a'}(t)} \cdot \sum_{\hat{a} \in A_{ia}^-} \dot{w}_{\hat{a}}(t) \qquad \text{(G.3)} \\
&= -K_{i_a} \left(1 - \frac{1}{K_{i_a}} \cdot \frac{\sum_{\hat{a} \in A_{ia}^-} \dot{w}_{\hat{a}}}{\sum_{a' \in A_{ia}^+} w_{a'}} \right) w_a \qquad \text{(G.4)} \\
&\qquad + K_{i_a} \cdot \left(g_{i_a} + \sum_{\hat{a} \in A_{ia}^-} w_{\hat{a}}(t) \right) \cdot \frac{\exp(-\beta z_a(w(t)))}{\sum_{a' \in A_{ia}^+} \exp(-\beta z_{a'}(w(t)))}, \qquad \text{(G.5)}
\end{aligned}$$

for each $a \in A$. More formally, we define each component $h : \mathcal{W} \to \mathbb{R}^{|A|}$ recursively as follows. First, for each $a \in A_o^+$, we set:

$$h_a(w) := K_o \left(-w_a + g_o \cdot \frac{\exp(-\beta z_a(w))}{\sum_{a' \in A_o^+} \exp(-\beta z_{a'}(w))} \right).$$

Suppose now that, for some arc $a \in A$, the component $h_a : \mathcal{W} \to \mathbb{R}$ of h has been defined for each $\hat{a} \in A_{i_a}^-$. Then, we set:

$$h_a(w) := -K_{i_a} \left(1 - \frac{1}{K_{i_a}} \cdot \frac{\sum_{\hat{a} \in A_{ia}^-} h_{\hat{a}}(w)}{\sum_{a' \in A_{ia}^+} w_{a'}} \right) w_a + K_{i_a} \cdot \sum_{a' \in A_{i_a}^-} w_{a'} \cdot \frac{\exp(-\beta z_a(w))}{\sum_{a' \in A_o^+} \exp(-\beta z_{a'}(w))}.$$

By iterating through the above definition forward through the Condensed DAG G from origin to destination, we can completely specify each h_a in a well-posed manner (For a more rigorous characterization of this iterative procedure, see Chapter G.1, Proposition G.1.1). We then define the w-dynamics corresponding to the ξ-dynamics (8.13) by:

$$\dot{w} = h(w). \qquad \text{(G.6)}$$

Now, recall the objective $F : \mathcal{W} \times \mathbb{R}^{|A_O|} \to \mathbb{R}$ of the optimization problem that characterizes $\bar{w}^\beta$, first stated in Theorem 8.2.1 as Equation (8.11), reproduced below:

$$F(w) := \sum_{[a] \in A_0} \int_0^{w_{[a]}} \ell_{[a]}(z)\, dz + \frac{1}{\beta} \sum_{i \neq d} \left[\sum_{a \in A_i^+} w_a \ln w_a - \left(\sum_{a \in A_i^+} w_a \right) \ln \left(\sum_{a \in A_i^+} w_a \right) \right].$$

Roughly speaking, our main approach is to show that *F is a Lyapunov function for the best-response dynamics in* (G.6). Let $\mathcal{W}_s$ denote the tangent space to $\mathcal{W}$, and let $\Pi_{\mathcal{W}_s}$ denote the orthogonal projection onto the linear subspace $\mathcal{W}_s$. Under the continuous-time flow dynamics (8.13) and (8.14), if $w \neq \bar{w}^\beta$:

$$\frac{d}{dt}(F \circ w)(t) = \dot{w}(t)^\top \nabla_w F(w(t)) \tag{G.7}$$

$$= \dot{w}(t)^\top \Pi_{\mathcal{W}_s} \nabla_w F(w(t)) \tag{G.8}$$

$$= \dot{w}(t)^\top \Pi_{\mathcal{W}_s} \Big(\nabla_w f(w(t)) + \nabla \chi^\beta(w(t)) \Big)$$

$$= \dot{w}(t)^\top \Pi_{\mathcal{W}_s} \Big(\big(\ell_{[a]}(w_{[a]}(t)) \big)_{a \in A} + \nabla \chi^\beta(w(t)) \Big) \tag{G.9}$$

$$= \dot{w}(t)^\top \Pi_{\mathcal{W}_s} \left[- \nabla \chi^\beta \left(\left(\left(g_{i_a} + \sum_{a' \in A_{i_a}^-} w_{a'}(t) \right) \cdot \frac{\exp(-\beta z_a(w(t)))}{\sum_{\bar{a} \in A_i^+} \exp(-\beta z_{\bar{a}}(w(t)))} \right)_{a \in A} \right) + \nabla \chi^\beta(w(t)) \right] \tag{G.10}$$

$$= \dot{w}(t)^\top \Pi_{\mathcal{W}_s} \left[- \nabla \chi^\beta \left(\left(\left(g_{i_a} + \sum_{a' \in A_{i_a}^-} w_{a'}(t) \right) \cdot \frac{\exp(-\beta z_a(w(t)))}{\sum_{\bar{a} \in A_i^+} \exp(-\beta z_{\bar{a}}(w(t)))} \right)_{a \in A} \right) \right. \tag{G.11}$$

$$\left. + \nabla \chi^\beta \left(\left(\left(1 - \frac{\sum_{a' \in A_{i_a}^-} h_{a'}(w(t))}{K_{i_a} \cdot \sum_{\hat{a} \in A_{i_a}^+} w_{\hat{a}}(t)} \right) \cdot w_a(t) \right)_{a \in A} \right) \right]$$

$$= \left[\left(- K_{i_a} \left(1 - \frac{\sum_{a' \in A_{i_a}^-} h_{a'}(w(t))}{K_{i_a} \cdot \sum_{\hat{a} \in A_{i_a}^+} w_{\hat{a}}(t)} \right) w_a(t) \right. \right. \tag{G.12}$$

$$\left. \left. + K_{i_a} \left(g_{i_a} + \sum_{a' \in A_{i_a}^+} w_{a'}(t) \right) \cdot \frac{\exp(-\beta z_a(w(t)))}{\sum_{\bar{a} \in A_{i_a}^+} \exp(-\beta z_{\bar{a}}(w(t)))} \right)_{a \in A} \right]^\top \tag{G.13}$$

$$\left[- \nabla \chi^\beta \left(\left(\left(g_{i_a} + \sum_{a' \in A_{i_a}^-} w_{a'}(t) \right) \cdot \frac{\exp(-\beta z_a(w(t)))}{\sum_{\bar{a} \in A_i^+} \exp(-\beta z_{\bar{a}}(w(t)))} \right)_{a \in A} \right) \right.$$

$$\left. + \nabla \chi^\beta \left(\left(\left(1 - \frac{\sum_{a' \in A_{i_a}^-} h_{a'}(w(t))}{K_{i_a} \cdot \sum_{\hat{a} \in A_{i_a}^+} w_{\hat{a}}(t)} \right) \cdot w_a(t) \right)_{a \in A} \right) \right]$$

$$< 0. \tag{G.14}$$

We explain the equalities (G.7) = (G.8), (G.9) = (G.10), (G.10) = (G.11), and (G.12) < (G.14) below.

Verifying (G.7) = (G.8) From the equations leading up to (G.4), we have, for each $w \in \mathcal{W}$:

$$w_a = \left(g_{i_a} + \sum_{\hat{a} \in A_{i_a}^-} w_{a'} \right) \cdot \xi_a,$$

and so:

$$\dot{w}_a(t) = \left(g_{i_a} + \sum_{\hat{a}\in A^-_{i_a}} w_{\hat{a}}(t)\right)\cdot\dot{\xi}_a + \sum_{\hat{a}\in A^-_{i_a}} \dot{w}_{a'}\cdot\xi_a$$
$$= \left(g_{i_a} + \sum_{\hat{a}\in A^-_{i_a}} w_{\hat{a}}(t)\right)\cdot K_{i_a}\left(-\xi_a + \frac{\exp(-\beta z_a(w(t)))}{\sum_{a'\in A^+_{i_a}}\exp(-\beta z_{a'}(w(t)))}\right) + \sum_{\hat{a}\in A^-_{i_a}} \dot{w}_{a'}\cdot\xi_a$$

Fix any node $i \in I$ in the Condensed DAG, and consider the sum of the above equation over the arc subset A^+_i:

$$\sum_{a'\in A^+_i} \dot{w}_{a'}(t) = \sum_{\hat{a}\in A^-_i} \dot{w}_{\hat{a}}(t).$$

Since $w(0) \in \mathcal{W}$ by assumption, we have the initial condition $(\sum_{\hat{a}\in A^+_i} w_{\hat{a}} - \sum_{a'\in A^-_{i_a}} w_{a'} - g_i)(0) = 0$ for the above linear time-invariant differential equation. We thus conclude that, for each $t \geqslant 0$:

$$\sum_{\hat{a}\in A^+_i} w_{\hat{a}}(t) - \sum_{a'\in A^-_{i_a}} w_{a'}(t) - g_i = 0.$$

Since this holds for any arbitrary node $i \in I$, we have $w(t) \in \mathcal{W}$ for all $t \geqslant 0$.

Verifying (G.9) = (G.10) We will show that:

$$\Pi_{\mathcal{W}_s}\left[\left(\ell_{[a]}(w_{[a]}(t))\right)_{a\in A} + \nabla\chi^\beta\left(\left(\left(g_{i_a} + \sum_{a'\in A^-_{i_a}} w_{a'}(t)\right)\cdot\frac{\exp(-\beta z_a(w(t)))}{\sum_{\bar{a}\in A^+_i}\exp(-\beta z_{\bar{a}}(w(t)))}\right)_{a\in A}\right)\right] \tag{G.15}$$
$$= 0, \tag{G.16}$$

which would a fortiori establish the desired claim (G.9) = (G.10). To do so, first note that, for each $i \neq d$, $a \in A^+_i$:

$$\frac{\partial\chi^\beta}{\partial w_a}(w) = \frac{1}{\beta}\cdot\left[\ln w_a + 1 - \ln\left(\sum_{a\in A^+_i} w_a\right) - 1\right] = \frac{1}{\beta}\ln\left(\frac{w_a}{\sum_{a\in A^+_i} w_a}\right). \tag{G.17}$$

Thus, we have:

$$\begin{aligned}
&\frac{\partial \chi^\beta}{\partial w_a}\left(\left(\left(g_{i_a}+\sum_{a'\in A_{i_a}^-} w_{a'}\right)\cdot\frac{\exp(-\beta z_a(w))}{\sum_{\bar a\in A_{i_a}^+}\exp(-\beta z_{\bar a}(w))}\right)_{a\in A}\right)\\
=&\frac{1}{\beta}\ln\left(\frac{\exp(-\beta z_a(w))}{\sum_{\bar a\in A_i^+}\exp(-\beta z_{\bar a}(w))}\right)\\
=&-z_a(w)-\frac{1}{\beta}\ln\left(\sum_{\bar a\in A_{i_a}^+}\exp(-\beta z_{\bar a}(w))\right)\\
=&-z_a(w)+\varphi_{i_a}(w).
\end{aligned}$$

Concatenating these partial derivatives to form the gradient, we can now verify (G.15) by observing that:

$$\begin{aligned}
&\Pi_{\mathcal{W}_s}\left[\left(\ell_{[a]}(w_{[a]})\right)_{a\in A}+\nabla\chi^\beta\left(\left(\left(g_{i_a}+\sum_{a'\in A_{i_a}^-} w_{a'}\right)\cdot\frac{\exp(-\beta z_{\hat a}(w))}{\sum_{\bar a\in A_i^+}\exp(-\beta z_{\bar a}(w))}\right)_{\hat a\in A}\right)_{a\in A}\right]\\
=&\Pi_{\mathcal{W}_s}\left(\ell_{[a]}(w_{[a]})-z_a(w)+\varphi_{i_a}(w)\right)_{a\in A}\\
=&\Pi_{\mathcal{W}_s}\left(\varphi_{i_a}(w)-\varphi_{j_a}(w)\right)_{a\in A}\\
=&\Pi_{\mathcal{W}_s}\left[\sum_{a\in A}\varphi_{i_a}(w)e_a-\sum_{a\in A}\varphi_{j_a}(w)e_a\right]\\
=&\Pi_{\mathcal{W}_s}\left[-\sum_{\hat a\in A_d^-}\varphi_{j_{\hat a}}(w)e_{\hat a}+\sum_{a'\in A_o^+}\varphi_{i_{a'}}(w)e_{a'}\right.\\
&\qquad\left.+\sum_{i\neq\{o,d\}}\left(\sum_{a'\in A_i^+}\varphi_i(w)e_{a'}-\sum_{\hat a\in A_i^-}\varphi_i(w)e_{\hat a}\right)\right]\\
=&\Pi_{\mathcal{W}_s}\left[0+\varphi_o(w)e_{A_o^+}+\sum_{i\neq\{o,d\}}\varphi_i(w)\left(e_{A_i^-}-e_{A_i^+}\right)\right]\\
=&0,
\end{aligned}$$

where the last equality follows by definition of $\Pi_{\mathcal{W}_s}$.

Verifying (G.10) = (G.11) We will show that:

$$\nabla\chi^\beta(w)=\nabla\chi^\beta\left(\left(\left(1-\frac{\sum_{a'\in A_{i_a}^-}\dot w_{a'}}{K_{i_a}\cdot\sum_{\hat a\in A_{i_a}^+}w_{\hat a}}\right)\cdot w_a\right)_{a\in A}\right),$$

which is equivalent to showing that (G.10) = (G.11). From (G.17), we have for each $a \in A$:

$$\begin{aligned}
&\frac{\partial \chi^\beta}{\partial w_a}\left(\left(\left(1-\frac{\sum_{a'\in A_{i_a}^-} h_{a'}(w)}{K_{i_a}\cdot\sum_{\hat{a}\in A_{i_a}^+} w_{\hat{a}}}\right)\cdot w_a\right)_{a\in A}\right)\\
=&\frac{1}{\beta}\ln\left(\frac{\left(1-\frac{\sum_{a'\in A_{i_a}^-} h_{a'}(w)}{K_{i_a}\cdot\sum_{\hat{a}\in A_{i_a}^+} w_{\hat{a}}}\right)w_a}{\sum_{\bar{a}\in A_{i_a}^+}\left(1-\frac{\sum_{a'\in A_{i_{\bar{a}}}^-} h_{a'}(w)}{K_{i_a}\cdot\sum_{\hat{a}\in A_{i_{\bar{a}}}^+} w_{\hat{a}}}\right)w_{\bar{a}}}\right)\\
=&\frac{1}{\beta}\ln\left(\frac{\left(1-\frac{\sum_{a'\in A_{i_a}^-} h_{a'}(w)}{K_{i_a}\cdot\sum_{\hat{a}\in A_{i_a}^+} w_{\hat{a}}}\right)w_a}{\left(1-\frac{\sum_{a'\in A_{i_a}^-} h_{a'}(w)}{K_{i_a}\cdot\sum_{\hat{a}\in A_{i_a}^+} w_{\hat{a}}}\right)\cdot\sum_{\bar{a}\in A_{i_a}^+} w_{\bar{a}}}\right)\\
=&\frac{1}{\beta}\ln\left(\frac{w_a}{\sum_{\bar{a}\in A_{i_a}^+} w_{\bar{a}}}\right)\\
=&\frac{\partial \chi^\beta}{\partial w_a}(w).
\end{aligned}$$

The second equality above follows because, for each $\bar{a} \in A_{i_a}^+$, we have $i_{\bar{a}} = i_a$. This verifies that (G.10) = (G.11).

Verifying (G.12) $<$ (G.14), $\forall w \neq \bar{w}^\beta$ Suppose $\frac{d}{dt}(F \circ w) = 0$ at some $\tilde{w} \in \mathcal{W}$. From (G.12), and by the definition of the convex function χ:

$$\begin{aligned}
0 =& \frac{d}{dt}(F\circ w)\\
=& \sum_{i\in I\setminus\{d\}}\left(\left[-\left(1-\frac{\sum_{\hat{a}\in A_{i_a}^-} h_{\hat{a}}(w)}{K_{i_a}\cdot\sum_{a'\in A_{i_a}^+}\tilde{w}_{a'}}\right)\tilde{w}_a+\left(g_{i_a}+\sum_{a'\in A_{i_a}^+}\tilde{w}_{a'}\right)\frac{\exp(-\beta\cdot z_a(\tilde{w}))}{\sum_{a'\in A_{i_a}^+}\exp(-\beta\cdot z_{a'}(\tilde{w}))}\right]_{a\in A}\right)^\top\\
&\cdot\frac{1}{\beta}\left(\nabla\chi_i^\beta\left(\left[\left(1-\frac{\sum_{\hat{a}\in A_{i_a}^-} h_{\hat{a}}(\tilde{w})}{K_{i_a}\cdot\sum_{a'\in A_{i_a}^+}\tilde{w}_{a'}}\right)\tilde{w}_a\right]_{a\in A}\right)\right.\\
&\left.-\nabla\chi_i^\beta\left(\left[\left(g_{i_a}+\sum_{a'\in A_{i_a}^+}\tilde{w}_{a'}\right)\cdot\frac{\exp(-\beta\cdot z_a(\tilde{w}))}{\sum_{a'\in A_{i_a}^+}\exp(-\beta\cdot z_{a'}(\tilde{w}))}\right]_{a\in A}\right)\right),
\end{aligned}$$

where, for each $i \in I\setminus\{d\}$, the convex map $\chi_i^\beta : \mathbb{R}^{|A_i^+|} \to \mathbb{R}$ is defined by:

$$\chi_i^\beta(\{w_a : a \in A_i^+\}) = \sum_{a\in A_i^+} w_a \ln w_a - \left(\sum_{a\in A_i^+} w_a\right)\ln\left(\sum_{a\in A_i^+} w_a\right).$$

The convexity of each χ_i^β implies that each of the above summands must be non-positive; since they sum to 0, each summand must be 0. (By assumption, $K_{i_a} > (g_o/C_w)\max\{K_{i_{\hat{a}}} : \hat{a} \in A_{i_a}^-\}$, so the input arguments to th $\nabla\chi_i(\cdot)$ maps are all strictly positive.)

Now, for each $i \in I\backslash\{d\}$ and each $w \in \mathbb{R}^{A_i^+}$, we have $N(\nabla^2\chi_i^\beta(w)) = \text{span}\{w\}$. In words, χ_i^β increases linearly only along rays emanating from the origin. In the context of the above summands, this implies that, for each $i \in I\backslash\{d\}$, there exists constants $Q_i \in \mathbb{R}$ such that, for each $a \in A_i^+$:

$$Q_i\left(1 - \frac{\sum_{\hat{a}\in A_{i_a}^-} h_{\hat{a}}(w)}{K_{i_a}\cdot\sum_{a'\in A_{i_a}^+} w_{a'}}\right)w_a = \left(g_{i_a} + \sum_{a'\in A_{i_a}^+} w_{a'}\right)\cdot\frac{\exp(-\beta\cdot z_a(w))}{\sum_{a'\in A_{i_a}^+}\exp(-\beta\cdot z_{a'}(w))}.$$

By definition of $h : \mathcal{W} \to \mathbb{R}^{|A|}$, for each $a \in A_o^+$:

$$h_a(w) = -\tilde{w}_a + g_o\cdot\frac{\exp(-\beta z_a(w))}{\sum_{a'\in A_o^+}\exp(-\beta z_{a'}(w))} = (Q_o - 1)w_a$$

and for each $a \in A_i^+$ with $i \neq o$:

$$\begin{aligned} h_a(w) &:= -\left(1 - \frac{\sum_{\hat{a}\in A_{i_a}^-} h_{\hat{a}}(w)}{\sum_{a'\in A_{i_a}^+} w_{a'}}\right)w_a + \sum_{a'\in A_{i_a}^-} w_{a'}\cdot\frac{\exp(-\beta z_a(w))}{\sum_{a'\in A_o^+}\exp(-\beta z_{a'}(w))} \\ &= (Q_o - 1)\left(1 - \frac{\sum_{\hat{a}\in A_{i_a}^-} h_{\hat{a}}(w)}{\sum_{a'\in A_{i_a}^+} w_{a'}}\right)w_a. \end{aligned}$$

By the flow continuity equations:

$$0 = \sum_{a'\in A_o^+} h_{a'}(w) = (Q_o - 1)\cdot\sum_{a'\in A_i^+} w_{a'} = (Q_o - 1)g_o,$$

so $Q_o = 1$, and thus $h_a(w) = 0$ for each $a \in A_o^+$. Now, suppose there exists some depth $\ell \in [\ell(G) - 1]$ such that $\ell_a(w) = 0$ for each $a \in A$ such that $\ell_a \leqslant \ell$. Then, for each $a \in A$ such that $\ell_a = \ell + 1$, the flow continuity equations give:

$$0 = \sum_{a'\in A_i^+} h_{a'}(w) - \sum_{\hat{a}\in A_i^-} h_{\hat{a}}(w) = \sum_{a'\in A_i^+} h_{a'}(w) = (Q_{i_a} - 1)\cdot\sum_{a'\in A_{i_a}^+} w_{a'}.$$

Thus, $Q_{i_a} = 1$, so $h_a(w) = 0$. This completes the recursion step, and shows that $h(w) = 0$, i.e., $w = \bar{w}^\beta$.

In summary, we established that the map F strictly decreases along any trajectory that starts in $\mathcal{W}\backslash\{\bar{w}^\beta\}$ and follows the best-response dynamics (G.6). The convergence of the dynamics (G.6) to the Condensed DAG equilibrium (8.2.1) now follows by invoking either Sandholm, Corollary 7.B.6 [362], or Sastry, Proposition 5.22 and Theorem 5.23 (LaSalle's Principle and its corollaries) [365]. □

Proof of Lemma 8.3.2 To prove Lemma 8.3.2, we require the following results. We first establish bounds on the trajectory of discrete-time and continuous-time traffic flow dynamics.

Lemma G.3.1.

1. *Consider the discrete-time dynamics:*

$$Y[n+1] = (1-\delta[n])Y[n] + \delta[n]F[n],$$

where, for each $n \geqslant 0$, we have $\delta[n] \in (0,1)$ and $Y[0], F[n] \in [a,b]$ for some $a, b \in \mathbb{R}$ satisfying $a < b$. Then $Y[n] \in [a,b]$ for each $n \geqslant 0$.

2. *Consider the continuous-time dynamics:*

$$\dot{y}(t) = K(-y(t) + f(t)),$$

where $K > 0$ and, for each $t \geqslant 0$, we have $y(0), f(t) \in [a,b]$ for some $a, b \in \mathbb{R}$ satisfying $a < b$. Then $y(t) \in [a,b]$ for each $t \geqslant 0$.

Proof.

1. Suppose there exists some $N \geqslant 0$ such that $Y[n] \in [a,b]$ for each $n \leqslant N$. (Since $Y[0] \in [a,b]$ by assumption, this is certainly true for $n = 0$). Then:

$$\begin{aligned} Y[n+1] &= (1-\delta[n])Y[n] + \delta[n]F[n] \\ &\in [(1-\delta[n]) \cdot a + \delta[n] \cdot a, (1-\delta[n]) \cdot b \\ &\qquad + \delta[n] \cdot b] \\ &= [a,b]. \end{aligned}$$

   This completes the induction step, and thus the proof of this part of the lemma.

2. We compute:

$$\frac{d}{dt}(e^{Kt}y(t)) = e^{Kt}\big(\dot{y}(t) + ay(t)\big) = ae^{Kt}f(t).$$

   Integrating from time 0 to time t, we have, for each $t \geqslant 0$:

$$e^{Kt}y(t) - y(0) = \int_0^t ae^{K\tau}f(\tau)\,d\tau.$$

   Rearranging terms, we obtain, for each $t \geqslant 0$:

$$\begin{aligned} y(t) &= e^{-Kt}y(0) + e^{-Kt}\int_0^t ae^{a\tau}f(\tau)\,d\tau \\ &\in \left[e^{-Kt}a + (1-e^{-Kt})a, e^{-Kt}b + (1-e^{-Kt})b\right] \\ &= [a,b]. \end{aligned}$$

□

Before proceeding, we rewrite the discrete ξ-dynamics (PBR) as a Markov process with a martingale difference term:

$$\xi_a[n+1] = \xi_a[n] + \mu\left(K_{i_a}\left(-\xi_a[n] + \frac{\exp(-\beta\big[z_a(W[n])\big])}{\sum\limits_{a'\in A_{i_a}^+}\exp(-\beta\big[z_{a'}(W[n])\big])}\right) + M_a[n+1]\right),$$

with:

$$M_a[n+1] := \left(\frac{1}{\mu}\eta_{i_a}[n+1]-1\right)\cdot K_{i_a}\left(-\xi_a[n] + \frac{\exp(-\beta\big[z_a(W[n])\big])}{\sum_{a'\in A_{i_a}^+}\exp(-\beta\big[z_{a'}(W[n])\big])}\right).$$

The following lemma bounds the magnitude of the discrete-time flow $W[n] \in \mathbb{R}^{|w|}$ and the martingale difference terms $M[n] \in \mathbb{R}^{|w|}$.

Lemma G.3.2. *The discrete-time dynamics* (PBR) *and* (8.12) *satisfy:*

1 For each $a \in A$: $\{M_a[n+1] : n \geqslant 0\}$ is a martingale difference sequence with respect to the filtration $\mathcal{F}_n := \sigma\Big(\cup_{a\in A}\left(W_a[1], \xi[1], \cdots, W_a[n], \xi[n]\right)\Big)$.

2 There exist $C_w, C_m > 0$ such that, for each $a \in A$ and each $n \geqslant 0$, we have:

$$\begin{aligned} W_a[n] &\in [C_w, g_o], \\ |M_a[n]| &\leqslant C_m. \end{aligned}$$

The continuous-time dynamics (8.13) *and* (8.14) *satisfy:*

3 For each $a \in A$, $n \geqslant 0$: $w_a(t) \in [C_w, g_o]$.

Proof.

1. We have:

$$\begin{aligned} &\mathbb{E}[M_a[n+1]|\mathcal{F}_n] \\ &= \left(\frac{1}{\mu}\mathbb{E}[\eta_{i_a}[n+1]]-1\right)\cdot K_{i_a}\left(-\xi_a[n] + \frac{\exp(-\beta\big[z_a(W[n])\big])}{\sum_{a'\in A_{i_a}^+}\exp(-\beta\big[z_{a'}(W[n])\big])}\right) \\ &= 0. \end{aligned}$$

2. We separate the proof of this part of the lemma into the following steps.

- First, we show that for each $a \in A$, $n \geqslant 0$, we have $\xi_a[n] \in (0, 1]$.

  Fix $a \in A$ arbitrarily. Then $\xi_a[0] \in (0, 1]$ by assumption, and for each $n \geqslant 0$:

$$\frac{\exp(-\beta\Big[z_a(W[n])\Big])}{\sum_{a' \in A_{ia}^+} \exp(-\beta\Big[z_{a'}(W[n])\Big])} \in (0, 1],$$

  since the exponential function takes values in $(0, \infty)$. Thus, by Lemma G.3.1, we have $\xi_a[n] \in (0, 1]$ for each $n \geqslant 0$.

- Second, we show that for each $a \in A$, $n \geqslant 0$, we have $W_a[n] \in (0, g_o]$.

  Note that (8.12), together with the assumption that $W[0] \in \mathcal{W}$, implies that $W[n] \in \mathcal{W}$ for each $n \geqslant 0$. Now, fix $a \in A$, $n \geqslant 0$ arbitrarily. Let $\mathbf{R}(a) \subseteq \mathbf{R}$ denote the set of all routes passing through a, and for each $r \in \mathbf{R}(a)$, let $a_{r,k}$ denote the k-th arc in r. Then, by the conservation of flow encoded in $\mathbf{R}$:

$$\begin{aligned} W_a[n] &= g_o \cdot \sum_{r \in \mathbf{R}(a)} \prod_{k=1}^{|r|} \xi_{a_{r,k}} \\ &\leqslant g_o \cdot \sum_{r \in \mathbf{R}} \prod_{k=1}^{|r|} \xi_{a_{r,k}} \\ &= g_o. \end{aligned}$$

  Similarly, since $\xi_a[n] \in (0, 1]$ for each $a \in A$, $n \geqslant 0$, we have:

$$W_a[n] = g_o \cdot \sum_{r \in \mathbf{R}(a)} \prod_{k=1}^{|r|} \xi_{a_{r,k}} > 0.$$

- Third, we show that there exists $C_z > 0$ such that $|z_a(W[n])| \leqslant C_z$ for each $a \in A$, $n \geqslant 0$. Fix $a \in A_d^- = \{a \in A : m_a = 1\}$ arbitrarily. Then, from (8.6):

$$\begin{aligned} z_a(w) = \ell_{[a]}(w_{[a]}) \in [0, \ell_{[a]}(g_o)], \\ \implies |z_a(w)| \leqslant \ell_{[a]}(g_o) := C_{z,1}. \end{aligned}$$

  Now, suppose that at some height $k \in [m(G) - 1]$, there exists some $C_{z,k} > 0$ such that, for each $n \geqslant 0$, and each $a \in A$ satisfying $m_a \leqslant k$ and each $n \geqslant 0$, we have $|z_a(w)| \leqslant C_{z,k}$. Then, for each $n \geqslant 0$, and each $a \in A$ satisfying $m_a = k + 1$ (at least one such $a \in A$ must exist, by Proposition G.1.2, Part 4):

$$\begin{aligned} z_a(w) &= \ell_{[a]}(w_{[a]}) - \frac{1}{\beta} \ln \left(\sum_{a' \in A_{ja}^+} e^{-\beta \cdot z_{a'}(w)} \right) \\ &\leqslant \ell_{[a]}(g_o) - \frac{1}{\beta} \ln \left(|A_{ja}^+| e^{-\beta \cdot C_z} \right) \\ &= \ell_{[a]}(g_o) + C_z, \end{aligned}$$

and:

$$\begin{aligned} z_a(w) &= \ell_{[a]}(w_{[a]}) - \frac{1}{\beta}\ln\left(\sum_{a'\in A_{ja}^+} e^{-\beta\cdot z_{a'}(w)}\right) \\ &\geqslant 0 + 0 - \frac{1}{\beta}\ln\left(|A_{ja}^+|e^{\beta\cdot C_z}\right) \\ &= -\frac{1}{\beta}\ln|A| - C_z, \end{aligned}$$

from which we conclude that:

$$\begin{aligned} |z_a(w)| &\leqslant \max\left\{\ell_{[a]}(g_o) + C_z, \frac{1}{\beta}\ln|A| + C_z\right\} \\ &:= C_{z,k+1}, \end{aligned}$$

with $C_{z+1} \geqslant C_z$. This completes the induction step, and the proof is completed by taking $C_z := C_{z,m(G)}$.

- Fourth, we show that there exists some $C_\xi > 0$ such that $\xi_a[n] \geqslant C_\xi$ for each $a \in A$, $n \geqslant 0$.

  Define:

$$C_\xi := \min\left\{\min_{a'\in A}\xi_{a'}[0], \frac{1}{|A|}e^{-2\beta C_z}\right\} \in (0,1).$$

  By definition of C_ξ, we have $\xi_a[0] \geqslant C_\xi$. Moreover, for each $n \geqslant 0$, we have:

$$\begin{aligned} &\frac{\exp(-\beta\Big[z_a(W[n])\Big])}{\sum_{a'\in A_{ia}^+}\exp(-\beta\Big[z_{a'}(W[n])\Big])} \\ \geqslant &\frac{e^{-\beta C_z}}{|A_{ia}^+|\cdot e^{\beta C_z}} \\ \geqslant &\frac{1}{|A|}e^{-2\beta C_z} \\ \geqslant &C_\xi. \end{aligned}$$

  Thus, by Lemma G.3.1, $\xi_a[n] \geqslant C_\xi$ for each $n \geqslant 0$.

- Fifth, we show that there exists $C_w > 0$ such that, for each $a \in A$, $n \geqslant 0$, we have $W_a[n] \geqslant C_w$.

  Fix $a \in A$, $n \geqslant 0$. Let $\mathbf{R}(a)$ denote the set of all routes in the Condensed DAG containing a, and let $r \in \mathbf{R}(a)$ be arbitrarily given. By unwinding the recursive

definition of $W_a[n]$ from the flow dispersion probability values $\{\xi_a[n] : a \in A, n \geqslant 0\}$, we have:

$$\begin{aligned} W_a[n] &= g_o \cdot \sum_{\substack{r' \in \mathbf{R} \\ a \in r'}} \prod_{a' \in r'} \xi_{a'}[n] \\ &\geqslant g_o \cdot \prod_{a' \in r} \xi_{a'}[n] \\ &\geqslant g_o \cdot (C_\xi)^{|r|} \\ &\geqslant g_o \cdot (C_\xi)^{\ell(G)} \\ &:= C_w. \end{aligned}$$

- Sixth, we show that there exists $C_m > 0$ such that, for each $a \in A$, $n \geqslant 0$, we have $M_a[n] \geqslant C_m$.

  Define, for convenience, $C_\mu := \max\{\overline{\mu} - \mu, \mu - \underline{\mu}\}$. Since $\eta_{i_a}[n] \in [\underline{\mu}, \overline{\mu}]$, we have from (8.16) that for each $a \in A$, $n \geqslant 0$:

$$\begin{aligned} &M_a[n+1] \\ &:= \left(\frac{1}{\mu}\eta_{i_a}[n+1] - 1\right) \\ &\quad \cdot K_{i_a}\left(-\xi_a[n] + \frac{\exp(-\beta\Big[z_a(W[n])\Big])}{\sum_{a' \in A^+_{i_a}} \exp(-\beta\Big[z_{a'}(W[n])\Big])}\right). \end{aligned}$$

  Thus, by the triangle inequality:

$$\begin{aligned} |M_a[n+1]| &\leqslant \frac{1}{\mu} C_\mu K_{i_a} \cdot (1+1) \\ &\leqslant \frac{2}{\mu} C_\mu \cdot \max_{i \in I \setminus \{d\}} K_i \\ &:= C_m. \end{aligned}$$

3. We separate the proof of this part of the lemma into the following steps.

- First, we show that for each $a \in A$, $t \geqslant 0$, we have $\xi_a(t) \in (0, 1]$.

  Fix $a \in A$. By assumption, $\xi_a(0) \in (0, 1]$, and at each $t \geqslant 0$:

$$\frac{\exp(-\beta z_a(w))}{\sum_{a' \in A^+_{i_a}} \exp(-\beta z_{a'}(w))} \in (0, 1].$$

  Thus, by Lemma G.3.1, we conclude that $\xi_a(t) \in (0, 1]$ for each $t \geqslant 0$.

- Second, we show that $w_a(t) \in [0, g_o]$ for each $t \geqslant 0$.

  The proof here is nearly identical to the proof that $W_a[n] \in (0, g_o)$ in the second bullet point of the second part of this lemma, and is omitted for brevity.

- Third, we show that $|z_a(w_a(t))| \leqslant C_z$ for each $t \geqslant 0$.

  The proof here is nearly identical to the proof that $|z_a(W_a[n])| \leqslant C_z$ in the fourth bullet point of the second part of this lemma, and is omitted for brevity.

- Fourth, we show that there exists some $C_\xi > 0$ such that $\xi_a(t) \geqslant C_\xi$ for each $a \in A$, $t \geqslant 0$.

  Define:

$$\begin{aligned} C_\xi &:= \min\left\{\min\{\xi_{a'}(0) : a' \in A\}, \frac{1}{|A|}e^{-2\beta C_z}\right\} \\ &\in (0,1). \end{aligned}$$

  By definition of C_ξ, we have $\xi_a(0) \geqslant C_\xi$. Moreover, for each $n \geqslant 0$, we have:

$$\begin{aligned} &\frac{\exp(-\beta\Big[z_a(W[n])\Big])}{\sum_{a' \in A_{i_a}^+} \exp(-\beta\Big[z_{a'}(W[n])\Big])} \\ \geqslant &\frac{e^{-\beta C_z}}{|A_{i_a}^+| \cdot e^{\beta C_z}} \\ \geqslant &\frac{1}{|A|}e^{-2\beta C_z} \\ \geqslant &C_\xi. \end{aligned}$$

  Thus, by Lemma G.3.1, we have $\xi_a(t) \geqslant C_\xi$ for each $t \geqslant 0$.

- Fifth, we show that there exists $C_w > 0$ such that, for each $a \in A$, $t \geqslant 0$, we have $w_a(t) \geqslant C_w$.

  The proof here is nearly identical to the proof that $W_a[n] \geqslant C_w$ in the fourth bullet point of the second part of this lemma, and is omitted for brevity.

□

Remark G.3.1. *Crucially, the constants introduced and used in the above proof, i.e.,*

$$C_w, C_m, C_\xi$$

(and naturally, g_o), do not depend on the node-dependent update rates K_i. This is a critical observation, since each K_i must be chosen to be large enough such that the term:

$$1 - \frac{\sum_{a' \in A_{i_a}^-} h_{a'}(w(t))}{K_{i_a} \cdot \sum_{\hat{a} \in A_{i_a}^+} w_{\hat{a}}(t)}$$

which appears in (G.12)*, is always strictly positive, i.e., that:*

$$K_{i_a} > \frac{\sum_{\hat{a} \in A^+_{i_a}} w_{\hat{a}}(t)}{\sum_{a' \in A^-_{i_a}} h_{a'}(w(t))} \tag{G.18}$$

for all $t \geqslant 0$*, regardless of the initial value of* $w(0) \in \mathcal{W}$*. The numerator in the right-hand-side expression of* (G.18) *can be straightforwardly (if loosely) upper bounded by* $|A| g_o$*. However, the denominator in the right-hand-side expression of* (G.18) *must be lower-bounded recursively in increasing order of depth, which requires* K_{i_a} *to depend on* $\{K_i : i \in I, \ell_i < \ell_{i_a}\}$*, as well as on the constants* C_w, C_m, C_ξ*, and* g_o*. Thus, the fact that* C_w, C_m, C_ξ*, and* g_o *are established independently of the values of* K_i *allows circular reasoning to be avoided.*

The lemma below establishes the final part of Lemma 8.3.2. Below, we restrict the domains of the maps $\bar{\xi}^\beta$ and z_a to reflect the bounds of the traffic flow trajectory w established in the above lemma, i.e., $\bar{\xi}^\beta, z_a : \mathcal{W}' \to \mathbb{R}$, with the flow restricted to:

$$\mathcal{W}' := \mathcal{W} \cap [C_w, g_o]^{|A|}$$

and the toll restricted to $[0, C_p]^{|A_O|}$.

Lemma G.3.3. *The continuous-time dynamics* (G.6) *satisfies:*

1. *For each* $a \in A$*, the restriction of the latency-to-go map* $z_a : \mathcal{W} \to \mathbb{R}^{|A_O|} \to \mathbb{R}$ *to* $\mathcal{W}'$ *is Lipschitz continuous.*

2. *The map from the probability transitions* $\xi \in \prod_{i \in I \setminus \{d\}} \Delta(A^+_i)$ *and the traffic flows* $w \in \mathcal{W}$ *is Lipschitz continuous.*

3. *For each* $a \in A$*, the restriction of the continuous dynamics transition map* $\rho_a : \mathbb{R}^{|A|} \times \mathbb{R}^{|A_O|} \to \mathbb{R}^{|A|}$*, defined recursively by:*

$$\rho_a(\xi) := K_{i_a} \left(-\xi_a + \frac{\exp(-\beta z_a(w))}{\sum_{a' \in A^+_{i_a}} \exp(-\beta z_{a'}(w))} \right)$$

   to $\mathcal{W}'$ *is Lipschitz continuous.*

Proof.

1. We shall establish the Lipschitz continuity of z_a, for each $a \in A$, by providing uniform bounds on its partial derivatives across all values of its arguments $w \in \mathcal{W}'$.

   The proof follows by induction on the height index $k \in [m(G)]$. For each $a \in A$, let $\tilde{z}_a : \mathbb{R}^{|A|} \to \mathbb{R}$ be the continuous extension of $z_a : \mathcal{W} \to \mathbb{R}$ to the Euclidean space $\mathbb{R}^{|A|}$ containing $\mathcal{W}$. By definition of Lipschitz continuity, if $\tilde{z}_a$ is Lipschitz for some $a \in A$, then so is z_a. For each $a \in A^-_d = \{a \in A : m_a = 1\}$ and any $w \in \mathbb{R}^{|A|}$:

$$\tilde{z}_a(w) = \ell_{[a]}(w_{[a]}).$$

Thus, for any $\hat{a} \in A$, and any $w \in \mathbb{R}^{|A|}$, $p \in \mathbb{R}^{|A_O|}$:

$$\frac{\partial \tilde{z}_a}{\partial w_{\hat{a}}}(w) = \frac{d\ell_{[a]}}{dw}(w_{[a]}) \cdot \mathbf{1}\{\hat{a} \in [a]\} \in [0, C_{ds}].$$

We set $C_{z,1} := C_{ds}$.

Now, suppose that there exists some depth $k \in [m(G) - 1]$ and some constant $C_{z,k} > 0$ such that, for any $a \in A$ satisfying $m_a \leqslant k$, and any $w \in \mathcal{W}$, $n \geqslant 0$, the map $\tilde{z}_a : \mathbb{R}^{|A|} \to \mathbb{R}$ is continuously differentiable, with:

$$\left| \frac{\partial \tilde{z}_a}{\partial w_{\hat{a}}}(w) \right| \leqslant C_{z,k}.$$

Continuing with the induction step, fix $a \in A$ such that $m_a = k + 1$ (there exists at least one such link, by Proposition G.1.1, Part 4). From Proposition G.1.1, Part 2, we have $m_{a'} \leqslant k$ for each $a' \in A^+_{i_a}$. Thus, the induction hypothesis implies that, for any $\hat{a} \in A$:

$$\tilde{z}_a(w) = \ell_{[a]}(w_{[a]}) - \frac{1}{\beta} \sum_{a' \in A^+_{j_a}} e^{-\beta z_{a'}(w)}.$$

Computing partial derivatives with respect to each component of w, we obtain:

$$\begin{aligned} \frac{\partial \tilde{z}_a}{\partial w_{\hat{a}}}(w) &= \frac{d\ell_{[a]}}{dw}(w_{[a]}) \cdot \mathbf{1}\{\hat{a} \in [a]\} \\ &\quad + \sum_{a' \in A^+_{ja}} e^{-\beta \tilde{z}_{a'}(w)} \cdot \frac{\partial \tilde{z}_{a'}}{\partial w_{\hat{a}}}(w), \\ \implies \left| \frac{\partial \tilde{z}_a}{\partial w_{\hat{a}}}(w) \right| &\leqslant C_{ds} + |A| \cdot C_{z,k}. \end{aligned}$$

We can complete the induction step by taking $C_{z,k+1} := C_{ds} + |A| \cdot C_{z,k}$.

This establishes that, for each $a \in A$, the map z_a is continuously differentiable, with partial derivatives uniformly bounded by a uniform constant, $C_z := C_{z,m(G)}$. This establishes the Lipschitz continuity of the map z_a for each $a \in A$, and thus proves this part of the proposition.

2. Recall that the map from traffic distributions probabilities (ξ) to traffic flows (w) is given as follows, for each $a \in A$. Recall that $\mathbf{R}(a)$ denotes the set of all routes in the Condensed DAG that contain the arc a:

$$w_a = \left(g_{i_a} + \sum_{\hat{a} \in A^-_i} w_a \right) \cdot \xi_a = g_o \cdot \sum_{r \in \mathbf{R}(a)} \prod_{k=1}^{|r|} \xi_{a_{r,k}},$$

where $a_{r,k}$ denotes the k-th arc along a given route $r \in \mathbf{R}$, for each $k \in |r|$. Thus, the map from ξ to w is continuously differentiable. Moreover, the domain of this map is compact; indeed, for each $a \in A$, we have $\xi_a \in [0, 1]$, and for each non-destination node $i \neq d$, we have $\sum_{a \in A_i^+} \xi_a = 1$. Therefore, the map $\xi \mapsto w$ has continuously differentiable derivatives with magnitude bounded above by some constant uniform in the compact set of realizable probability distributions ξ. This is equivalent to stating that the map $\xi \mapsto w$ is Lipschitz continuous.

3. Above, we have established that the maps z_a and $\xi \mapsto w$ are Lipschitz continuous. Since the addition and composition of Lipschitz maps is Lipschitz, it suffices to verify that the map $\hat{\rho} : \mathbb{R}^{|A|} \to \mathbb{R}^{|A|}$, defined element-wise by:

$$\hat{\rho}_a(z) := \frac{e^{-\beta z_a}}{\sum_{a' \in A_{ia}^+} e^{-\beta z_{a'}}}, \qquad \forall\, a \in A$$

is Lipschitz continuous. We do so below by computing, and establishing a uniform bound for, its partial derivatives. For each $\hat{a} \in A$:

$$\begin{aligned}
&\frac{\partial \hat{\rho}_a}{\partial z_{\bar{a}}} \\
&= \frac{\left(\sum_{a' \in A_{ia}^+} e^{-\beta z_{a'}} \cdot (-\beta) e^{-\beta z_a} \cdot \frac{\partial z_a}{\partial z_{\bar{a}}} - e^{-\beta z_a} \cdot \sum_{a' \in A_{ia}^+} (-\beta) e^{-\beta z_{a'}} \frac{\partial z_{a'}}{\partial z_{\bar{a}}} \right)}{\left(\sum_{a' \in A_{ia}^+} e^{-\beta z_{a'}} \right)^2} \\
&= -\frac{e^{-\beta z_a}}{\sum_{a' \in A_{ia}^+} e^{-\beta z_{a'}}} \cdot \beta \cdot \frac{\partial z_a}{\partial z_{\hat{a}}} + \frac{\beta e^{-\beta z_a}}{\left(\sum_{a' \in A_{ia}^+} e^{-\beta z_{a'}} \right)^2} \cdot \sum_{a' \in A_{ia}^+} e^{-\beta z_{a'}} \frac{\partial z_{a'}}{\partial z_{\bar{a}}}.
\end{aligned}$$

Observe that:

$$\begin{aligned}
\sum_{a' \in A_{ia}^+} e^{-\beta z_{a'}} \frac{\partial z_{a'}}{\partial z_{\bar{a}}} &= \sum_{a' \in A_{ia}^+} e^{-\beta z_{a'}} \cdot \mathbf{1}\{a' = \hat{a}\} \\
&\leqslant \max_{a' \in A_{ia}^+} e^{-\beta z_{a'}}.
\end{aligned}$$

This, together with triangle inequality, then gives:

$$\left| \frac{\partial \hat{\rho}_a}{\partial z_{\bar{a}}} \right| = \beta + \beta = 2\beta.$$

This concludes the proof for this part of the proposition.

□

We present the proof of Theorem 8.3.1, restated as follows: For any $\delta > 0$:

$$\limsup_{n\to\infty} \mathbb{E}\Big[\|\xi[n] - \bar{\xi}^{\beta}\|_2^2\Big] \leqslant O(\mu),$$
$$\limsup_{n\to\infty} \mathbb{P}\Big(\|\xi[n] - \bar{\xi}^{\beta}\|_2^2 \geqslant \delta\Big) \leqslant O\left(\frac{\mu}{\delta}\right).$$

Proof of Theorem 8.3.1 Here, we conclude the proof of Theorem 8.3.1.

Proof. (**Proof of Theorem 8.3.1**) Lemma G.3.2 asserts that $M[n]$ is bounded (uniformly in $n \geqslant 0$), while Lemma G.3.3 establishes that $\rho : \mathbb{R}^{|A|} \to \mathbb{R}$ is Lipschitz continuous. The proof of Theorem 8.3.1 now follows by applying the stochastic approximation results in Borkar [57], Chapters 2 and 9. □

Appendix H

Appendix for Chapter 9

H.1 Calibration of Latency Functions

Here, we present the methodology used to compute the latency function of all freeways in Figure 9.3. Recall from Section 9.3, we need to compute the average travel time and average flow on every edge for every day. To achieve this goal, we utilize morning rush hour data from the PeMS dataset, spanning from January 2019 to June 2019. Let's denote the set of all weekdays in this time-frame by $\mathcal{T}$. For every edge $e \in E$ and day $t \in \mathcal{T}$, let's denote the average travel time by $\hat{\ell}_e^t$ and the average edge flow by $\hat{w}_e^t$. In order to estimate these quantities, we use PeMS data during the morning rush hours $\mathcal{H} = [6\text{am} - 7\text{am}, 7\text{am} - 8\text{am}, 8\text{am} - 9\text{am}, 9\text{am} - 10\text{am}, 10\text{am} - 11\text{am}, 11\text{am} - 12\text{noon}]$. Let S_e be the number of sensors fitted on edge $e \in E$ which provide average hourly flow and average speed information.

First, we demonstrate how to use the raw data from sensors to compute the average travel time on every edge. We compute an estimate of the time required to travel the edge e at hour h by accumulating the average time required to travel between sensors on that link as follows:

$$\hat{\ell}_e^{ht} = \sum_{s=1}^{S_e-1} \frac{L_e^s}{v_e^{sht}}, \quad \forall\, e \in E, h \in \mathcal{H}, t \in \mathcal{T}, \tag{H.1}$$

where L_e^s is the distance between sensor s and $s+1$ on edge $e \in E$ and v_e^{sht} is the average speed of traffic passing over the sensor s on edge e during hour h on day t. Next, we compute the average hourly flow on an edge as follows:

$$\hat{w}_e^{ht} = \frac{\sum_{s=1}^{S_e-1} L_e^s \tilde{w}_e^{sht}}{\sum_{s=1}^{S_e-1} L_e^s}, \quad \forall\, e \in E, h \in \mathcal{H}, t \in \mathcal{T},$$

where $\tilde{w}_e^{sht}$ is the hourly average flow of traffic passing over sensor s on edge e during hour h on day t. We use the hourly average edge flows $\hat{w}_e^{ht}$ and the hourly average travel times

$\hat{\ell}_e^{ht}$ to compute the average travel time on any edge $e \in E$ as follows:

$$\hat{\ell}_e^t = \frac{\sum_{h\in\mathcal{H}} \hat{w}_e^{ht}\hat{\ell}_e^{ht}}{\sum_{h\in\mathcal{H}} \hat{w}_e^{ht}}, \quad e \in E, t \in \mathcal{T}.$$

Similarly, we compute the average of the hourly flows as follows:

$$\hat{w}_e^t = \frac{1}{|\mathcal{H}|}\sum_{h\in\mathcal{H}} \hat{w}_e^{ht}, \quad t \in \mathcal{T}, e \in E.$$

H.2 Calibration of User Demand

We outline our method for calculating the daily demand of travelers moving between various origin-destination pairs from January 2019 to June 2019. Our approach involves three main steps:

Step 1: Estimating relative demand between nodes using the Safegraph dataset: We leverage the Safegraph dataset to obtain the relative demand of travelers traveling between different nodes in the Bay Area. Specifically, the Neighborhood Patterns dataset from Safegraph provides the average daily count of mobile devices moving between different census block groups (CBGs) on workdays for each month. This is then aggregated over the set of nodes after adjusting for sampling bias.

More formally, let's denote the set of CBGs in the Bay Area by $\mathcal{C}$. The SafeGraph dataset provides the average daily count of travelers $N^{cc'}$ traveling from CBG c to c'. However, the SafeGraph dataset exhibits sampling bias[1] because different CBGs are sampled at different rates. We correct for sampling bias in this data by modifying the counts $N^{cc'}$ using the population data provided by the ACS. That is, we compute the corrected count of travelers traveling from CBG c to c' as follows

$$\tilde{N}^{cc'} = N^{cc'}\frac{R^c}{\sum_{c\in\mathcal{C}} R^c} \cdot \frac{\sum_{c\in\mathcal{C}}\sum_{c'\in\mathcal{C}} N^{cc'}}{\sum_{c'\in\mathcal{C}} N^{cc'}},$$

where R^c is the number of residents in CBG c as reported by the ACS dataset.

Step 2: Calibrating type-specific demands with ACS dataset. Given the the adjusted count of travelers we compute the demand of travelers from o-d pair $k \in K$ by aggregating the demand over set of nodes as follows

$$\tilde{D}^k = \sum_{c\in k_o}\sum_{c'\in k_d} \tilde{N}_{\mathcal{C}}^{cc'},$$

[1] as referred in `https://colab.research.google.com/drive/1u15afRytJMsizySFqA2EPlXSh3KTmNTQ`

where $k_o, k_d \in N$ are the origin and destination nodes of the o-d pair $k \in K$. To obtain the demand in terms of units of flow we compute

$$D^{ik} = \frac{\tilde{D}^k}{\sum_{k' \in K} \tilde{D}^{k'} \mathbb{1}(k'_o = k_o)} \frac{A^{ik_o}}{|\mathcal{H}|},$$

where A^{ik_o} is the total driving population of type i at node k_o as given by the ACS dataset and $|\mathcal{H}|$ is the number of hours in morning rush hours (6 am to 12 noon).

Step 3: Incorporating daily variability with the PeMS dataset. We convert the monthly demand estimates obtained in Step 2 into daily demand data by scaling it proportional to the total daily flow from PeMS dataset. More formally, we compute the average total edge load over all workdays from January 2019 to June 2019 as follows

$$\overline{w} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \sum_{e \in E} \hat{w}_e^t, \tag{H.2}$$

where $\hat{w}_e^t$ is the average edge load on day t on edge e, which is obtained in Appendix B using PeMS data. Next, to obtain the daily demand, we scale the monthly demand obtain in Step 2 as follows:

$$D_t^{ik} = \frac{\sum_{e \in E} \hat{w}_e^t}{\overline{w}} \cdot D^{ik}, \quad \forall\, t \in \mathcal{T}, i \in I, k \in K. \tag{H.3}$$

H.3 Computing Pricing Schemes Lying on the Pareto Curve in Figure 9.11

Here, we provide a method to compute the Pareto efficient congestion pricing schemes that trade-off between minimizing average travel time and optimizing the equity objective (as in (9.8)).

Before presenting our method to compute Pareto front, we recap the methodology delineated in Section 9.2. On a high level, the procedure in Section 9.2 comprises of two steps:

- **Step 1:** Characterize the set of tolls that will implement the best possible average travel time $S(w^\dagger)$
- **Step 2:** On the set of tolls characterized in Step 1, compute the tolls that optimize the joint equity-welfare objective.

However, the above methodology does not provide a way to compute pricing schemes that trade-off some amount of average travel time in order to improve on equity. Particularly, for

any $S^* < S(w^\dagger)$ there is no direct method to characterize the set of pricing schemes that will implement an edge flow vector that would result in average travel time of S^* (as in Step 1 above). We provide a new procedure that builds up on the tools presented in Section 9.2 to *estimate* this set of tolls which can be use to the tolls that optimize the equity-welfare objective as in Step 2.

Our approach is stated below:

- Sample N vectors $\{\gamma^i\}_{i\in[N]} \subset \mathbb{R}^{|E|}$ such that for every $i \in [N]$, $\gamma^i \sim \mathsf{Unif}([0,1]^{|E|})$.
- For each $i \in [N]$, solve the following weighted average time minimization problem

$$\min_{w\in W} \sum_{e\in E} \gamma_e^i w_e \ell_e(w_e),$$

  where W is the set of all feasible edge flows given demand D as highlighted below

$$W = \{w \in \mathbb{R}^{|E|} : \exists\ q \in \mathcal{Q}(D) \text{ s.t. } w_e = \sum_{i\in I} f_e^i(q)\}.$$

  Let's denote $w^{\dagger\gamma^i}$ to be the optimal value of the above optimization problem.
- For each $i \in [N]$, use $w^{\dagger\gamma^i}$ in place of $w^\dagger$ in the optimization process to compute hom and het pricing schemes (as highlighted in Section 9.2).
- Compute the average travel time and equity objective corresponding to each of $w^{\dagger\gamma^i}$ and compute the Pareto efficient solutions amongst these N solutions.

In Figure 9.11), the blue triangles are Pareto efficient solutions obtained by taking $N = 100$ in the above procedure.

Remark H.3.1. *Note that this procedure only provides an estimate of Pareto front and not the exact Pareto front. This is because by definition $S^*_{\gamma^i} := S(w^{\dagger\gamma^i}) \leqslant S(w^\dagger)$. Following similar analysis as in Proposition 9.2.2, we can compute the set of pricing schemes that will implement $w^{\dagger\gamma^i}$ on the transportation network. Unlike $S(w^\dagger)$, the set of edge flow vectors that result in the average travel time of $S^*_{\gamma^i}$ is not unique, we cannot characterize the entire set of tolls that could result in average travel time $S^*_{\gamma^i}$. Thus, our procedure relies on taking large values of N so that we can get better estimate of this set. s Extending our approach to derive better estimates of Pareto front is an interesting direction of future research that is bound to help planner in making important design decisions about congestion pricing.*

H.4 Proofs for Section 9.2

Proof of Proposition 1.

(1) To establish this result, we first show that for any given set of tolls p, the optimization problem (9.6) is a convex optimization problem. Next, using KKT conditions for optimality we show that the optimal solution to (9.6) satisfy the requirements of Nash equilibrium posited in Definition 9.1.1.

To show that the (9.6) is a convex optimization problem, we note that the constraint set is convex as it is a product simplex which is a convex set. Next, we show that the objective function is convex. Since the objective is differentiable, it is sufficient to show that

$$\sum_{i\in I}\sum_{k\in K}\sum_{r\in R^k}\left(\frac{\partial\Phi(q,p)}{\partial q_r^{ik}}-\frac{\partial\Phi(\tilde{q},p)}{\partial q_r^{ik}}\right)\left(q_r^{ik}-\tilde{q}_r^{ik}\right)\geqslant 0\quad\forall\, q,\tilde{q}\in\mathcal{Q}, q\neq\tilde{q}. \tag{H.4}$$

To see this, we note that

$$\begin{aligned}\frac{\partial\Phi(q,p)}{\partial q_r^{ik}}&=\sum_{e\in E}\ell_e(w_e(q))\frac{\partial w_e(q)}{\partial q_r^{ik}}+\sum_{i\in I}\sum_{e\in E}\frac{(p_e^i+g_e)}{\theta^i}\frac{\partial w_e^i(q)}{\partial q_r^{ik}}\\&=\sum_{e\in E}\ell_e(w_e(q))\mathbb{1}(e\in r)+\sum_{e\in E}\frac{(p_e^i+g_e)}{\theta^i}\mathbb{1}(e\in r)=c_r^i(q,p).\end{aligned}$$

Consequently, for any $q,\tilde{q}\in\mathcal{Q}$ such that $q\neq\tilde{q}$ it holds that

$$\begin{aligned}&\sum_{i\in I}\sum_{k\in K}\sum_{r\in R^k}\left(\frac{\partial\Phi(q,p)}{\partial q_r^{ik}}-\frac{\partial\Phi(\tilde{q},p)}{\partial q_r^{ik}}\right)\left(q_r^{ik}-\tilde{q}_r^{ik}\right)\\&=\sum_{i\in I}\sum_{k\in K}\sum_{r\in R^k}\left(\sum_{e\in E}(\ell_e(w_e(q))-\ell_e(w_e(\tilde{q})))\mathbb{1}(e\in r)\right)\left(q_r^{ik}-\tilde{q}_r^{ik}\right),\\&=\sum_{e\in E}(\ell_e(w_e(q))-\ell_e(w_e(\tilde{q})))\sum_{i\in I}\sum_{k\in K}\sum_{r\in R^k}\mathbb{1}(e\in r)\left(q_r^{ik}-\tilde{q}_r^{ik}\right),\\&=\sum_{e\in E}(\ell_e(w_e(q))-\ell_e(w_e(\tilde{q})))\,(w_e(q)-w_e(\tilde{q}))\geqslant 0.\end{aligned}$$

where the last inequality follows because ℓ_e is strictly increasing function. Thus, we have established the (9.6) is convex optimization problem.

Next, we analyze the KKT conditions associated with (9.6). Define the Lagrangian

$$\mathcal{L}(q,\lambda,\mu;p)=\Phi(q,p)+\sum_{i\in I}\sum_{k\in K}\lambda^{ik}(D^{ik}-\sum_{r\in R^k}q_r^{ik})-\sum_{i\in I}\sum_{k\in K}\sum_{r\in R^k}\mu_r^{ik}q_r^{ik}.$$

Since (9.6) is a convex optimization problem and the strong form of Slater's conditions hold as the feasible set is a product-simplex, we obtain the following first-order

necessary and sufficient condition of optimality:

$$\frac{\partial \mathcal{L}(q^*(p), \lambda^*, \mu^*; p)}{\partial q_r^{ik}} = 0 \quad \forall\, i \in I, k \in K, r \in R^k, \tag{C1}$$

$$\sum_{r \in R^k} q_r^{*ik}(p) = D^{ik} \quad \forall\, i \in I, k \in K, \tag{C2}$$

$$\mu_r^{*ik} q_r^{*ik}(p) = 0 \quad \forall\, i \in I, k \in K, r \in R^k, \tag{C3}$$

$$\mu_r^{*ik} \geqslant 0, q_r^{*ik}(p) \geqslant 0 \quad \forall\, i \in I, k \in K, r \in R^k. \tag{C4}$$

Note that (C1) can be equivalently written as

$$0 = \frac{\partial \mathcal{L}(q^*(p), \lambda^*, \mu^*; p)}{\partial q_r^{ik}} = \frac{\partial \Phi(q^*(p), p)}{\partial q_r^{ik}} - \lambda^{ik} - \mu_r^{ik} = c_r^i(q^*(p), p) - \lambda^{ik} - \mu_r^{ik}$$

Additionally, using (C4) we obtain that $c_r^i(q^*(p), p) \geqslant \lambda^{ik}$, for every $i \in I, k \in K, r \in R^k$. Furthermore, from (C3) we obtain that if for some $i \in I, k \in K, r \in R^k$, $q_r^{*ik} > 0$ then $c_r^i(q^*(p), p) = \lambda^{ik}$. This is precisely the conditions stated in Definition 9.1.1.

(2) Using the first-order necessary conditions for constrained optimality, we observe that,

$$\sum_{i \in I} \sum_{k \in K} \sum_{r \in R^k} \frac{\partial S(q^\dagger)}{\partial q_r^{ik}} \left(\tilde{q}_r^{ik} - q_r^{\dagger, ik} \right) \geqslant 0 \quad \forall\, \tilde{q} \in \mathcal{Q}. \tag{H.5}$$

Similarly, it holds that

$$\sum_{i \in I} \sum_{k \in K} \sum_{r \in R^k} \frac{\partial S(\bar{q}^\dagger)}{\partial q_r^{ik}} \left(\tilde{q}_r^{ik} - \bar{q}_r^{\dagger, ik} \right) \geqslant 0 \quad \forall\, \tilde{q} \in \mathcal{Q}. \tag{H.6}$$

Selecting $\tilde{q} = \bar{q}^\dagger$ in (H.5), and selecting $\tilde{q} = q^\dagger$ in (H.6) and substracting the resulting inequality we obtain

$$\sum_{i \in I} \sum_{k \in K} \sum_{r \in R^k} \left(\frac{\partial S(q^\dagger)}{\partial q_r^{ik}} - \frac{\partial S(\bar{q}^\dagger)}{\partial q_r^{ik}} \right) \left(q_r^{\dagger, ik} - \bar{q}_r^{\dagger, ik} \right) \leqslant 0. \tag{H.7}$$

Suppose there exists $q^\dagger, \bar{q}^\dagger \in Q^\dagger$ such that there exists $e \in E$ such that $w_e(q^\dagger) \neq w_e(\bar{q}^\dagger)$. Then we will show that (H.7) is violated.

Note that for any $q \in \mathcal{Q}$,

$$\begin{aligned} \frac{\partial S(q)}{\partial q_r^{ik}} &= \sum_{e \in E} \frac{\partial w_e(q)}{\partial q_r^{ik}} \ell_e(w_e(q)) + \sum_{e \in E} w_e(q) \nabla \ell_e(w_e(q)) \frac{\partial w_e(q)}{\partial q_r^{ik}} \\ &= \sum_{e \in E} \mathbb{1}(e \in r) \ell_e(w_e(q)) + \sum_{e \in E} w_e(q) \nabla \ell_e(w_e(q)) \mathbb{1}(e \in r). \end{aligned}$$

Using this, we compute the left-hand side of (H.7),

$$
\begin{aligned}
&\sum_{i\in I}\sum_{k\in K}\sum_{r\in R^k}\left(\frac{\partial S(q^\dagger)}{\partial q_r^{ik}}-\frac{\partial S(\bar{q}^\dagger)}{\partial q_r^{ik}}\right)\left(q_r^{\dagger,ik}-\bar{q}_r^{\dagger,ik}\right)\\
=&\sum_{i\in I}\sum_{k\in K}\sum_{r\in R^k}\sum_{e\in E}\mathbb{1}(e\in r)(\ell_e(w_e(q^\dagger))-\ell_e(w_e(\bar{q}^\dagger)))\left(q_r^{\dagger,ik}-\bar{q}_r^{\dagger,ik}\right)\\
&+\sum_{i\in I}\sum_{k\in K}\sum_{r\in R^k}\sum_{e\in E}(w_e(q^\dagger)\nabla\ell_e(w_e(q^\dagger))-w_e(\bar{q}^\dagger)\nabla\ell_e(w_e(\bar{q}^\dagger)))\mathbb{1}(e\in r)\left(q_r^{\dagger,ik}-\bar{q}_r^{\dagger,ik}\right)\\
=&\sum_{e\in E}(\ell_e(w_e(q^\dagger))-\ell_e(w_e(\bar{q}^\dagger)))\sum_{i\in I}\sum_{k\in K}\sum_{r\in R^k}\mathbb{1}(e\in r)\left(q_r^{\dagger,ik}-\bar{q}_r^{\dagger,ik}\right)\\
&+\sum_{e\in E}(w_e(q^\dagger)\nabla\ell_e(w_e(q^\dagger))-w_e(\bar{q}^\dagger)\nabla\ell_e(w_e(\bar{q}^\dagger)))\sum_{i\in I}\sum_{k\in K}\sum_{r\in R^k}\mathbb{1}(e\in r)\left(q_r^{\dagger,ik}-\bar{q}_r^{\dagger,ik}\right)\\
=&\sum_{e\in E}(\ell_e(w_e(q^\dagger))-\ell_e(w_e(\bar{q}^\dagger)))\left(w_e(q^\dagger)-w_e(\bar{q}^\dagger)\right)\\
&\quad+\sum_{e\in E}(w_e(q^\dagger)\nabla\ell_e(w_e(q^\dagger))-w_e(\bar{q}^\dagger)\nabla\ell_e(w_e(\bar{q}^\dagger)))\left(w_e(q^\dagger)-w_e(\bar{q}^\dagger)\right)
\end{aligned}
$$

Note that $\sum_{e\in E}(\ell_e(w_e(q^\dagger))-\ell_e(w_e(\bar{q}^\dagger)))\left(w_e(q^\dagger)-w_e(\bar{q}^\dagger)\right)\geqslant 0$, due to the monotonicity of latency function. Moreover, note that

$$
\begin{aligned}
&\sum_{e\in E}(w_e(q^\dagger)\nabla\ell_e(w_e(q^\dagger))-w_e(\bar{q}^\dagger)\nabla\ell_e(w_e(\bar{q}^\dagger)))\left(w_e(q^\dagger)-w_e(\bar{q}^\dagger)\right)\\
=&\sum_{e\in E}(w_e(q^\dagger)\nabla\ell_e(w_e(q^\dagger))-w_e(\bar{q}^\dagger)\nabla\ell_e(w_e(q^\dagger))\\
&\quad+w_e(\bar{q}^\dagger)\nabla\ell_e(w_e(q^\dagger))-w_e(\bar{q}^\dagger)\nabla\ell_e(w_e(\bar{q}^\dagger)))\cdot\left(w_e(q^\dagger)-w_e(\bar{q}^\dagger)\right)\\
=&\sum_{e\in E}\nabla\ell_e(w_e(q^\dagger))(w_e(q^\dagger)-w_e(\bar{q}^\dagger))^2+\sum_{e\in E}w_e(\bar{q}^\dagger)(\nabla\ell_e(w_e(q^\dagger))\\
&\quad-\nabla\ell_e(w_e(\bar{q}^\dagger)))\left(w_e(q^\dagger)-w_e(\bar{q}^\dagger)\right).
\end{aligned}
$$

Note that $\sum_{e\in E}\nabla\ell_e(w_e(q^\dagger))(w_e(q^\dagger)-w_e(\bar{q}^\dagger))^2>0$ due to the hypothesis that there exists at least one edge where $w_e(q^\dagger)\neq w_e(\bar{q}^\dagger)$ and the fact that the latency function is strictly increasing. Moreover,

$$
\sum_{e\in E}w_e(\bar{q}^\dagger)(\nabla\ell_e(w_e(q^\dagger))-\nabla\ell_e(w_e(\bar{q}^\dagger)))\left(w_e(q^\dagger)-w_e(\bar{q}^\dagger)\right)\geqslant 0
$$

as $\ell_e(\cdot)$ is assumed to be convex. Thus, we obtain

$$
\sum_{i\in I}\sum_{k\in K}\sum_{r\in R^k}\left(\frac{\partial S(q^\dagger)}{\partial q_r^{ik}}-\frac{\partial S(\bar{q}^\dagger)}{\partial q_r^{ik}}\right)\left(q_r^{\dagger,ik}-\bar{q}_r^{\dagger,ik}\right)>0,
$$

which contradicts (H.7).

□

Proof of Proposition 9.2.2.

(1) First, we prove that given any optimal solution $(p^\dagger, z^\dagger)$ of $(\mathcal{P}_{\mathsf{hom}})$, $p^\dagger$ induces the socially optimal edge flow vector $w^\dagger$. Consider any optimal solution of $(\mathcal{D}_{\mathsf{hom}})$, denoted as $q^\dagger$. From strong duality theory, we know that $(p^\dagger, z^\dagger, q^\dagger)$ must satisfy complementary slackness conditions associated with the constraints in $(\mathcal{P}_{\mathsf{hom}})$ and $(\mathcal{D}_{\mathsf{hom}})$. In particular, the complementary slackness condition for $(\mathcal{P}_{\mathsf{hom}})$-$(\mathcal{D}_{\mathsf{hom}})$ indicates that for any $i \in I$, $k \in K$, and $r \in R^k$,

$$q_r^{\dagger ik} > 0, \quad \Rightarrow \quad z^{\dagger ik} = \theta^i \ell_r(w^\dagger) + \sum_{e \in r}(p_e^\dagger + g_e) = \theta^i c_r^i(q^\dagger, p^\dagger).$$

Additionally, from $(\mathcal{P}_{\mathsf{hom}})$, we have for all $i \in I, k \in K, r' \in R^k$,

$$z^{\dagger ik} \leqslant \theta^i \ell_{r'}(w^\dagger) + \sum_{e \in r'}(p_e^\dagger + g_e) = \theta^i c_{r'}^i(q^\dagger, p^\dagger).$$

Consequently,

$$\forall i \in I, k \in K, r \in R^k, \quad q_r^{\dagger ik} > 0, \quad \Rightarrow \quad c_r^i(q^\dagger, p^\dagger) \leqslant c_{r'}^i(q^\dagger, p^\dagger), \quad \forall r' \in R_k. \tag{H.8}$$

That is, the flow vector $q^\dagger$ only takes routes with the minimum cost given the socially optimal edge flow vector $w^\dagger$. We next prove that $w^\dagger$ is indeed induced by $q^\dagger$, i.e. constraint $(\mathcal{D}_{\mathsf{hom}.a})$ is tight with the optimal solution.

For notational brevity, we denote $\hat{w}_e = \sum_{i \in I}\sum_{k \in K}\sum_{r \in \{R^k | r \ni e\}} q_r^{\dagger ik}$ as the edge flow induced by $q^\dagger$. Suppose for the sake of contradiction that for some non-empty subset of edges $E^\dagger \subseteq E$,

$$\begin{aligned}
&\forall\ e \in E^\dagger, \quad \hat{w}_e = \sum_{i \in I}\sum_{k \in K}\sum_{r \in \{R^k | r \ni e\}} q_r^{\dagger ik} < w_e^\dagger, \\
&\forall\ e \in E \backslash E^\dagger, \quad \hat{w}_e = \sum_{i \in I}\sum_{k \in K}\sum_{r \in \{R^k | r \ni e\}} q_r^{\dagger ik} = w_e^\dagger.
\end{aligned}$$

Then,

$$\begin{aligned}
\sum_{e \in E} \hat{w}_e \ell_e(\hat{w}_e) &= \sum_{e \in E^\dagger} \hat{w}_e \ell_e(\hat{w}_e) + \sum_{e \in E \backslash E^\dagger} \hat{w}_e \ell_e(\hat{w}_e) \\
&< \sum_{e \in E^\dagger} w_e^\dagger \ell_e(w_e^\dagger) + \sum_{e \in E \backslash E^\dagger} w_e^\dagger \ell_e(w_e^\dagger) = \sum_{e \in E} w_e^\dagger \ell_e(w_e^\dagger),
\end{aligned}$$

where the inequality is due the the fact that ℓ_e is a strictly increasing function. This contradicts with the fact that $w^\dagger$ minimizes the social cost function. Therefore, we must have $\hat{w}_e = \sum_{i \in I}\sum_{k \in K}\sum_{r \in \{R^k | r \ni e\}} f_r^{\dagger ik} = w_e^\dagger$, for every $e \in E$.

Following from the fact that $q^\dagger$ satisfies (H.8) and induces the socially optimal edge flow vector $w^\dagger$, we can conclude that $w^\dagger$ is an equilibrium edge flow vector induced by the flow

vector $q^\dagger$ associated under the toll price $p^\dagger$. Hence, the optimal solution $p^\dagger$ of $(\mathcal{P}_{\mathsf{hom}})$ indeed implements the socially optimal edge flow.

We now prove the other direction. Suppose that there exists a hom toll vector $\tilde{p}$ that induces the socially optimal edge flow $w^\dagger$ in equilibrium, then there exists $\tilde{z}$ such that $(\tilde{z}, \tilde{p})$ is an optimal solution to $(\mathcal{P}_{\mathsf{hom}})$. We denote $\tilde{q}$ as a Nash equilibrium strategy distribution given toll $\tilde{p}$. Then, such $\tilde{q}$ is a feasible solution of $(\mathcal{D}_{\mathsf{hom}})$, and $(\mathcal{D}_{\mathsf{hom}.a})$ holds with equality.

Next, we define $\tilde{z}^{ik} = \min_{r \in R^k} \theta^i \tilde{\ell}_r(w^\dagger) + \sum_{e \in r}(\tilde{p}_e + g_e)$. This ensures that

$$\tilde{z}^{ik} \leqslant \theta^i \tilde{\ell}_r(w^\dagger) + \sum_{e \in r}(\tilde{p}_e + g_e), \quad \forall k \in K, r \in R^k, i \in I.$$

Therefore, $(\tilde{p}, \tilde{z})$ is a feasible solution of the primal problem $(\mathcal{P}_{\mathsf{hom}})$. Moreover, we note that $(\tilde{p}, \tilde{z}, \tilde{q})$ satisfies the complementary slackness condition associated with $(\mathcal{P}_{\mathsf{hom}})$ and $(\mathcal{D}_{\mathsf{hom}})$. Thus, $(\tilde{p}, \tilde{z})$ is an optimal solution to $(\mathcal{P}_{\mathsf{hom}})$ and $\tilde{q}$ is an optimal solution to $(\mathcal{D}_{\mathsf{hom}})$.

(2) The proof of this part follows an analogous procedure as that in part (1). We denote an optimal solution of $(\mathcal{P}_{\mathsf{het}})$ as $(p^\dagger, z^\dagger)$, and an optimal solution of $(\mathcal{D}_{\mathsf{het}})$ as $q^\dagger$. From the complementary slackness condition associated with $(\mathcal{P}_{\mathsf{het}})$-$(\mathcal{D}_{\mathsf{het}})$, we know that if $q_r^{\dagger ik} > 0$ for some $i \in I, k \in K, r \in R^k$, then $z^{\dagger ik} = \theta^i \tilde{\ell}_r(w^\dagger) + \sum_{e \in r}(p_e^{\dagger i} + g_e) = \theta^i c_r^i(q^\dagger, p^\dagger)$. Moreover, we know that for every $i \in I, k \in K, r' \in R^k$,

$$z^{\dagger ik} \leqslant \theta^i \tilde{\ell}_{r'}(w^\dagger) + \sum_{e \in r'}(p_e^{\dagger i} + g_e) = \theta^i c_{r'}^i(q, p^\dagger),$$

which implies that $c_r^i(q^\dagger, p^\dagger) \leqslant c_{r'}^i(q^\dagger, p^\dagger)$, i.e. $q^\dagger$ sends flow on routes with the minimum cost associated with the heterogeneous toll $p^\dagger$ and the socially optimal edge flow $w^\dagger$. Moreover, following the same procedure as that in the hom case, we can argue that $q^\dagger$ induces the socially optimal edge flow $f^\dagger$ (i.e. $(\mathcal{D}_{\mathsf{het}.a})$ is tight), otherwise we arrive at a contradiction that $f^\dagger$ is not socially optimal. Therefore, we can conclude that $q^\dagger$ is an equilibrium strategy distribution that induces the socially optimal (type-specific) edge flow $f^\dagger$ given the het toll vector $p^\dagger$.

On the other hand, suppose that there exists a het toll vector $\tilde{p}$ that induces the socially optimal edge flow $w^\dagger$in equilibrium, then we define $\tilde{z}^{ik} = \min_{r \in R^k} \theta^i \ell_r(w^\dagger) + \sum_{e \in r}(\tilde{p}_e^i + g_e)$ for all $k \in K$, $r \in R^k$, and $i \in I$. Analogous to the case with hom toll, we can argue that $(\tilde{p}, \tilde{z})$ (resp. $\tilde{q}$) is a feasible solution of $(\mathcal{P}_{\mathsf{het}})$ (resp. $(\mathcal{D}_{\mathsf{het}})$), and satisfies complementary slackness conditions. Consequently, we know that $(\tilde{p}, \tilde{z})$ (resp. $\tilde{q}$) is an optimal solution of $(\mathcal{P}_{\mathsf{het}})$ (resp. $(\mathcal{D}_{\mathsf{het}})$). $\square$

Appendix I

Appendix for Chapter 10

I.1 Technical Results in the Proof of Theorem 10.4.1

First, we list some fundamental facts regarding projections onto convex, compact subsets of an Euclidean space. Below, for any fixed convex, compact subset $\Omega \subset \mathbb{R}^d$, we denote the projection operator onto Ω by $\mathcal{P}_\Omega(x) := \text{argmin}_{z\in\Omega}\|x - z\|_2$ for each $x \in \mathbb{R}^d$. Note that $\mathcal{P}_\Omega(x)$ is well-defined (i.e., exists and is unique) for each $x \in \mathbb{R}^d$, if $\Omega \subset \mathbb{R}^d$ were convex and compact.

Proposition I.1.1. *Let $\Omega \subset \mathbb{R}^d$ be compact and convex, and fix $x, y \in \mathbb{R}^d$ arbitrarily. Then:*

$$\begin{aligned}
&\left\|\mathcal{P}_\Omega(x) - \mathcal{P}_\Omega(y)\right\|_2^2 \leqslant \Big(\mathcal{P}_\Omega(x) - \mathcal{P}_\Omega(y)\Big)^\top (x - y),\\
&\|\mathcal{P}_\Omega(x) - \mathcal{P}_\Omega(y)\|_2 \leqslant \|x - y\|_2.
\end{aligned}$$

Proof. From [314], Lemma 3.1.4 (see also [349], Lemma 7.4), we have:

$$\begin{aligned}
\Big(\mathcal{P}_\Omega(x) - \mathcal{P}_\Omega(y)\Big)^\top \Big(x - \mathcal{P}_\Omega(x)\Big) \geqslant 0,\\
\Big(\mathcal{P}_\Omega(y) - \mathcal{P}_\Omega(x)\Big)^\top \Big(y - \mathcal{P}_\Omega(y)\Big) \geqslant 0.
\end{aligned}$$

Adding the two expressions and rearranging terms, we obtain:

$$\begin{aligned}
&\Big(\mathcal{P}_\Omega(x) - \mathcal{P}_\Omega(y)\Big)^\top \Big((x - y) - (\mathcal{P}_\Omega(x) - \mathcal{P}_\Omega(y))\Big) \geqslant 0,\\
\Rightarrow &\|\mathcal{P}_\Omega(x) - \mathcal{P}_\Omega(y)\|_2^2 \leqslant \Big(\mathcal{P}_\Omega(x) - \mathcal{P}_\Omega(y)\Big)^\top (x - y),
\end{aligned}$$

as given in the first claim. The Cauchy Schwarz inequality then implies:

$$\begin{aligned}
\|\mathcal{P}_\Omega(x) - \mathcal{P}_\Omega(y)\|_2^2 &\leqslant \Big(\mathcal{P}_\Omega(x) - \mathcal{P}_\Omega(y)\Big)^\top (x - y)\\
&\leqslant \|\mathcal{P}_\Omega(x) - \mathcal{P}_\Omega(y)\|_2 \cdot \|x - y\|_2.
\end{aligned}$$

If $\mathcal{P}_\Omega(x) = \mathcal{P}_\Omega(y)$, then the second claim becomes $0 \leqslant \|x - y\|_2$, which is clearly true. Otherwise, dividing both sides above by $\|\mathcal{P}_\Omega(x) - \mathcal{P}_\Omega(y)\|_2$ gives the second claim. □

Lemma I.1.2. *Let $\Omega \subset \mathbb{R}^d$ be a compact, convex subset of $\mathbb{R}^d$, and consider the update $z_{k+1} = \mathcal{P}_\Omega(z_k - \eta F(z_{k+1}) + \gamma_k)$, where $z_k, z_{k+1}, \gamma_k \in \mathbb{R}^d$. Then, for each $z \in \Omega$:*

$$\begin{aligned}&\langle F(z_{k+1}), z_{k+1} - z\rangle \\ \leqslant &\frac{1}{2\eta}\|z_k - z\|^2 - \frac{1}{2\eta}\|z_{k+1} - z\|^2 - \frac{1}{2\eta}\|z_{k+1} - z_k\|^2 + \frac{1}{\eta}\langle \gamma_k, z_{k+1} - z\rangle .\end{aligned}$$

Proof. Note that:

$$\begin{aligned}\|z_{k+1} - z\|^2 &= \|z_{k+1} - z_k + z_k - z\|^2 \\ &= \|z_{k+1} - z_k\|^2 + \|z_k - z\|^2 + 2\langle z_{k+1} - z_k, z_k - z\rangle \\ &= \|z_{k+1} - z_k\|^2 + \|z_k - z\|^2 + 2\langle z_{k+1} - z_k, z_k - z_{k+1} + z_{k+1} - z\rangle \\ &= \|z_k - z\|^2 - \|z_{k+1} - z_k\|^2 + 2\langle z_{k+1} - z_k, z_{k+1} - z\rangle\end{aligned}$$

By definition of z_{k+1}, and optimality conditions for the projection operator:

$$\begin{aligned}&\langle z_{k+1} - z, z_{k+1} - z_k + \eta F(z_{k+1}) - \gamma_k\rangle \leqslant 0, \\ \implies &\langle z_{k+1} - z_k, z_{k+1} - z\rangle \leqslant \langle \gamma_k, z_{k+1} - z\rangle - \eta \cdot \langle F(z_{k+1}), z_{k+1} - z\rangle .\end{aligned}$$

Substituting back, we obtain:

$$\begin{aligned}\|z_{k+1} - z\|^2 &= \|z_k - z\|^2 - \|z_{k+1} - z_k\|^2 + 2\langle z_{k+1} - z_k, z_{k+1} - z\rangle \\ &\leqslant \|z_k - z\|^2 - \|z_{k+1} - z_k\|^2 + 2\langle \gamma_k, z_{k+1} - z\rangle - 2\eta \cdot \langle F(z_{k+1}), z_{k+1} - z\rangle .\end{aligned}$$

Rearranging and dividing by η gives the claim in the lemma. □

Next, we state the properties of the mean and variance of the zeroth-order gradient estimator defined in Section 10.4 ([66], Lemma C.1). Below, we define the δ-smoothed loss function $L^\delta : \mathbb{R}^d \to \mathbb{R}$ by $L^\delta(u) := \mathbb{E}_{\overline{v}\sim \mathsf{Unif}(B^d)}[L(u + \delta\overline{v})]$, where $\mathcal{S}^{d-1}$ denotes the $(d-1)$-dimensional unit sphere in $\mathbb{R}^d$, B^d denotes the d-dimensional unit open ball in $\mathbb{R}^d$, and $\mathsf{Unif}(\cdot)$ denotes the continuous uniform distribution over a set. Similarly, we define $L_i^\delta : \mathbb{R}^d \to \mathbb{R}$ by $L_i^\delta(u) := \mathbb{E}_{\overline{v}\sim \mathsf{Unif}(B^d)}[L_i(u + \delta\overline{v})]$, for each $i \in [n] := \{1, \cdots, n\}$. We further define $\delta \cdot \mathcal{S}^{d-1} := \{\delta v : v \in \mathcal{S}^{d-1}\}$ and $\delta \cdot B^d := \{\delta\overline{v} : \overline{v} \in B^d\}$. Finally, we use $\mathsf{vol}_d(\cdot)$ to denote the volume of a set in d dimensions.

Proposition I.1.3. *Let $\hat{F}(u; \delta, v) = \frac{d}{\delta} \cdot L(u + \delta v)v$ and $F(u) = \nabla L(u)$. Then the following holds:*

$$\mathbb{E}_{v\sim \mathsf{Unif}(\mathcal{S}^{d-1})}\left[\hat{F}(u; \delta, v)\right] = \nabla L^\delta(u), \tag{I.1}$$

$$\|\nabla L^\delta(u) - F(u)\|_2 \leqslant \ell\delta, \tag{I.2}$$

$$\|\hat{F}(u; \delta, v)\|_2 \leqslant dG + \frac{dM_L}{\delta}, \tag{I.3}$$

$$\|\hat{F}(u; \delta, v) - F(u)\| \leqslant \min\left\{(d+1)G + \frac{dM_L}{\delta}, \ell\delta + 2dG + \frac{2dM_L}{\delta}\right\}. \tag{I.4}$$

Proof. First, to establish (I.1), observe that since $L^\delta(u) = \mathbb{E}_{v\sim\mathsf{Unif}(B^d)}[L(u+\delta v)]$ and $\hat{F}(u;\delta,v) = \frac{d}{\delta}\cdot L(u+\delta v)v$ for each $u\in\mathbb{R}^d$, $\delta>0$, and $v\in\mathcal{S}^{d-1}$:

$$\begin{aligned}
\nabla L^\delta(u) &= \nabla \mathbb{E}_{\overline{v}\sim\mathsf{Unif}(B^d)}\Big[L(u+\delta\overline{v})\Big] \\
&= \nabla \mathbb{E}_{\overline{v}\sim\mathsf{Unif}(\delta\cdot B^d)}\Big[L(u+\overline{v})\Big] \\
&= \frac{1}{\mathsf{vol}_d(\delta\cdot B^d)}\cdot\nabla\left(\int_{\delta\cdot B^d} L(u+\overline{v})\,d\overline{v}\right) \\
&= \frac{1}{\mathsf{vol}_d(\delta\cdot B^d)}\cdot\int_{\delta\cdot\mathcal{S}^{d-1}} L(u+v)\cdot\frac{v}{\|v\|_2}\,dv, \qquad \text{(I.5)} \\
\mathbb{E}_{v\sim\mathsf{Unif}(\mathcal{S}^{d-1})}\Big[\hat{F}(u;\delta,v)\Big] &= \frac{d}{\delta}\cdot\mathbb{E}_{v\sim\mathsf{Unif}(\mathcal{S}^{d-1})}\Big[L(u+\delta v)v\Big] \\
&= \frac{d}{\delta}\cdot\mathbb{E}_{v\sim\mathsf{Unif}(\delta\cdot\mathcal{S}^{d-1})}\left[L(u+v)\cdot\frac{v}{\|v\|_2}\right] \\
&= \frac{d}{\delta}\cdot\frac{1}{\mathsf{vol}_{d-1}(\delta\cdot\mathcal{S}^{d-1})}\cdot\int_{\delta\cdot\mathcal{S}^{d-1}} L(u+v)\cdot\frac{v}{\|v\|_2}\,dv,
\end{aligned}$$

where (I.5) follows because Stokes' Theorem (see, e.g., Lee, Theorem 16.11 [236]) implies that:

$$\nabla\int_{\delta\cdot B^d} L(u+\overline{v})\,d\overline{v} = \int_{\delta\cdot\mathcal{S}^{d-1}} L(u+v)\cdot\frac{v}{\|v\|_2}\,dv.$$

The equality (I.1) now follows by observing that the surface-area-to-volume ratio of $\delta\cdot B^d$ is d/δ.

Next, to establish (I.2), we note that:

$$\begin{aligned}
\|\nabla L^\delta(u) - F(u)\|_2 &= \Big\|\nabla\mathbb{E}_{\overline{v}\sim\mathsf{Unif}(B^d)}\Big[L^\delta(u)-L(u)\Big]\Big\|_2 \\
&= \frac{1}{\mathsf{vol}_d(B^d)}\cdot\left\|\nabla\left(\int_{B^d}\Big[L(u+\delta\overline{v})-L(u)\Big]\,d\overline{v}\right)\right\|_2 \\
&\leqslant \frac{1}{\mathsf{vol}_d(B^d)}\cdot\left\|\int_{B^d}\Big[F(u+\delta\overline{v})-F(u)\Big]\,d\overline{v}\right\|_2 \qquad \text{(I.6)} \\
&\leqslant \frac{1}{\mathsf{vol}_d(B^d)}\cdot\int_{B^d}\Big\|F(u+\delta\overline{v})-F(u)\Big\|_2\,d\overline{v} \\
&\leqslant \frac{1}{\mathsf{vol}_d(B^d)}\cdot\int_{B^d}\ell\delta\cdot\|\overline{v}\|_2\,d\overline{v} \\
&\leqslant \ell\delta,
\end{aligned}$$

where (I.6) follows by differentiating under the integral sign (see, e.g., Rudin, Theorem 9.42 [356]), and the remaining inequalities follow from the fact that F is ℓ-Lipschitz.

Next, we establish (I.3) by using the triangle inequality and the M_L-boundedness of $L(\cdot)$ on $\mathcal{X} \times \mathcal{Y}$, and the G-Lipschitzness of $L(\cdot)$:

$$\begin{aligned}|\hat{F}(u;\delta,v)| &= \frac{d}{\delta}|L(u+\delta v)| \cdot \|v\|_2 \\ &\leqslant \frac{d}{\delta} \cdot \Big(|L(u)| + |L(u+\delta v) - L(u)|\Big) \cdot 1 \\ &\leqslant \frac{d}{\delta} \cdot (M_L + \delta G).\end{aligned}$$

We can then use (I.3) to establish (I.4) by observing that:

$$|\hat{F}(u;\delta,v) - F(u)| \leqslant |\hat{F}(u;\delta,v)| + |F(u)| \leqslant (d+1)G + \frac{dM_L}{\delta}.$$

and, from (I.3):

$$\begin{aligned}&|\hat{F}(u;\delta,v) - F(u)| \\ \leqslant &\Big|\hat{F}(u;\delta,v) - \mathbb{E}_v[\hat{F}(u;\delta,v)|u]\Big| + \Big|\mathbb{E}_v[\hat{F}(u;\delta,v)|u] - F(u)\Big| \\ \leqslant &\Big|\hat{F}(u;\delta,v) - \mathbb{E}_v[\hat{F}(u;\delta,v)|u]\Big| + \Big|\nabla L^{\delta}(u) - F(u)\Big| \\ \leqslant &2\left(dG + \frac{dM_L}{\delta}\right) + \ell\delta.\end{aligned}$$

This concludes the proof. □

Below, we present technical lemmas that allow us to analyze the convergence rate of the correlated iterates $\{u_i^t\}$ in our random reshuffling-based OGDA Algorithm (Algorithm 9).

Let $\sigma^0, \cdots, \sigma^{t-1}$ denote the permutations drawn from epoch 0 to epoch $t-1$, and let $\{u_i^t(\sigma^t)\}_{1\leqslant i\leqslant n}$ and $\{u_i^t(\tilde{\sigma}^t)\}_{1\leqslant i\leqslant n}$ denote the iterates obtained at epoch t, when the permutations σ^t and $\tilde{\sigma}^t$ are used for the epoch t, respectively. Moreover, let $\mathcal{D}_{i,t}$ denote the distribution of $\{u_i^t(\sigma^t)\}_{1\leqslant i\leqslant n}$ under σ^t, and for $1 \leqslant r \leqslant n$ let $\mathcal{D}_{i,t}^{(r)}$ denote the distribution of $\{u_i^t(\sigma^t)\}_{1\leqslant i\leqslant n}$ with σ^t conditioned on the event $\{\sigma_{i-1}^t = r\}$.

We use the *p-Wasserstein distance between probability distributions on* $\mathbb{R}^d$, defined below, to characterize the distance between $\mathcal{D}_{i,t}$ and $\mathcal{D}_{i,t}^{(r)}$. This is used in the coupling-based techniques employed to establish non-asymptotic convergence results for our random reshuffling algorithm. Note the difference between the p-Wasserstein distance for probability distributions on $\mathbb{R}^d$, and the *Wasserstein distance on* $\mathcal{Z} := \mathbb{R}^d \times \{+1, -1\}$ *associated with a metric* $c : \mathcal{Z} \times \mathcal{Z} \to [0, \infty)$, defined in Chapter 10.3 (Definition 10.1.1).

Definition I.1.1 (***p*-Wasserstein distance between distributions on $\mathbb{R}^d$**)**.** *Let μ, ν be probability distributions over $\mathbb{R}^d$ with finite p-th moments, for some $p \geqslant 1$, and let $\Pi(\mu,\nu)$ denote the set of all couplings (joint distributions) between μ and ν. The p-Wasserstein distance between μ and ν, denoted $\mathcal{W}_p(\mu,\nu)$, is defined by:*

$$\mathcal{W}_p(\mu,\nu) = \inf_{(X,X')\sim\pi\in\Pi(\mu,\nu)} \Big(\mathbb{E}_\pi\Big[\|X - X'\|^p\Big]\Big)^{1/p}.$$

The following proposition characterizes the 1-Wasserstein distance as a measure of the gap between Lipschitz functions of random variables.

Proposition I.1.4 (**Kantorovich Duality**). *If μ, ν are probability distributions over $\mathbb{R}^d$ with finite second moments, then:*

$$\mathcal{W}_1(\mu, \nu) = \sup_{g \in \mathrm{Lip}(1)} \mathbb{E}_{X \sim \mu}[g(X)] - \mathbb{E}_{Y \sim \nu}[g(Y)],$$

where $\mathrm{Lip}(1) := \{g : \mathbb{R}^d \to \mathbb{R} : g \text{ is 1-Lipschitz}\}$.

Using [442, Lemma C.2], we now bound the difference between the unbiased gap $\mathbb{E}[\Delta(u_i^t)]$ and the biased gap $\mathbb{E}[L_{\sigma_i^t}(x_{i+1}^t, y^\star) - L_{\sigma_i^t}(x^\star, y_{i+1}^t)]$ using the Wasserstein metric.

Lemma I.1.5. *Let $u^\star := (x^\star, y^\star) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} = \mathbb{R}^d$ denote a saddle point of the min-max optimization problem* (10.9)*. Then, for each $t \in [T]$ and $i \in [n]$, the iterates $\{u_i^t\} = \{(x_i^t, y_i^t)\}$ of the OGDA-RR algorithm satisfy:*

$$\left|\mathbb{E}[\Delta(u_{i+1}^t)] - \mathbb{E}\Big[L_{\sigma_i^t}(x_{i+1}^t, y^\star) - L_{\sigma_i^t}(x^\star, y_{i+1}^t)\Big]\right| \leqslant \frac{G}{n} \sum_{r=1}^{n} \mathcal{W}_2\Big(\mathcal{D}_{i+1,t}, \mathcal{D}_{i+1,t}^r\Big).$$

Proof. Since σ^t and $\tilde{\sigma}^t$ are independently generated permutations of $[n]$, the iterates

$$\{u_i^t\}_{1 \leqslant i \leqslant n} = \{u_i^t(\sigma^t)\}_{1 \leqslant i \leqslant n}$$

and

$$\{u_i^t(\tilde{\sigma}^t)\}_{1 \leqslant i \leqslant n}$$

are i.i.d. Thus, we have:

$$\mathbb{E}[\Delta(u_{i+1}^t)] = \mathbb{E}\Big[L_{\sigma_i^t}(x_{i+1}^t(\tilde{\sigma}^t), y^\star) - L_{\sigma_i^t}(x^\star, y_{i+1}^t(\tilde{\sigma}^t))\Big],$$

and thus:

$$
\begin{aligned}
&\Big|\mathbb{E}[\Delta(u_{i+1}^t)] - \mathbb{E}\Big[L_{\sigma_i^t}(x_{i+1}^t, y^\star) - L_{\sigma_i^t}(x^\star, y_{i+1}^t)\Big]\Big| \\
=&\Big|\mathbb{E}\Big[L_{\sigma_i^t}(x_{i+1}^t(\tilde\sigma^t), y^\star) - L_{\sigma_i^t}(x^\star, y_{i+1}^t(\tilde\sigma^t))\Big] - \mathbb{E}\Big[L_{\sigma_i^t}(x_{i+1}^t, y^\star) - L_{\sigma_i^t}(x^\star, y_{i+1}^t)\Big]\Big| \\
=&\Bigg|\frac{1}{n}\sum_{r=1}^n \mathbb{E}\Big[L_r(x_{i+1}^t(\tilde\sigma^t), y^\star) - L_r(x^\star, y_{i+1}^t(\tilde\sigma^t))\Big] \qquad \text{(I.7)} \\
&\qquad\qquad - \frac{1}{n}\sum_{r=1}^n \mathbb{E}\Big[L_r(x_{i+1}^t, y^\star) - L_r(x^\star, y_{i+1}^t)\Big|\sigma_i^t = r\Big]\Bigg| \\
\leqslant&\frac{1}{n}\sum_{r=1}^n \Big|\mathbb{E}\Big[L_r(x_{i+1}^t(\tilde\sigma^t), y^\star) - L_r(x^\star, y_{i+1}^t(\tilde\sigma^t))\Big] - \mathbb{E}\Big[L_r(x_{i+1}^t, y^\star) - L_r(x^\star, y_{i+1}^t)\Big|\sigma_i^t = r\Big]\Big| \\
\leqslant&\frac{1}{n}\sum_{r=1}^n \sup_{g\in\mathrm{Lip}(G)}\Big(\mathbb{E}\Big[g(x_{i+1}^t(\tilde\sigma^t), y_{i+1}^t(\tilde\sigma^t))\Big] - \mathbb{E}\Big[g(x_{i+1}^t, y_{i+1}^t)|\sigma_i^t = r\Big]\Big) \qquad \text{(I.8)} \\
\leqslant&\frac{1}{n}\sum_{r=1}^n G\cdot\mathcal{W}_1(\mathcal{D}_{i+1,t}, \mathcal{D}_{i+1,t}^{(r)}) \qquad \text{(I.9)} \\
\leqslant&\frac{1}{n}\sum_{r=1}^n G\cdot\mathcal{W}_2(\mathcal{D}_{i+1,t}, \mathcal{D}_{i+1,t}^{(r)}), \qquad \text{(I.10)}
\end{aligned}
$$

where (I.7) follows by properties of the conditional expectation on $\{\sigma_i^t = r\}$ and the fact that σ^t and $\tilde\sigma^t$ are independent, (I.8) follows from the fact that L is Lipschitz, (I.9) follows from Proposition I.1.4, and (I.10) follows from the fact that $\mathcal{W}_1(\mu,\nu) \leqslant \mathcal{W}_2(\mu,\nu)$ for any two probability distributions μ, ν. □

The next lemma bounds the difference in the iterates $\{u_i^t(\sigma^t)\}$ and $\{u_i^t(\tilde\sigma^t)\}$ (assuming, as before, that $\sigma^0, \cdots, \sigma^{t-1}$ were fixed and identical for both sequences.)

Lemma I.1.6. *Denote, with a slight abuse of notation, $u_i^t := u_i^t(\sigma^t)$ and $\tilde u_i^t := u_i^t(\tilde\sigma^t)$. Then:*

$$
\|u_{i+1}^t - \tilde u_{i+1}^t\|_2 \leqslant \left(6nd + 14n + 2\cdot\sum_{i=1}^n \mathbf{1}\{\sigma_i^t \neq \tilde\sigma_i^t\}\right) G\cdot\eta^t + 6ndM_L\cdot\frac{\eta^t}{\delta^t}.
$$

Proof. Our proof strategy is to bound the differences between zeroth-order and first-order

OGDA updates, and between the OGDA and proximal point updates. To this end, we define:

$$
\begin{aligned}
u_{i+1}^t &= \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}\Big(u_i^t - \eta^t \hat{F}_{\sigma_i^t}(u_i^t;\delta^t,v_i^t) - \eta^t \hat{F}_{\sigma_{i-1}^t}(u_i^t;\delta^t,v_i^t) + \eta^t \hat{F}_{\sigma_{i-1}^t}(u_{i-1}^t;\delta^t,v_{i-1}^t)\Big),\\
\tilde{u}_{i+1}^t &= \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}\Big(\tilde{u}_i^t - \eta^t \hat{F}_{\tilde{\sigma}_i^t}(\tilde{u}_i^t;\delta^t,v_i^t) - \eta^t \hat{F}_{\tilde{\sigma}_{i-1}^t}(\tilde{u}_i^t;\delta^t,v_i^t) + \eta^t \hat{F}_{\tilde{\sigma}_{i-1}^t}(\tilde{u}_{i-1}^t;\delta^t,v_{i-1}^t)\Big),\\
v_{i+1}^t &= \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}\Big(u_i^t - \eta^t F_{\sigma_i^t}(u_i^t) - \eta^t F_{\sigma_{i-1}^t}(u_i^t) + \eta^t F_{\sigma_{i-1}^t}(u_{i-1}^t)\Big),\\
\tilde{v}_{i+1}^t &= \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}\Big(\tilde{u}_i^t - \eta^t F_{\tilde{\sigma}_i^t}(\tilde{u}_i^t) - \eta^t F_{\tilde{\sigma}_{i-1}^t}(\tilde{u}_i^t) + \eta^t F_{\tilde{\sigma}_{i-1}^t}(\tilde{u}_{i-1}^t)\Big),\\
w_{i+1}^t &= \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}\Big(u_i^t - \eta^t F_{\sigma_i^t}(w_{i+1}^t)\Big),\\
\tilde{w}_{i+1}^t &= \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}\Big(\tilde{u}_i^t - \eta^t F_{\tilde{\sigma}_i^t}(\tilde{w}_{i+1}^t)\Big).
\end{aligned}
$$

By the triangle inequality:

$$
\begin{aligned}
\|u_{i+1}^t - \tilde{u}_{i+1}^t\|_2 \leqslant \|u_{i+1}^t - v_{i+1}^t\|_2 + \|v_{i+1}^t - w_{i+1}^t\|_2 + \|w_{i+1}^t - \tilde{w}_{i+1}^t\|_2 \\
+ \|\tilde{w}_{i+1}^t - \tilde{v}_{i+1}^t\|_2 + \|\tilde{v}_{i+1}^t - \tilde{u}_{i+1}^t\|_2.
\end{aligned} \tag{I.11}
$$

Observe that bounding the fourth term is equivalent to bounding the second term, and bounding the fifth term is equivalent to bounding the first term.

To bound the first term on the right hand side, we use Proposition I.1.3 to conclude that:

$$
\begin{aligned}
\|u_{i+1}^t - v_{i+1}^t\|_2 &\leqslant \eta^t \cdot \|\hat{F}_{\sigma_i^t}(u_i^t;\delta^t,v_i^t) - F_{\sigma_i^t}(u_i^t)\| + \eta^t \cdot \|\hat{F}_{\sigma_{i-1}^t}(u_i^t;\delta^t,v_i^t) - F_{\sigma_{i-1}^t}(u_i^t)\| \\
&\quad + \eta^t \cdot \|\hat{F}_{\sigma_{i-1}^t}(u_{i-1}^t;\delta^t,v_{i-1}^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t)\| \\
&\leqslant 3(d+1)G\eta^t + 3dM_L \cdot \frac{\eta^t}{\delta^t}
\end{aligned} \tag{I.12}
$$

For the second term, we use the G-Lipschitzness of L_r, for each $r \in [n]$ to conclude that:

$$
\begin{aligned}
\|v_{i+1}^t - w_{i+1}^t\|_2 &\leqslant \eta^t \cdot |F_{\sigma_i^t}(u_i^t)| + \eta^t \cdot |F_{\sigma_{i-1}^t}(u_i^t)| + \eta^t \cdot |F_{\sigma_{i-1}^t}(u_{i-1}^t)| + \eta^t \cdot |F_{\sigma_i^t}(w_{i+1}^t)| \\
&\leqslant 4G \cdot \eta^t.
\end{aligned} \tag{I.13}
$$

For the third term, we observe that if $\sigma_i^t \neq \tilde{\sigma}_i^t$, then:

$$
\begin{aligned}
\|w_{i+1}^t - \tilde{w}_{i+1}^t\|_2 &\leqslant \|u_i^t - \tilde{u}_i^t\|_2 + \eta^t \cdot \|F_{\sigma_i^t}(w_{i+1}^t) - F_{\tilde{\sigma}_i^t}(\tilde{w}_{i+1}^t)\|_2 \\
&\leqslant \|u_i^t - \tilde{u}_i^t\|_2 + 2G \cdot \eta^t.
\end{aligned} \tag{I.14}
$$

On the other hand, if $\sigma_i^t = \tilde{\sigma}_i^t$, then:

$$
\begin{aligned}
w_{i+1}^t &= \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}\Big(u_i^t - \eta^t F_{\sigma_i^t}(w_{i+1}^t)\Big),\\
\tilde{w}_{i+1}^t &= \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}\Big(\tilde{u}_i^t - \eta^t F_{\sigma_i^t}(\tilde{w}_{i+1}^t)\Big),
\end{aligned}
$$

so we have:

$$\begin{aligned}
&\|w_{i+1}^t - \tilde{w}_{i+1}^t\|_2^2 \\
\leqslant &(w_{i+1}^t - \tilde{w}_{i+1}^t)^\top \Big((u_i^t - \eta \cdot F_{\sigma_i^t}(w_{i+1}^t)) - (\tilde{u}_i^t - \eta \cdot F_{\sigma_i^t}(\tilde{w}_{i+1}^t))\Big) &\text{(I.15)}\\
= &(w_{i+1}^t - \tilde{w}_{i+1}^t)^\top (u_i^t - \tilde{u}_i^t) - \eta (w_{i+1}^t - \tilde{w}_{i+1}^t)^\top \Big(F_{\sigma_i^t}(w_{i+1}^t)) - F_{\sigma_i^t}(\tilde{w}_{i+1}^t))\Big) \\
\leqslant &(w_{i+1}^t - \tilde{w}_{i+1}^t)^\top (u_i^t - \tilde{u}_i^t) &\text{(I.16)}\\
\leqslant &\|w_{i+1}^t - \tilde{w}_{i+1}^t\|_2 \cdot \|u_i^t - \tilde{u}_i^t\|_2, &\text{(I.17)}
\end{aligned}$$

so $\|w_{i+1}^t - \tilde{w}_{i+1}^t\|_2 \leqslant \|u_i^t - \tilde{u}_i^t\|_2$. Here, (I.15) follows from the definitions of w_{i+1}^t and $\tilde{w}_{i+1}^t$, as well as Proposition I.1.1, while (I.16) holds because the monotonicity of F_i, for each $i \in [n]$, implies that $(w_{i+1}^t - \tilde{w}_{i+1}^t)^\top \Big(F_{\sigma_i^t}(w_{i+1}^t) - F_{\sigma_i^t}(\tilde{w}_{i+1}^t)\Big) \geqslant 0$. Putting together (I.12), (I.13), (I.14), (I.17), we have:

$$\begin{aligned}
\|u_{i+1}^t - \tilde{u}_{i+1}^t\|_2 \leqslant \|u_i^t - \tilde{u}_i^t\|_2 + (6d+14)G \cdot \eta^t + 6dM_L \cdot \frac{\eta^t}{\delta^t} \\
+ 2G \cdot \mathbf{1}\{\sigma_i^t \neq \tilde{\sigma}_i^t\} \cdot \eta^t,
\end{aligned}$$

where the indicator $\mathbf{1}(A)$ returns 1 if the given event A occurs, and 0 otherwise.

Since $u_0^t = \tilde{u}_0^t$, we can iteratively apply the above inequality to obtain that, for any and epoch t and $i \in [n]$:

$$\|u_{i+1}^t - \tilde{u}_{i+1}^t\|_2 \leqslant (6d+14)nG \cdot \eta^t + 6ndM_L \cdot \frac{\eta^t}{\delta^t} + 2\eta_t G \cdot \sum_{i=1}^{n} \mathbf{1}\{\sigma_i^t \neq \tilde{\sigma}_i^t\}.$$

□

Remark I.1.1. *In the theorems and lemmas below, we will be concerned with the case where σ^t and $\tilde{\sigma}^t$ have the following specific relationship. Let $\mathcal{R}_n$ denote the set of all random permutations over the set $[n]$. For each $l, m \in [n]$, let $S_{l,m} : \mathcal{R}_n \to \mathcal{R}_n$ denote the map that swaps, for each input permutation σ, the l-th and m-th entries. For each $r, i \in [n]$, define the map $\omega_{r,i} : \mathcal{R}_n \to \mathcal{R}_n$ as follows:*

$$\omega_{r,i}(\sigma) = \begin{cases} \sigma, & \text{if } \sigma_{i-1} = r, \\ S_{i-1,j}(\sigma), & \text{if } \sigma_j = r \text{ and } j \neq i-1. \end{cases}.$$

Intuitively, $\omega_{r,i}$ performs a single swap such that the $(i-1)$-th position of the permutation is r. Clearly, if σ^t is a random permutation (i.e., selected from a uniform distribution over $\mathcal{R}_n$), then $\omega_{r,i}(\sigma^s)$ has the same distribution as $\sigma^t | (\sigma_{i-1}^t = r)$. Based on this construction, we have $u_i(\sigma^t) \sim \mathcal{D}_{i,t}$ and $u_i(\omega_{r,i}(\sigma^t)) \sim \mathcal{D}_{i,t}^{(r)}$. This gives a coupling between $\mathcal{D}_{s,t}$ and $\mathcal{D}_{s,t}^{(r)}$.

Since σ^t and $\tilde{\sigma}^t$ differ by at most two entries, by iteratively applying Lemma I.1.6, we have:

$$\begin{aligned}\|u_{i+1}^t - \tilde{u}_{i+1}^t\|_2 &\leqslant n\left((6d+14)G\cdot\eta^t + 6dM_L\cdot\frac{\eta^t}{\delta^t}\right) + 4G\cdot\eta^t \\ &= (6nd+14n+4)G\cdot\eta^t + 6ndM_L\cdot\frac{\eta^t}{\delta^t},\end{aligned}$$

as claimed.

Lemma I.1.7. *If $\eta^t \leqslant 1/(2\ell)$ for each $t \in \{0,1,\cdots,T-1\}$, the iterates $\{u_i^t\} = \{(x_i^t, y_i^t)\}$ of the OGDA-RR algorithm satisfy, for each $u \in \mathcal{X}\times\mathcal{Y}$:*

$$\begin{aligned}&2\eta^t\cdot\mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_{i+1}^t), u_{i+1}^t - u\Big\rangle\Big] \\ \leqslant&\mathbb{E}\Big[\|u_i^t - u\|_2^2\Big] - \mathbb{E}\Big[\|u_{i+1}^t - u\|_2^2\Big] - \frac{1}{2}\mathbb{E}\Big[\|u_{i+1}^t - u_i^t\|_2^2\Big] + \frac{1}{2}\mathbb{E}\Big[\|u_i^t - u_{i-1}^t\|_2^2\Big] \\ &+ 2\eta^t\cdot\mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_{i+1}^t) - F_{\sigma_i^t}(u_i^t), u_{i+1}^t - u\Big\rangle\Big] \\ &- 2\eta^t\cdot\mathbb{E}\Big[\Big\langle F_{\sigma_{i-1}^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t), u_i^t - u\Big\rangle\Big] \\ &+ 6C_1\cdot\left(\eta^t\delta^t + (\eta^t)^2\delta^t + (\eta^t)^2 + \frac{(\eta^t)^2}{\delta^t} + \frac{(\eta^t)^2}{(\delta^t)^2}\right),\end{aligned}$$

where $C_1 := d^2\max\Big\{6G\ell D, 18G^2 + 6M_L\ell D, 30M_LG, 12M_L^2\Big\}$ is a constant independent of the sequences $\{\eta^t\}$ and $\{\delta^t\}$.

Proof. The iterates of the OGDA-RR algorithm are given by:

$$\begin{aligned}u_{i+1}^t &= \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}\Big(u_i^t - \eta^t\hat{F}_{\sigma_i^t}(u_i^t;\delta^t,v_i^t) - \eta^t\hat{F}_{\sigma_{i-1}^t}(u_i^t;\delta^t,v_i^t) \\ &\qquad - \eta^t\hat{F}_{\sigma_{i-1}^t}(u_{i-1}^t;\delta^t,v_{i-1}^t)\Big) \\ &= \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}\Big(u_i^t - \eta^t F_{\sigma_i^t}(u_{i+1}^t) + \eta^t\Big(\gamma_i^t + E_{i,1}^t + E_{i,2}^t + E_{i,3}^t\Big)\Big),\end{aligned} \tag{I.18}$$

where we have defined:

$$\begin{aligned}\gamma_i^t &:= F_{\sigma_i^t}(u_{i+1}^t) - F_{\sigma_i^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_i^t) + F_{\sigma_{i-1}^t}(u_{i-1}^t), \\ E_{i,1}^t &:= F_{\sigma_i^t}(u_i^t) - \hat{F}_{\sigma_i^t}(u_i^t;\delta^t,v_i^t), \\ E_{i,2}^t &:= F_{\sigma_{i-1}^t}(u_i^t) - \hat{F}_{\sigma_{i-1}^t}(u_i^t;\delta^t,v_i^t), \\ E_{i,3}^t &:= F_{\sigma_{i-1}^t}(u_{i-1}^t) - \hat{F}_{\sigma_{i-1}^t}(u_{i-1}^t;\delta^t,v_{i-1}^t).\end{aligned}$$

First, by applying Lemma I.1.2 we have:

$$\begin{aligned} & 2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_{i+1}^t), u_{i+1}^t - u \Big\rangle\Big] \\ \leqslant & \mathbb{E}\Big[\|u_i^t - u\|_2^2\Big] - \mathbb{E}\Big[\|u_{i+1}^t - u\|_2^2\Big] - \mathbb{E}\Big[\|u_{i+1}^t - u_i^t\|_2^2\Big] \\ & \quad + 2\eta^t \cdot \mathbb{E}\Big[\Big\langle \gamma_i^t, u_{i+1}^t - u \Big\rangle\Big] + \sum_{k=1}^{3} 2\eta^t \cdot \mathbb{E}\Big[\Big\langle E_{i,k}^t, u_{i+1}^t - u \Big\rangle\Big]. \end{aligned} \tag{I.19}$$

Below, we proceed to bound the inner product terms on the right-hand-side of (I.19). First, we bound $\Big\langle \gamma_i^t, u_{i+1}^t - u \Big\rangle$:

$$\begin{aligned} \Big\langle \gamma_i^t, u_{i+1}^t - u \Big\rangle &= \Big\langle F_{\sigma_i^t}(u_{i+1}^t) - F_{\sigma_i^t}(u_i^t), u_{i+1}^t - u \Big\rangle \\ &\quad - \Big\langle F_{\sigma_{i-1}^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t), u_{i+1}^t - u \Big\rangle \\ &= \Big\langle F_{\sigma_i^t}(u_{i+1}^t) - F_{\sigma_i^t}(u_i^t), u_{i+1}^t - u \Big\rangle \\ &\quad - \Big\langle F_{\sigma_{i-1}^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t), u_i^t - u \Big\rangle \\ &\quad - \Big\langle F_{\sigma_{i-1}^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t), u_{i+1}^t - u_i^t \Big\rangle \\ &\leqslant \Big\langle F_{\sigma_i^t}(u_{i+1}^t) - F_{\sigma_i^t}(u_i^t), u_{i+1}^t - u \Big\rangle \\ &\quad - \Big\langle F_{\sigma_{i-1}^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t), u_i^t - u \Big\rangle \\ &\quad + \frac{1}{2}\ell \cdot \|u_i^t - u_{i-1}^t\|_2^2 + \frac{1}{2}\ell \cdot \|u_{i+1}^t - u_i^t\|_2^2. \end{aligned} \tag{I.20}$$

Note that the final inequality follows by applying Young's inequality, and noting that F is ℓ-Lipschitz. Next, we bound $\langle E_{i,1}^t, u_{i+1}^t - u \rangle$:

$$\begin{aligned} & \mathbb{E}\Big[\langle E_{i,1}^t, u_{i+1}^t - u \rangle\Big] \\ = & \mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_i^t) - \hat{F}_{\sigma_i^t}(u_i^t, \delta^t, v_i^t), u_{i+1}^t - u \Big\rangle\Big] \\ = & \mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_i^t) - \nabla L_{\sigma_i^t}^{\delta^t}(u_i^t), u_{i+1}^t - u \Big\rangle\Big] \\ & \quad + \mathbb{E}\Big[\Big\langle \mathbb{E}\big[\hat{F}_{\sigma_i^t}(u_i^t; \delta^t, v_i^t | u_i^t\big] - \hat{F}_{\sigma_i^t}(u_i^t, \delta^t, v_i^t), u_{i+1}^t - u \Big\rangle\Big] \\ = & \mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_i^t) - \nabla L_{\sigma_i^t}^{\delta^t}(u_i^t), u_{i+1}^t - u \Big\rangle\Big] \\ & \quad + \mathbb{E}\Big[\Big\langle \mathbb{E}_v\big[\hat{F}_{\sigma_i^t}(u_i^t; \delta^t, v | u_i^t\big] - \hat{F}_{\sigma_i^t}(u_i^t, \delta^t, v_i^t), u_i^t - u \Big\rangle\Big] \\ & \quad + \mathbb{E}\Big[\Big\langle \mathbb{E}_v\big[\hat{F}_{\sigma_i^t}(u_i^t; \delta^t, v | u_i^t\big] - \hat{F}_{\sigma_i^t}(u_i^t, \delta^t, v_i^t), u_{i+1}^t - u_i^t \Big\rangle\Big], \end{aligned} \tag{I.21}$$

where the first equality above follows by applying Proposition I.1.3, (I.1), and we have used the shorthand $\mathbb{E}_v := \mathbb{E}_{v\sim\mathsf{Unif}(\mathcal{S}^{d-1})}$. (Recall that $L^\delta(u) := \mathbb{E}_{v\sim\mathsf{Unif}(\mathcal{S}^{d-1})}\Big[L(u+\delta v)\Big]$) Next, we upper bound each of the three quantities in (I.21). First, by Proposition I.1.3, (I.2), we have:

$$\begin{aligned}
&\mathbb{E}\bigg[\Big\langle F_{\sigma_i^t}(u_i^t) - \nabla L^{\delta^t}_{\sigma_i^t}(u_i^t), u_{i+1}^t - u\Big\rangle\bigg] \\
\leqslant&\mathbb{E}\Big[\|F_{\sigma_i^t}(u_i^t) - \nabla L^{\delta^t}_{\sigma_i^t}(u_i^t)\|_2 \cdot \|u_{i+1}^t - u\|_2\Big] \\
\leqslant&\ell D\cdot\delta^t, \qquad\qquad (\text{I.22})
\end{aligned}$$

with $C_1 > 0$ as given in Lemma I.1.7. Meanwhile, the law of iterated expectations can be used to bound the second quantity:

$$\begin{aligned}
&\mathbb{E}\bigg[\Big\langle \mathbb{E}_v\Big[\hat{F}_{\sigma_i^t}(u_i^t;\delta^t,v)|u_i^t\Big] - \hat{F}_{\sigma_i^t}(u_i^t,\delta^t,v_i^t), u_i^t - u\Big\rangle\bigg] \\
=&\mathbb{E}\bigg[\mathbb{E}_v\Big[\Big\langle \hat{F}_{\sigma_i^t}(u_i^t,\delta^t,v_i^t), u_i^t - u\Big\rangle\Big|u_i^t\Big]\bigg] - \mathbb{E}\bigg[\Big\langle \hat{F}_{\sigma_i^t}(u_i^t,\delta^t,v_i^t), u_i^t - u\Big\rangle\bigg] \\
=&0, \qquad\qquad (\text{I.23})
\end{aligned}$$

and we can upper-bound the third quantity as shown below. By using the compactness of $\mathcal{X}\times\mathcal{Y}$ and the continuity of L, we have:

$$\begin{aligned}
&\mathbb{E}\bigg[\Big\langle \mathbb{E}_v\Big[\hat{F}_{\sigma_i^t}(u_i^t;\delta^t,v)|u_i^t\Big] - \hat{F}_{\sigma_i^t}(u_i^t,\delta^t,v_i^t), u_{i+1}^t - u_i^t\Big\rangle\bigg] \\
\leqslant&\bigg(\Big\|\mathbb{E}_v\Big[\hat{F}_{\sigma_i^t}(u_i^t;\delta^t,v)|u_i^t\Big]\Big\|_2 + \|\hat{F}_{\sigma_i^t}(u_i^t,\delta^t,v_i^t)\|\bigg)\cdot\|u_{i+1}^t - u_i^t\|_2 \\
\leqslant&2\cdot\frac{d}{\delta^t}\cdot\sup_{\substack{u\in\mathcal{X}\times\mathcal{Y}\\ v\sim\mathsf{Unif}(\mathcal{S}^{d-1})}}|L(u_i^t+\delta^t v)|\cdot\|u_{i+1}^t - u_i^t\|_2, \\
\leqslant&2\cdot\frac{d}{\delta^t}\cdot(M_L + \delta^t G)\cdot\|u_{i+1}^t - u_i^t\|_2, \qquad\qquad (\text{I.24})
\end{aligned}$$

and using (I.22) and the bound for each $\|\hat{F}_{\sigma_i^t}\|_2$ given in (I.24), we have:

$$\begin{aligned}
&\|u_{i+1}^t - u_i^t\|_2 \\
\leqslant&\eta^t\cdot\|\hat{F}_{\sigma_i^t}(u_i^t;\delta^t,v_i^t) + \hat{F}_{\sigma_{i-1}^t}(u_i^t;\delta^t,v_i^t) - \hat{F}_{\sigma_{i-1}^t}(u_{i-1}^t;\delta^t,v_{i-1}^t)\| \\
\leqslant&\eta^t\cdot\|F_{\sigma_i^t}(u_i^t) + F_{\sigma_{i-1}^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t)\|_2 \\
&\quad+\eta^t d\cdot\|\hat{F}_{\sigma_i^t}(u_i^t;\delta^t,v_i^t) - F_{\sigma_i^t}(u_i^t)\|_2 \\
&\quad+\eta^t d\cdot\|\hat{F}_{\sigma_{i-1}^t}(u_i^t;\delta^t,v_i^t) - F_{\sigma_{i-1}^t}(u_i^t)\|_2 \\
&\quad+\eta^t d\cdot\|\hat{F}_{\sigma_{i-1}^t}(u_{i-1}^t;\delta^t,v_{i-1}^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t)\|_2 \\
\leqslant&3G\eta^t + 3\eta^t d\cdot\left(2(M_L + G\delta^t)\cdot\frac{1}{\delta^t} + \ell D\cdot\delta^t\right). \qquad\qquad (\text{I.25})
\end{aligned}$$

Substituting (I.25) back into (I.24), we have:

$$\begin{aligned}
&\mathbb{E}\Big[\Big\langle \mathbb{E}_v\big[\hat{F}_{\sigma_i^t}(u_i^t;\delta^t,v)|u_i^t\big] - \hat{F}_{\sigma_i^t}(u_i^t,\delta^t,v_i^t), u_{i+1}^t - u_i^t\Big\rangle\Big]\\
\leqslant& d^2\ell D 6\eta^t G\cdot\delta^t + 6d^2\eta^t(3G^2 + M_L\ell D) + 30d^2\eta^t M_L G\cdot\frac{1}{\delta^t} + 12d^2\eta^t M_L^2\cdot\Big(\frac{1}{\delta^t}\Big)^2\\
\leqslant& C_1\cdot\left(\eta^t\delta^t + \eta^t + \frac{\eta^t}{\delta^t} + \frac{\eta^t}{(\delta^t)^2}\right), \qquad \text{(I.26)}
\end{aligned}$$

where $C_1 := d^2\cdot\max\Big\{6G\ell D, 18G^2 + 6M_L\ell D, 30M_L G, 12M_L^2\Big\}$ is a constant independent of the sequences $\{\eta^t\}$ and $\{\delta^t\}$. The quantities $\mathbb{E}\Big[\langle E_{i,2}^t, u_{i+1}^t - u\rangle\Big]$ and $\mathbb{E}\Big[\langle E_{i,3}^t, u_{i+1}^t - u\rangle\Big]$ can be similarly bounded. Substituting (I.22), (I.23), (I.26) back into (I.21), and substituting (I.21) and (I.20) into (I.19), we find that:

$$\begin{aligned}
&2\eta^t\cdot\mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_{i+1}^t), u_{i+1}^t - u\Big\rangle\Big]\\
=&\mathbb{E}\big[\|u_i^t - u\|_2^2\big] - \mathbb{E}\big[\|u_{i+1}^t - u\|_2^2\big] - \mathbb{E}\big[\|u_{i+1}^t - u_i^t\|_2^2\big]\\
&\quad + 2\eta^t\cdot\mathbb{E}\Big[\Big\langle\gamma_i^t, u_{i+1}^t - u\Big\rangle\Big] + 2\eta^t\cdot\sum_{k=1}^{3}\mathbb{E}\Big[\Big\langle E_{i,k}^t, u_{i+1}^t - u\Big\rangle\Big]\\
\leqslant&\mathbb{E}\big[\|u_i^t - u\|_2^2\big] - \mathbb{E}\big[\|u_{i+1}^t - u\|_2^2\big] - \mathbb{E}\big[\|u_{i+1}^t - u_i^t\|_2^2\big]\\
&\quad + 2\eta^t\cdot\mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_{i+1}^t) - F_{\sigma_i^t}(u_i^t), u_{i+1}^t - u\Big\rangle\Big]\\
&\quad - 2\eta^t\cdot\mathbb{E}\Big[\Big\langle F_{\sigma_{i-1}^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t), u_i^t - u\Big\rangle\Big]\\
&\quad + \eta^t\ell\cdot\mathbb{E}\big[\|u_i^t - u_{i-1}^t\|_2^2\big] + \eta^t\ell\cdot\mathbb{E}\big[\|u_{i+1}^t - u_i^t\|_2^2\big]\\
&\quad + 6C_1\cdot\left(\eta^t\delta^t + (\eta^t)^2\delta^t + (\eta^t)^2 + \frac{(\eta^t)^2}{\delta^t} + \frac{(\eta^t)^2}{(\delta^t)^2}\right),
\end{aligned}$$

In particular, since by assumption $\eta^t \leqslant 1/(2\ell)$ for each $t\in\{0,1,\cdots,T-1\}$, then:

$$\begin{aligned}
&2\eta^t\cdot\mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_{i+1}^t), u_{i+1}^t - u\Big\rangle\Big]\\
\leqslant&\mathbb{E}\big[\|u_i^t - u\|_2^2\big] - \mathbb{E}\big[\|u_{i+1}^t - u\|_2^2\big] - \frac{1}{2}\mathbb{E}\big[\|u_{i+1}^t - u_i^t\|_2^2\big] + \frac{1}{2}\mathbb{E}\big[\|u_i^t - u_{i-1}^t\|_2^2\big]\\
&\quad + 2\eta^t\cdot\mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_{i+1}^t) - F_{\sigma_i^t}(u_i^t), u_{i+1}^t - u\Big\rangle\Big]\\
&\quad - 2\eta^t\cdot\mathbb{E}\Big[\Big\langle F_{\sigma_{i-1}^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t), u_i^t - u\Big\rangle\Big]\\
&\quad + 6C_1\cdot\left(\eta^t\delta^t + (\eta^t)^2\delta^t + (\eta^t)^2 + \frac{(\eta^t)^2}{\delta^t} + \frac{(\eta^t)^2}{(\delta^t)^2}\right),
\end{aligned}$$

□

Finally, to bound the step size terms above, we require the following lemma, which follows from standard calculus arguments.

Lemma I.1.8.

$$\sum_{t=1}^{T} t^{-\beta} \geqslant \frac{1}{1-\beta} T^{1-\beta}, \qquad \forall\, \beta < 1,$$
$$\sum_{t=1}^{T} t^{-(1+\beta)} \leqslant \frac{1}{\beta} + 1, \qquad \forall\, \beta > 0.$$

I.2 Proof of Theorem 10.4.1

By applying Lemma I.1.7 (note that $\eta^t \leqslant \eta^0 \leqslant \frac{1}{2\ell}$, for each $t \in \{0, 1, \cdots, T-1\}$) and using convex-concave nature of L_r (refer Proposition 1 in [304]), for each $r \in \{1, \cdots n\}$, we have:

$$\begin{aligned}
&2\eta^t \cdot \mathbb{E}\Big[L_{\sigma_i^t}(x_{i+1}^t, y^\star) - L_{\sigma_i^t}(x^\star, y_{i+1}^t)\Big] \\
\leqslant &2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_{i+1}^t), u_{i+1}^t - u^\star \Big\rangle\Big] \\
\leqslant &\mathbb{E}\Big[\|u_i^t - u^\star\|_2^2\Big] - \mathbb{E}\Big[\|u_{i+1}^t - u^\star\|_2^2\Big] - \frac{1}{2}\mathbb{E}\Big[\|u_{i+1}^t - u_i^t\|_2^2\Big] + \frac{1}{2}\mathbb{E}\Big[\|u_i^t - u_{i-1}^t\|_2^2\Big] \\
&\quad + 2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_{i+1}^t) - F_{\sigma_i^t}(u_i^t), u_{i+1}^t - u^\star \Big\rangle\Big] \\
&\quad - 2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_{i-1}^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t), u_i^t - u^\star \Big\rangle\Big] \\
&\quad + 6C_1 \cdot \left(\eta^t \delta^t + (\eta^t)^2 \delta^t + (\eta^t)^2 + \frac{(\eta^t)^2}{\delta^t} + \frac{(\eta^t)^2}{(\delta^t)^2}\right).
\end{aligned} \tag{I.27}$$

Meanwhile, Lemma I.1.5, Proposition I.1.4 (Kantorovich Duality), and Lemma I.1.6 imply that:

$$\begin{aligned}
\Big|\mathbb{E}[\Delta(u_{i+1}^t)] - \mathbb{E}\Big[L_{\sigma_i^t}(x_{i+1}^t, y^\star) - L_{\sigma_i^t}(x^\star, y_{i+1}^t)\Big]\Big| &\leqslant \frac{G}{n} \sum_{r=1}^{n} \mathcal{W}_2\Big(\mathcal{D}_{i+1,t}, \mathcal{D}_{i+1,t}^r\Big) \\
&\leqslant \frac{G}{n} \sum_{r=1}^{n} \sqrt{\mathbb{E}\Big[\Big\|u_{i+1}^t(\sigma^t) - u_{i+1}^t(\tilde{\sigma}^t)\Big\|_2^2\Big]} \\
&\leqslant G \cdot \left((6nd + 14n + 4)G \cdot \eta^t + 6ndM_L \cdot \frac{\eta^t}{\delta^t}\right).
\end{aligned}$$

Substituting back into (I.27), we have:

$$
\begin{aligned}
&2\eta^t \cdot \mathbb{E}\Big[\Delta(u_i^t)\Big] \\
\leqslant &2\eta^t \cdot \mathbb{E}\Big[L_{\sigma_i^t}(x_{i+1}^t, y^\star) - L_{\sigma_i^t}(x^\star, y_{i+1}^t)\Big] \\
&\qquad + G \cdot \left((12nd + 28n + 8)G \cdot (\eta^t)^2 + 12ndM_L \cdot \frac{(\eta^t)^2}{\delta^t}\right) \\
\leqslant &\mathbb{E}\Big[\|u_i^t - u^\star\|_2^2\Big] - \mathbb{E}\Big[\|u_{i+1}^t - u^\star\|_2^2\Big] - \frac{1}{2}\mathbb{E}\Big[\|u_{i+1}^t - u_i^t\|_2^2\Big] + \frac{1}{2}\mathbb{E}\Big[\|u_i^t - u_{i-1}^t\|_2^2\Big] \\
&\qquad + 2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_{i+1}^t) - F_{\sigma_i^t}(u_i^t), u_{i+1}^t - u\Big\rangle\Big] \\
&\qquad - 2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_{i-1}^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t), u_i^t - u\Big\rangle\Big] \\
&\qquad + 6C_1 \cdot \left(\eta^t\delta^t + (\eta^t)^2\delta^t + (\eta^t)^2 + \frac{(\eta^t)^2}{\delta^t} + \frac{(\eta^t)^2}{(\delta^t)^2}\right) \\
&\qquad + G \cdot \left((12nd + 28n + 8)G \cdot (\eta^t)^2 + 12ndM_L \cdot \frac{(\eta^t)^2}{\delta^t}\right). \qquad \text{(I.28)}
\end{aligned}
$$

We can now sum the above telescoping terms across the t-th epoch, as shown below:

$$
\begin{aligned}
&2 \cdot \sum_{i=1}^{n} \eta^t \cdot \mathbb{E}\Big[\Delta(u_i^t)\Big] \\
\leqslant &\mathbb{E}\Big[\|u_1^t - u^\star\|_2^2\Big] - \mathbb{E}\Big[\|u_1^{t+1} - u^\star\|_2^2\Big] + \frac{1}{2}\mathbb{E}\Big[\|u_1^t - u_0^t\|_2^2\Big] - \frac{1}{2}\mathbb{E}\Big[\|u_1^{t+1} - u_0^{t+1}\|_2^2\Big] \\
&\qquad + 2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_0^t}(u_1^t) - F_{\sigma_0^t}(u_0^t), u_1^t - u^\star\Big\rangle\Big] \\
&\qquad - 2\eta^t \cdot \mathbb{E}\Big[\Big\langle F_{\sigma_0^{t+1}}(u_1^{t+1}) - F_{\sigma_0^{t+1}}(u_0^{t+1}), u_1^{t+1} - u^\star\Big\rangle\Big] \\
&\qquad + 6nC_1 \cdot \left(\eta^t\delta^t + (\eta^t)^2\delta^t + (\eta^t)^2 + \frac{(\eta^t)^2}{\delta^t} + \frac{(\eta^t)^2}{(\delta^t)^2}\right) \\
&\qquad + nG \cdot \left((12nd + 28n + 8)G \cdot (\eta^t)^2 + 12ndM_L \cdot \frac{(\eta^t)^2}{\delta^t}\right).
\end{aligned}
$$

Meanwhile, we have for each $t = 0, 1, \cdots, T-1$, $i \in [n]$:

$$\begin{aligned}
&\mathbb{E}\Big[\Big\langle F_{\sigma_i^t}(u_{i+1}^t) - F_{\sigma_i^t}(u_i^t), u_{i+1}^t - u^\star \Big\rangle\Big] \\
\leqslant &\mathbb{E}\Big[\Big\|F_{\sigma_i^t}(u_{i+1}^t) - F_{\sigma_i^t}(u_i^t)\Big\| \cdot \Big\|u_{i+1}^t - u^\star\Big\|\Big] \\
= &\ell \cdot \mathbb{E}\Big[\|u_{i+1}^t - u_i^t\|\Big] \cdot D \\
\leqslant &\ell D \cdot \mathbb{E}\Big[\Big\| - \eta^t \hat{F}_{\sigma_i^t}(u_i^t; \delta^t, v_i^t) - \eta^t \hat{F}_{\sigma_{i-1}^t}(u_i^t; \delta^t, v_i^t) + \eta^t \hat{F}_{\sigma_{i-1}^t}(u_{i-1}^t; \delta^t, v_{i-1}^t)\Big\|\Big] \\
\leqslant &3\ell D \cdot \eta^t \cdot \left(dG + \frac{dM_L}{\delta^t}\right) \\
= &3\ell D d G \cdot \eta^t + 3\ell D d M_L \cdot \frac{\eta^t}{\delta^t},
\end{aligned}$$

where the final inequality follows from Proposition I.1.3, (I.3). We can upper bound

$$\mathbb{E}\Big[\Big\langle F_{\sigma_{i-1}^t}(u_i^t) - F_{\sigma_{i-1}^t}(u_{i-1}^t), u_i^t - u \Big\rangle\Big]$$

in a similar fashion. Substituting back into (I.28), we have:

$$\begin{aligned}
&2 \cdot \sum_{i=1}^n \eta^t \cdot \mathbb{E}\Big[\Delta(u_i^t)\Big] \\
\leqslant &\mathbb{E}\Big[\|u_1^t - u^\star\|_2^2\Big] - \mathbb{E}\Big[\|u_1^{t+1} - u^\star\|_2^2\Big] + \frac{1}{2}\mathbb{E}\Big[\|u_1^t - u_0^t\|_2^2\Big] - \frac{1}{2}\mathbb{E}\Big[\|u_1^{t+1} - u_0^{t+1}\|_2^2\Big] \\
&\quad + 6nC_1 \cdot \left(\eta^t\delta^t + (\eta^t)^2\delta^t + (\eta^t)^2 + \frac{(\eta^t)^2}{\delta^t} + \frac{(\eta^t)^2}{(\delta^t)^2}\right) \\
&\quad + nG \cdot \left((12nd + 28n + 8)G \cdot (\eta^t)^2 + 12ndM_L \cdot \frac{(\eta^t)^2}{\delta^t}\right) \\
&\quad + 6\ell D d G \cdot (\eta^t)^2 + 6\ell D d M_L \cdot \frac{(\eta^t)^2}{\delta^t} \\
\leqslant &\mathbb{E}\Big[\|u_1^t - u^\star\|_2^2\Big] - \mathbb{E}\Big[\|u_1^{t+1} - u^\star\|_2^2\Big] + \frac{1}{2}\mathbb{E}\Big[\|u_1^t - u_0^t\|_2^2\Big] - \frac{1}{2}\mathbb{E}\Big[\|u_1^{t+1} - u_0^{t+1}\|_2^2\Big] \qquad \text{(I.29)} \\
&\quad + 2C \cdot \left(\eta^t\delta^t + (\eta^t)^2\delta^t + (\eta^t)^2 + \frac{(\eta^t)^2}{\delta^t} + \frac{(\eta^t)^2}{(\delta^t)^2}\right),
\end{aligned}$$

where $C := \max\{3nC_1, (6nd + 14n + 4)nG, 6ndM_L, 3\ell DdG, 3\ell DdM_L\}$.

Finally, summing the above telescoping terms over $i \in [n]$ and $t \in \{0, 1, \cdots, T-1\}$, and

removing non-positive terms, we obtain:

$$\begin{aligned}
&\frac{\sum_{t=0}^{T-1}\sum_{i=1}^{n}\eta^t\cdot\mathbb{E}\left[\Delta(u_i^t)\right]}{\sum_{t=0}^{T-1}\sum_{i=1}^{n}\eta^t}\\
\leqslant&\frac{1}{2\cdot\sum_{t=0}^{T-1}\sum_{i=1}^{n}\eta^t}\left(\|u_0^0-u^\star\|_2-\mathbb{E}\left[\|u_n^{T-1}-u^\star\|_2\right]+\frac{1}{2}\|u_1^0-u_0^0\|_2-\frac{1}{2}\mathbb{E}\left[\|u_n^{T-1}-u_{n-1}^{T-1}\|_2\right]\right)\\
&\qquad+C\cdot\frac{1}{\sum_{t=0}^{T-1}\sum_{i=1}^{n}\eta^t}\cdot\sum_{t=0}^{T-1}\left(\eta^t\delta^t+(\eta^t)^2\delta^t+(\eta^t)^2+\frac{(\eta^t)^2}{\delta^t}+\frac{(\eta^t)^2}{(\delta^t)^2}\right)\\
\leqslant&\frac{1}{\sum_{t=0}^{T-1}\eta^t}\cdot\frac{3D}{4n}+C\cdot\frac{1}{n\sum_{t=0}^{T-1}\eta^t}\cdot\sum_{t=0}^{T-1}\left(\eta^t\delta^t+(\eta^t)^2\delta^t+(\eta^t)^2+\frac{(\eta^t)^2}{\delta^t}+\frac{(\eta^t)^2}{(\delta^t)^2}\right),
\end{aligned}\tag{I.30}$$

By definition, $\eta^t=\eta^0\cdot(t+1)^{-3/4-\chi}$ and $\delta^t=\delta^0\cdot(t+1)^{-1/4}$, so by Lemma I.1.8, we have:

$$\begin{aligned}
\sum_{t=0}^{T-1}\eta^t&=\eta^0\cdot\sum_{t=1}^{T}t^{-3/4-\chi}\geqslant 4\eta^0\cdot T^{1/4-\chi},\\
\sum_{t=0}^{T-1}\eta^t\delta^t&=\eta^0\delta^0\cdot\sum_{t=1}^{T}t^{-(1+\chi)}\leqslant\eta^0\delta^0\cdot\left(1+\frac{1}{\chi}\right),\\
\sum_{t=0}^{T-1}(\eta^t)^2&=(\eta^0)^2\cdot\sum_{t=1}^{T}t^{-3/2-2\chi}\leqslant(\eta^0)^2\cdot\left(1+\frac{1}{\frac{1}{2}+2\chi}\right)\leqslant 3\cdot(\eta^0)^2,\\
\sum_{t=0}^{T-1}(\eta^t)^2\delta^t&=(\eta^0)^2\delta^0\cdot\sum_{t=1}^{T}t^{-7/4-2\chi}\leqslant(\eta^0)^2\delta^0\cdot\left(1+\frac{1}{\frac{3}{4}+2\chi}\right)\leqslant\frac{7}{4}\cdot(\eta^0)^2\epsilon^0,\\
\sum_{t=0}^{T-1}\frac{(\eta^t)^2}{\delta^t}&=\frac{(\eta^0)^2}{\delta^0}\cdot\sum_{t=1}^{T}t^{-5/4-2\chi}\leqslant\frac{(\eta^0)^2}{\delta^0}\cdot\left(1+\frac{1}{\frac{1}{4}+2\chi}\right)\leqslant 5\cdot\frac{(\eta^0)^2}{\epsilon^0},\\
\sum_{t=0}^{T-1}\frac{(\eta^t)^2}{(\delta^t)^2}&=\frac{(\eta^0)^2}{(\delta^0)^2}\cdot\sum_{t=1}^{T}t^{-1-2\chi}\leqslant\frac{(\eta^0)^2}{(\delta^0)^2}\cdot\left(1+\frac{1}{2\chi}\right).
\end{aligned}$$

Substituting back into (I.30) and using the convexity of the gap function $\Delta(\cdot)$, we have:

$$\begin{aligned}
&\mathbb{E}\Big[\Delta(u^T)\Big] \\
&\leqslant \frac{\sum_{t=0}^{T-1}\sum_{i=1}^{n}\eta^t\cdot\mathbb{E}\Big[\Delta(u_i^t)\Big]}{\sum_{t=0}^{T-1}\sum_{i=1}^{n}\eta^t} \\
&\leqslant \frac{1}{\sum_{t=0}^{T-1}\eta^t}\cdot\frac{3}{4n}D + C\cdot\frac{1}{\sum_{t=0}^{T-1}\eta^t}\cdot\sum_{t=0}^{T-1}\left(\eta^t\delta^t + (\eta^t)^2\delta^t + (\eta^t)^2 + \frac{(\eta^t)^2}{\delta^t} + \frac{(\eta^t)^2}{(\delta^t)^2}\right) \\
&\leqslant \left(\frac{3}{16n}D + \frac{47}{4n}\cdot C\max\left\{\delta^0, \eta^0, \eta^0\delta^0, \frac{\eta^0}{\delta^0}, \frac{\eta^0}{(\delta^0)^2}\right\}\left(1+\frac{1}{\chi}\right)\right)T^{-1/4+\chi} \\
&\leqslant \delta.
\end{aligned}$$

where the final inequality follows by definition of T.

I.3 Additional Details on the Experimental Study

Algorithms

In our experiments, we compare the OGDA-RR algorithm (Algorithm 9) with three other zeroth-order algorithms—Optimistic Gradient Descent Ascent with Sampling with Replacement (OGDA-WR), Stochastic Gradient Descent Ascent with Random Reshuffling (SGDA-RR), and Stochastic Gradient Descent Ascent with Sampling with Replacement (SGDA-WR)—characterized by the update equations (I.32), (I.33), (I.34), respectively. For convenience, we have reproduced (10.13), the update equation for the OGDA-RR algorithm (Algorithm 9), as (I.31) below:

$$u_{i+1}^t = \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}\Big(u_i^t - \eta^t\hat{F}_{\sigma_i^t}(u_i^t;\delta^t,v_i^t) - \eta^t\hat{F}_{\sigma_{i-1}^t}(u_i^t;\delta^t,v_i^t) + \eta^t\hat{F}_{\sigma_{i-1}^t}(u_{i-1}^t;\delta^t,v_{i-1}^t)\Big), \quad \text{(I.31)}$$

$$u_{i+1}^t = \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}\Big(u_i^t - \eta^t\hat{F}_{j_i^t}(u_i^t;\delta^t,v_i^t) - \eta^t\hat{F}_{j_{i-1}^t}(u_i^t;\delta^t,v_i^t) + \eta^t\hat{F}_{j_{i-1}^t}(u_{i-1}^t;\delta^t,v_{i-1}^t)\Big), \quad \text{(I.32)}$$

$$u_{i+1}^t = \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}\Big(u_i^t - \eta^t\hat{F}_{\sigma_i^t}(u_i^t;\delta^t,v_i^t)\Big), \quad \text{(I.33)}$$

$$u_{i+1}^t = \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}\Big(u_i^t - \eta^t\hat{F}_{j_i^t}(u_i^t;\delta^t,v_i^t)\Big), \quad \text{(I.34)}$$

where the indices σ_i^t and j_i^t are as defined in Algorithms 17, 18, and 19.

Additional Experimental Results

In this section, we present more experimental findings, on both synthetic and real-world datasets, that reinforces the utility of the proposed algorithm. In all experimental results throughout this subsection, we take $\delta = 0.4$, $\kappa = 0.5$ and $\zeta = 0.05$.

Algorithm 17 OGDA-WR Algorithm

Require: Stepsizes η^t, δ^t, data points $\{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}$, initial value $u_0^{(0)}$, time horizon T

for $t = 0$ **to** $T-1$ **do**
  for $i = 0$ **to** $n-1$ **do**
    Sample $j_i^t \sim \mathsf{Unif}(\{1, \cdots, n\})$
    Sample $v_i^t \sim \mathsf{Unif}(\mathcal{S}^{d-1})$
    $u_{i+1}^t \leftarrow$ update from equation (I.32)
  end for
  $u_0^{(t+1)} \leftarrow u_n^t$
  $u_{-1}^{(t+1)} \leftarrow u_{n-1}^t$
end for

Ensure: $\tilde{u}^T := \frac{1}{n \cdot \sum_{t=0}^{T-1} \eta^t} \sum_{t=0}^{T-1} \sum_{i=1}^n \eta^t u_i^t$

Algorithm 18 SGDA-RR Algorithm

Require: Stepsizes η^t, δ^t, data points $\{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}$, initial value $u_0^{(0)}$, time horizon T

for $t = 0$ **to** $T-1$ **do**
  for $i = 0$ **to** $n-1$ **do**
    Sample $j_i^t \sim \mathsf{Unif}(\{1, \cdots, n\})$
    Sample $v_i^t \sim \mathsf{Unif}(\mathcal{S}^{d-1})$
    $u_{i+1}^t \leftarrow$ update from equation (I.33)
  end for
  $u_0^{(t+1)} \leftarrow u_n^t$
  $u_{-1}^{(t+1)} \leftarrow u_{n-1}^t$
end for

Ensure: $\tilde{u}^T := \frac{1}{n \cdot \sum_{t=0}^{T-1} \eta^t} \sum_{t=0}^{T-1} \sum_{i=1}^n \eta^t u_i^t$

Experimental Study On Synthetic Datasets

Figure I.1 compares the performance of 1-4 on a synthetic dataset (whose generating process is the same as that described in Section 10.5), with 4000 training points and 800 test points. Our proposed algorithm performs better empirically compared to most of its counterparts. Moreover, the proposed classifier, 1, is significantly more robust than a classifier obtained without considering adversarial perturbations. Note, however, that we cannot make any conclusive claims yet, because of the inherent randomness in these algorithms. Indeed, even if we fix the initialization, then there are two sources of randomness—the construction of the zeroth-order gradient estimator, and the sampling process that generates the data points.

To illustrate the variability in these algorithms' performance, we run each algorithm repeatedly on a data set with 500 synthetically generated data points, using the same initialization, and present confidence interval plots with ± 2 standard deviations for the resulting

Algorithm 19 SGDA-WR Algorithm

Require: Stepsizes η^t, δ^t, data points $\{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}$, initial value $u_0^{(0)}$, time horizon T

1: **for** $t = 0$ **to** $T-1$ **do**
2: $\sigma^t = (\sigma_1^t, \cdots, \sigma_n^t) \leftarrow$ a random permutation of the set $[n]$
3: **for** $i = 0$ **to** $n-1$ **do**
4: Sample $v_i^t \sim \mathsf{Unif}(\mathcal{S}^{d-1})$
5: $u_{i+1}^t \leftarrow$ update from equation (I.34)
6: **end for**
7: $u_0^{(t+1)} \leftarrow u_n^t$
8: $u_{-1}^{(t+1)} \leftarrow u_{n-1}^t$
9: **end for**

Ensure: $\tilde{u}^T := \frac{1}{n \cdot \sum_{t=0}^{T-1} \eta^t} \sum_{t=0}^{T-1} \sum_{i=1}^n \eta^t u_i^t$

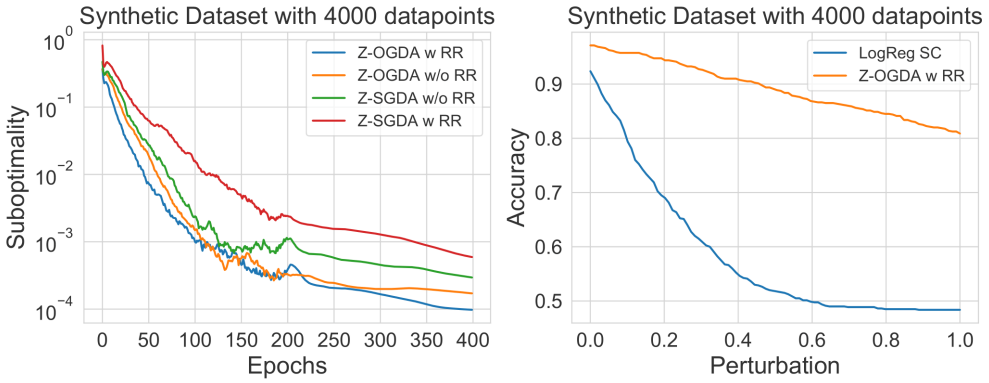


Figure I.1: Experimental results for a synthetic dataset with $n = 4000$. (Left pane)) Suboptimality iterates generated by the four algorithms 1, 2, 3, 4, respectively denoted as *Z-OGDA w RR*, *Z-OGDA w/o RR*, *Z-SGDA w RR*, *Z-SGDA w/o RR*. (Right pane) Comparison between decay in accuracy of strategic classification with logistic regression (trained with $\zeta = 0.05$) and Algorithm 1 with changes in perturbation.

performance (Figure I.2). On average, our proposed algorithm 1 outperforms the other algorithms 2-4. It is also interesting to point out that the performance of algorithms with random reshuffling is generally higher, and fluctuate less, compared to the performance of algorithms without random reshuffling.

We now illustrate the performance of our algorithm on two real-world data sets—the *GiveMeSomeCredit* dataset [1], and the *Porto Bank* data set[2].

[1]This dataset can be found at `https://www.kaggle.com/c/GiveMeSomeCredit`
[2]This dataset can be found at `https://archive.ics.uci.edu/ml/datasets/bank+marketing`

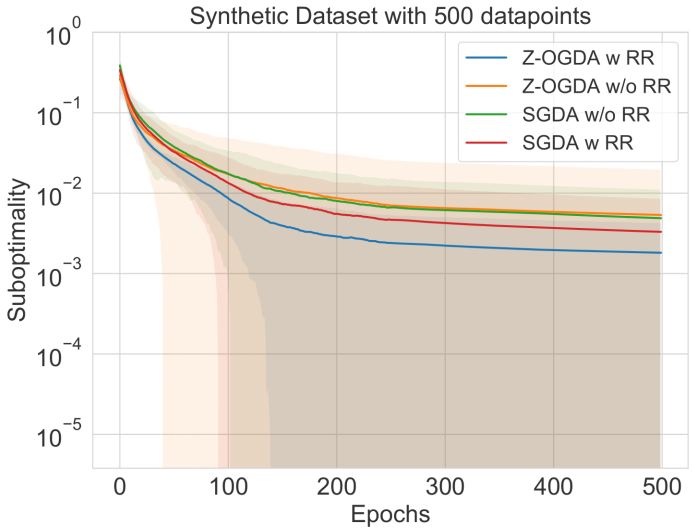


Figure I.2: Experimental results for a synthetic dataset with $n = 500$. Suboptimality iterates generated by the four algorithms 1, 2, 3, and 4 are respectively denoted as *Z-OGDA w RR*, *Z-OGDA w/o RR*, *Z-SGDA w RR*, and *Z-SGDA w/o RR*.

Experimental Study on Credit Dataset

In modern times, banks use machine learning to determine whether or not to finance a customer. This process can be encoded into a classification framework, by using features such as age, debt ratio, monthly income to classify a customer as either likely or unlikely to default. However, those algorithms generally do not account for strategic or adversarial behavior on the part of the agents.

To illustrate the effect of our algorithm on datasets of practical significance, we deploy our algorithms on the *GiveMeSomeCredit*(GMSC) dataset, while assuming that the underlying features are subject to strategic or adversarial perturbations. We use a subset of the dataset of size 2000 with balanced labels. In Figure I.3, we compare the empirical performance of our algorithm 1 with that of 2-4. The left pane shows that 1 performs well, and the right pane illustrates that our classifier is significantly more robust to adversarial perturbations in data, compared to the strategic classification-based logistic regression algorithm developed recently in the literature [124].

Experimental Study on Porto-Bank Dataset

Next, we present empirical results obtained by applying our algorithm to the *Porto-Bank* dataset, which describes marketing campaigns of term deposits at Portuguese financial institutions. The classification task in this scenario aims to predict whether a customer with given features (eg. age, job, marital status etc.) would enroll for term deposits.

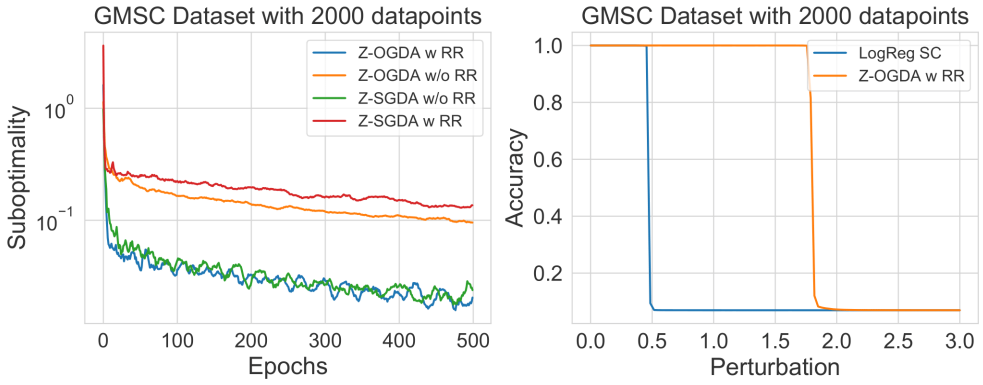


Figure I.3: Experimental results for a balanced GiveMeSomeCredit dataset with $n = 2000$. (Left pane) Suboptimality iterates generated by the four algorithms 1, 2, 3, 4, respectively denoted as *Z-OGDA w RR*, *Z-OGDA w/o RR*, *Z-SGDA w RR*, *Z-SGDA w/o RR*. (Right pane) Comparison between decay in accuracy of strategic classification with logistic regression (originally trained with $\zeta = 0.05$) and Algorithm 1 with changes in perturbation.

In Figure I.4, we present the performance of our proposed algorithm 1 on the Porto-Bank dataset. For ease of illustration, we consider a subset of the dataset with 2000 training data points, 800 test data points, and balanced labels. In Figure I.4, we compare the empirical performance of our Algorithm 1 with that of Algorithms 2-4. The left pane shows that Algorithm 1 performs well, while the right pane illustrates that our classifier is significantly more robust to adversarial perturbations in data, compared to the strategic classification-based logistic regression developed recently in the literature [124].

Effect of n, d on sample complexity

In this part, we demonstrate the empirical results that corroborates the theoretical dependence of sample complexity on n, d. For this purpose, we use synthetic dataset which is generated as per the method described in Section 6.1. Here we work in the setting where $n \in \{500, 1000, 1500, 2000\}$ and $d \in \{10, 15, 20, 25\}$. We fix the suboptimality to $\epsilon = 0.1$ and compute the number of samples required in each of the settings of n and d so that the iterates reach the $\epsilon-$suboptimality. We present the results in Figure I.5.

I.4 Logistic regression as a Generalized linear model

The goal in logistic regression is to maximize the log-likelihood of the conditional probability of y (the *label*) given x (the *feature*). In this model, it is assumed that:

$$P(Y = 1|x, \theta) = \frac{1}{1 + \exp(-\langle x, \theta \rangle)}.$$

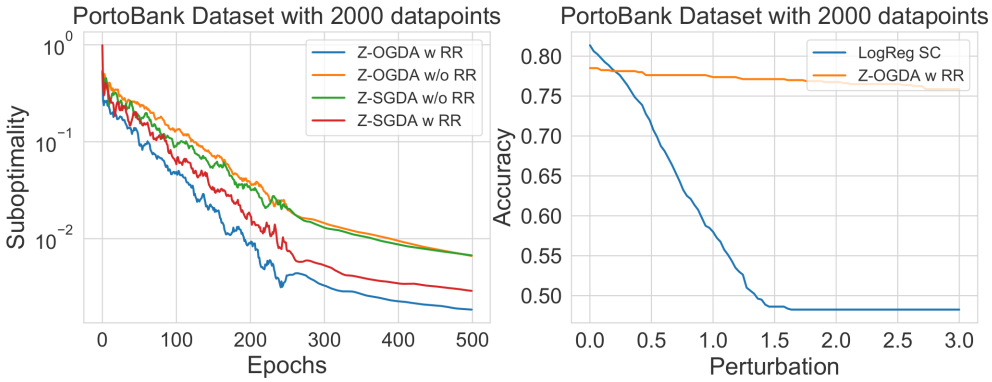


Figure I.4: Experimental results for a balanced PortoBank dataset with $n = 2000$. (Left pane) Suboptimality iterates generated by the four algorithms 1, 2, 3, 4, respectively denoted as *Z-OGDA w RR*, *Z-OGDA w/o RR*, *Z-SGDA w RR*, *Z-SGDA w/o RR*. (Right pane) Comparison between decay in accuracy of strategic classification with logistic regression (originally trained with $\zeta = 0.05$) and Algorithm 1 with change in perturbation.

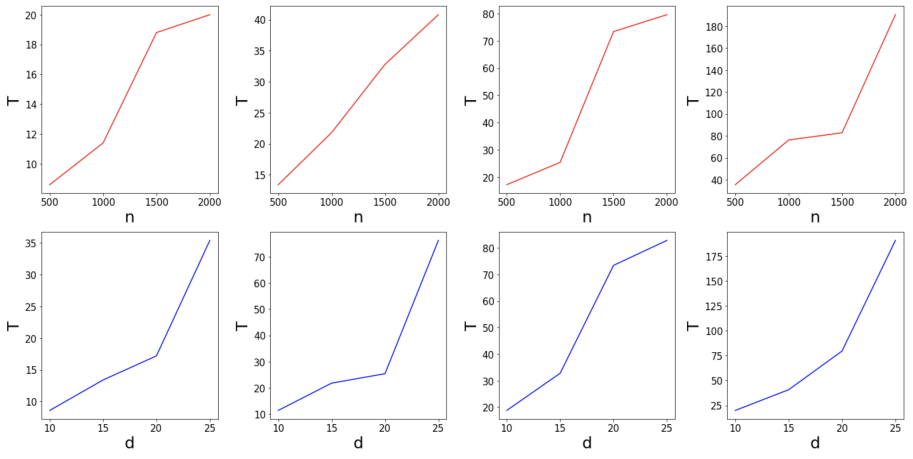


Figure I.5: Experimental results presenting the number of samples required to reach $\epsilon-$ suboptimality, with $\epsilon = 0.1$, for our algorithm 1 on synthetic dataset with varying values of $n \in \{500, 1000, 1500, 2000\}$ and $d \in \{10, 15, 20, 25\}$.

This implies that:

$$P(Y = -1|x, \theta) = \frac{\exp(-\langle x, \theta \rangle)}{1 + \exp(-\langle x, \theta \rangle)}.$$

Given a data point (x, y) the logistic loss is log-likelihood of observing y given x. For any θ and $y \in \{-1, 1\}$:

$$P(Y = y|x; \theta) = (P(Y = 1|x, \theta))^{\frac{1+y}{2}} (P(Y = -1|x, \theta))^{\frac{1-y}{2}}.$$

Now, the log-likelihood is given by:

$$\begin{aligned}
L(x, y; \theta) &= \log(P(Y = y|x; \theta)) \\
&= \frac{1+y}{2} \log\left(\frac{1}{1 + \exp(-\langle x, \theta \rangle)}\right) + \frac{1-y}{2} \log\left(\frac{\exp(-\langle x, \theta \rangle)}{1 + \exp(-\langle x, \theta \rangle)}\right) \\
&= -\frac{1-y}{2} \langle x, \theta \rangle + \left(\frac{1+y}{2} + \frac{1-y}{2}\right) \log\left(\frac{1}{1 + \exp(-\langle x, \theta \rangle)}\right) \\
&= -\frac{1-y}{2} \langle x, \theta \rangle - \log(1 + \exp(-\langle x, \theta \rangle)) \\
&= \frac{y}{2} \langle x, \theta \rangle - \frac{1}{2} \langle x, \theta \rangle + \langle x, \theta \rangle - \log(1 + \exp(\langle x, \theta \rangle)) \\
&= \frac{y}{2} \langle x, \theta \rangle + \frac{1}{2} \langle x, \theta \rangle - \log(1 + \exp(\langle x, \theta \rangle)).
\end{aligned}$$

The goal is to maximize the log likelihood, which is equivalent to minimizing the negative log likelihood. Thus the logistic regression minimizes the following loss:

$$\tilde{L}(x, y; \theta) = -L(x, y; \theta) = -\frac{y}{2} \langle x, \theta \rangle + \phi(\langle x, \theta \rangle),$$

where $\phi(\beta) = \log(1 + \exp(\beta)) - \frac{\beta}{2}$. If $y = 1$, then the above loss becomes:

$$\log(1 + \exp(\langle x, \theta \rangle)) - \langle x, \theta \rangle = \log(1 + \exp(-\langle x, \theta \rangle)).$$

Otherwise, if $y = -1$, then the above loss becomes $\log(1 + \exp(\langle x, \theta \rangle))$. Thus, the above loss is equivalent to $\log(1 + \exp(-y \langle x, \theta \rangle))$.

Appendix J

Appendix for Chapter 11

J.1 Technical Results

Properties of zeroth-order gradient estimator. In this section, we state relevant properties of the zeroth-order gradient estimator used in the proposed algorithm. Define the δ-smoothed loss function $\tilde{f}_\delta : \mathbb{R}^d \to \mathbb{R}$ by $\tilde{f}_\delta(x) := \mathbb{E}_{\overline{v}\sim\mathsf{Unif}(B^d)}[\tilde{f}(x+\delta\overline{v})]$, where $\mathcal{B}(\mathbb{R}^d)$ denotes the d-dimensional unit open ball in $\mathbb{R}^d$. We further define $\delta\cdot\mathcal{S}(\mathbb{R}^d) := \{\delta v : v \in \mathcal{S}(\mathbb{R}^d)\}$ and $\delta\cdot\mathcal{B}(\mathbb{R}^d) := \{\delta\overline{v} : \overline{v} \in \mathcal{B}(\mathbb{R}^d)\}$. Finally, we use $\mathsf{vol}_d(\cdot)$ to denote the volume of a set in d dimensions.

Lemma J.1.1 ([395, 140, 268])**.** *Let $\tilde{F}(x;\delta,v) = \frac{d}{\delta}\cdot\Big(\tilde{f}(\hat{x}) - \tilde{f}(x)\Big)\,v$ where $\hat{x} = x+\delta v$ and $F(x) = \nabla\tilde{f}_\delta(x)$. Then the following holds:*

$$\mathbb{E}_{v\sim\mathit{Unif}(\mathcal{S}(\mathbb{R}^d))}\Big[\tilde{F}(x;\delta,v)\Big] = \nabla\tilde{f}_\delta(x). \tag{J.1}$$

Proof. Observe that $\tilde{f}_\delta(x) = \mathbb{E}_{v\sim\mathsf{Unif}(B^d)}[\tilde{f}(\hat{x})]$ where $\hat{x} = x+\delta v$ and $\tilde{F}(x;\delta,v) = \frac{d}{\delta}\cdot\Big(\tilde{f}(x+\delta v) - \tilde{f}(x)\Big)\,v$ for each $x\in\mathbb{R}^d$, $\delta>0$, and $v\in\mathcal{S}(\mathbb{R}^d)$. We compute

$$\begin{aligned}
\nabla\tilde{f}_\delta(x) &= \nabla\mathbb{E}_{\overline{v}\sim\mathsf{Unif}(\mathcal{B}(\mathbb{R}^d))}\Big[\tilde{f}(x+\delta\overline{v})\Big] \\
&= \nabla\mathbb{E}_{\overline{v}\sim\mathsf{Unif}(\delta\cdot\mathcal{B}(\mathbb{R}^d))}\Big[\tilde{f}(x+\overline{v})\Big] \\
&= \frac{1}{\mathsf{vol}_d(\delta\cdot\mathcal{B}(\mathbb{R}^d))}\cdot\nabla\left(\int_{\delta\cdot B^d}\tilde{f}(x+\overline{v})\,d\overline{v}\right) \\
&= \frac{1}{\mathsf{vol}_d(\delta\cdot\mathcal{B}(\mathbb{R}^d))}\cdot\int_{\delta\cdot\mathcal{S}^{d-1}}\tilde{f}(x+v)\cdot\frac{v}{\|v\|_2}\,dv, \qquad \text{(J.2)}
\end{aligned}$$

where (J.2) follows by Stokes' Theorem [236] which implies that

$$\nabla\int_{\delta\cdot\mathcal{B}(\mathbb{R}^d)}\tilde{f}(x+\overline{v})\,d\overline{v} = \int_{\delta\cdot\mathcal{S}(\mathbb{R}^d)}\tilde{f}(x+v)\cdot\frac{v}{\|v\|_2}\,dv.$$

Next note that

$$\begin{aligned}\mathbb{E}_{v\sim\mathsf{Unif}(\mathcal{S}^{d-1})}\left[\tilde{F}(x;\delta,v)\right] &= \frac{d}{\delta}\cdot\mathbb{E}_{v\sim\mathsf{Unif}(\mathcal{S}(\mathbb{R}^d))}\left[\left(\tilde{f}(x+\delta v)-\tilde{f}(x)\right)v\right]\\ &= \frac{d}{\delta}\cdot\mathbb{E}_{v'\sim\mathsf{Unif}(\delta\cdot\mathcal{S}(\mathbb{R}^d))}\left[\tilde{f}(x+v')\cdot\frac{v'}{\|v'\|_2}\right]\\ &= \frac{d}{\delta}\cdot\frac{1}{\mathsf{vol}_{d-1}(\delta\cdot\mathcal{S}(\mathbb{R}^d))}\cdot\int_{\delta\cdot\mathcal{S}(\mathbb{R}^d)}\tilde{f}(x+v)\cdot\frac{v}{\|v\|_2}\,dv,\end{aligned}$$

The equality (J.1) now follows by observing that

$$\frac{\mathsf{vol}_{d-1}(\delta\cdot\mathcal{S}(\mathbb{R}^d))}{\mathsf{vol}_d(\delta\cdot\mathcal{B}(\mathbb{R}^d))}=\frac{d}{\delta}.$$

That is, surface-area-to-volume ratio of sphere in $\mathbb{R}^d$ or radius δ is d/δ. This concludes the proof.

Discrete-Gronwall Inequality.

Lemma J.1.2 ([60])**.** *For a non-negative sequence (s_n) such that*

$$s_{n+1}\leqslant C+L\left(\sum_{m=0}^{n}s_m\right), \tag{J.3}$$

the following inequality holds: $s_{n+1}\leqslant\tilde{C}\exp(Ln)$ where $\tilde{C}=\max\{s_0,C\}$.

Proof. The proof is taken from [60]. We provide details here for the sake of completeness. Define $S_n=\sum_{m=0}^{n}s_m$. Then (J.3) can be equivalently written as

$$S_{n+1}-S_n\leqslant\tilde{C}+LS_n,$$

where $\tilde{C}=\max\{s_0,C\}$. Thus, it follows that

$$\begin{aligned}S_{n+1} &\leqslant (1+L)S_n+\tilde{C}\\ &\leqslant (1+L)^{n+1}S_0+\sum_{k=0}^{n}\tilde{C}(1+L)^{n-k}\\ &\leqslant \tilde{C}\sum_{k=0}^{n+1}(1+L)^{n+1-k}\\ &\leqslant \tilde{C}\sum_{k=0}^{n+1}\exp(L(n+1-k))\\ &\leqslant \tilde{C}\int_0^{n+1}\exp(L(n+1-\tau))d\tau\\ &= \frac{\tilde{C}}{L}\left(\exp(L(n+1))-1\right).\end{aligned}$$

Substituting this in (J.3) we obtain

$$x_{n+1} \leqslant \tilde{C} + LS_n \leqslant \tilde{C} + \tilde{C}\left(\exp(L(n+1)) - 1\right) = \tilde{C}\exp(L(n+1)).$$

This concludes the proof.

J.2 Proof of Lemmas

Lemma 1 (Restated). *If $\bar{\eta} \leqslant d/(2\tilde{\ell})$ then*

$$\begin{aligned}&\mathbb{E}\left[\tilde{f}(x_{t+1})\right] \\ &\leqslant \mathbb{E}\left[\tilde{f}(x_t)\right] - \frac{\eta_t}{2}\mathbb{E}\left[\|\nabla\tilde{f}(x_t)\|^2\right] + \eta_t\mathbb{E}\left[\|\mathcal{E}_t^{(1)}\|^2\right] + \eta_t\mathbb{E}\left[\|\mathcal{E}_t^{(3)}\|^2\right] + \tilde{\ell}\eta_t^2\mathbb{E}\left[\|\mathcal{E}_t^{(2)}\|^2\right].\end{aligned} \tag{J.4}$$

Proof. Note that by smoothness of $\tilde{f}$ function

$$\begin{aligned}\tilde{f}(x_{t+1}) &\leqslant \tilde{f}(x_t) + \left\langle\nabla\tilde{f}(x_t), x_{t+1} - x_t\right\rangle + \frac{\tilde{\ell}}{2}\|x_{t+1} - x_t\|^2 \\ &= \tilde{f}(x_t) - \eta_t\left\langle\nabla\tilde{f}(x_t), \hat{F}(x_t;\delta,v_t)\right\rangle + \frac{\tilde{\ell}\eta_t^2}{2}\|\hat{F}(x_t;\delta_t,v_t)\|^2 \\ &= \tilde{f}(x_t) - \eta_t\left\langle\nabla\tilde{f}(x_t), \nabla\tilde{f}(x_t) + \mathcal{E}_t^{(1)} + \mathcal{E}_t^{(2)} + \mathcal{E}_t^{(3)}\right\rangle + \frac{\tilde{\ell}\eta_t^2}{2}\|\nabla\tilde{f}(x_t) + \mathcal{E}_t^{(1)} + \mathcal{E}_t^{(2)} + \mathcal{E}_t^{(3)}\|^2,\end{aligned} \tag{J.5}$$

where we define

$$\begin{aligned}\mathcal{E}_t^{(1)} &:= \nabla\tilde{f}_\delta(x_t) - \nabla\tilde{f}(x_t) \\ \mathcal{E}_t^{(2)} &:= \frac{d}{\delta_t}(\tilde{f}(\hat{x}_t) - \tilde{f}(x_t))v_t - \nabla\tilde{f}_{\delta_t}(x_t) \\ \mathcal{E}_t^{(3)} &:= \frac{d}{\delta_t}\left((f(\hat{x}_t, y_t^{(K)})v_t - \tilde{f}(\hat{x}_t)v_t) - (f(x_t, \tilde{y}_t^{(K)})v_t - \tilde{f}(x_t)v_t)\right).\end{aligned}$$

Taking expectation on both sides of (J.5) we obtain

$$\begin{aligned}\mathbb{E}\left[\tilde{f}(x_{t+1})\right] &\leqslant \mathbb{E}\left[\tilde{f}(x_t)\right] - \eta_t\mathbb{E}\left[\left\langle\nabla\tilde{f}(x_t), \nabla\tilde{f}(x_t) + \mathcal{E}_t^{(1)} + \mathcal{E}_t^{(2)} + \mathcal{E}_t^{(3)}\right\rangle\right] \\ &\quad + \frac{\tilde{\ell}\eta_t^2}{2}\mathbb{E}\left[\|\nabla\tilde{f}(x_t) + \mathcal{E}_t^{(1)} + \mathcal{E}_t^{(2)} + \mathcal{E}_t^{(3)}\|^2\right] \\ &= \mathbb{E}\left[\tilde{f}(x_t)\right] - \eta_t\mathbb{E}\left[\left\langle\nabla\tilde{f}(x_t), \nabla\tilde{f}(x_t) + \mathcal{E}_t^{(1)} + \mathcal{E}_t^{(3)}\right\rangle\right] \\ &\quad + \frac{\tilde{\ell}\eta_t^2}{2}\mathbb{E}\left[\|\nabla\tilde{f}(x_t) + \mathcal{E}_t^{(1)} + \mathcal{E}_t^{(2)} + \mathcal{E}_t^{(3)}\|^2\right] \\ &\leqslant \mathbb{E}\left[\tilde{f}(x_t)\right] - \eta_t\mathbb{E}\left[\left\langle\nabla\tilde{f}(x_t), \nabla\tilde{f}(x_t) + \mathcal{E}_t^{(1)} + \mathcal{E}_t^{(3)}\right\rangle\right] \\ &\quad + \tilde{\ell}\eta_t^2\left(\mathbb{E}\left[\|\nabla\tilde{f}(x_t) + \mathcal{E}_t^{(1)} + \mathcal{E}_t^{(3)}\|^2\right] + \mathbb{E}\left[\|\mathcal{E}_t^{(2)}\|^2\right]\right),\end{aligned}$$

where the first equality and last inequality follows by noting that $\mathbb{E}\left[\mathcal{E}_t^{(2)}|x_t\right] = 0$ from Lemma 10. Next, choosing $\eta_t \leqslant \frac{1}{2\tilde{\ell}}$ we obtain

$$\begin{aligned}
&\mathbb{E}\left[\tilde{f}(x_{t+1})\right] \leqslant \mathbb{E}\left[\tilde{f}(x_t)\right] - \frac{\eta_t}{2}\Bigg(2\mathbb{E}\left[\left\langle \nabla\tilde{f}(x_t), \nabla\tilde{f}(x_t) + \mathcal{E}_t^{(1)} + \mathcal{E}_t^{(3)}\right\rangle\right] \\
&- \mathbb{E}\left[\|\nabla\tilde{f}(x_t) + \mathcal{E}_t^{(1)} + \mathcal{E}_t^{(3)}\|^2\right]\Bigg) + \tilde{\ell}\eta_t^2\left(\mathbb{E}\left[\|\mathcal{E}_t^{(2)}\|^2\right]\right) \\
&= \mathbb{E}\left[\tilde{f}(x_t)\right] - \frac{\eta_t}{2}\left(\mathbb{E}\left[\|\nabla\tilde{f}(x_t)\|^2\right] - \mathbb{E}\left[\|\mathcal{E}_t^{(1)} + \mathcal{E}_t^{(3)}\|^2\right]\right) + \tilde{\ell}\eta_t^2\mathbb{E}\left[\|\mathcal{E}_t^{(2)}\|^2\right],
\end{aligned}$$

where the equality follows by completing the squares.

Lemma 2 (Restated)**.** *The errors* $\mathbb{E}\left[\|\mathcal{E}_t^{(i)}\|^2\right]$ *for* $i \in \{1,2,3\}$ *are bounded as follows*

$$\begin{aligned}
&\mathbb{E}\left[\|\mathcal{E}_t^{(1)}\|^2\right] \leqslant \frac{\tilde{\ell}^2\delta_t^2 d^2}{4}, \quad \mathbb{E}\left[\|\mathcal{E}_t^{(2)}\|^2\right] \leqslant 4d^2\tilde{L}^2, \\
&\mathbb{E}\left[\|\mathcal{E}_t^{(3)}\|^2\right] \leqslant \frac{d^2}{\delta_t^2}L_2^2\left(2\alpha^t e_0 + 2C_6\sum_{k=0}^{t-1}\alpha^{t-k}\eta_k^2 + C_4\sum_{k=0}^{t-1}\alpha^{t-k}\delta_k^2\right).
\end{aligned} \tag{J.6}$$

Proof. We begin with bounding the terms $\mathbb{E}\left[\|\mathcal{E}_t^{(i)}\|^2\right]$ for $i \in \{1,2,3\}$. We note that

$$\begin{aligned}
&\mathbb{E}\left[\|\mathcal{E}_t^{(1)}\|^2\right] = \mathbb{E}\left[\|\nabla\tilde{f}_{\delta_t}(x_t) - \nabla\tilde{f}(x_t)\|^2\right] \\
&= \mathbb{E}\left[\left\|\mathbb{E}\left[\frac{d}{\delta_t}(\tilde{f}(\hat{x}_t) - \tilde{f}(x_t))v_t - \nabla\tilde{f}(x_t)\middle| x_t\right]\right\|^2\right] \\
&= \frac{d^2}{\delta_t^2}\mathbb{E}\left[\left\|\mathbb{E}\left[(\tilde{f}(\hat{x}_t) - \tilde{f}(x_t))v_t - \delta_t\mathbb{E}[v_t v_t^\top]\nabla\tilde{f}(x_t)\middle| x_t\right]\right\|^2\right] \\
&= \frac{d^2}{\delta_t^2}\mathbb{E}\left[\left\|\mathbb{E}\left[v_t\left(\tilde{f}(\hat{x}_t) - \tilde{f}(x_t) - \delta_t v_t^\top\nabla\tilde{f}(x_t)\right)\middle| x_t\right]\right\|^2\right] \\
&\leqslant \frac{d^2}{\delta_t^2}\mathbb{E}\left[\|\tilde{f}(\hat{x}_t) - \tilde{f}(x_t) - \delta_t v_t^\top\nabla\tilde{f}(x_t)\|^2\right] \\
&\leqslant \frac{d^2}{\delta_t^2}\frac{\tilde{\ell}^2\delta_t^4}{4} = \frac{\tilde{\ell}^2\delta_t^2 d^2}{4},
\end{aligned} \tag{J.7}$$

where the first inequality is due to Jensen's inequality and the last inequality is due to

$\tilde{\ell}-$smoothness of $\tilde{f}$. Next, we bound $\|\mathcal{E}_t^{(2)}\|$ as follows

$$\begin{aligned}\|\mathcal{E}_t^{(2)}\| &= \left\|\frac{d}{\delta_t}(\tilde{f}(\hat{x}_t)-\tilde{f}(x_t))v_t - \nabla\tilde{f}_{\delta_t}(x_t)\right\| \\ &\leqslant \left\|\frac{d}{\delta_t}(\tilde{f}(\hat{x}_t)-\tilde{f}(x_t))v_t\right\| + \left\|\mathbb{E}\left[\frac{d}{\delta_t}(\tilde{f}(\hat{x}_t)-\tilde{f}(x_t))v_t\middle| x_t\right]\right\| \\ &\leqslant 2\left\|\frac{d}{\delta_t}(\tilde{f}(\hat{x}_t)-\tilde{f}(x_t))v_t\right\| \leqslant 2\frac{d}{\delta_t}\tilde{L}\|\hat{x}_t - x_t\| \leqslant 2d\tilde{L}.\end{aligned} \tag{J.8}$$

Finally, we bound $\|\mathcal{E}_t^{(3)}\|$. Note that

$$\begin{aligned}\|\mathcal{E}_t^{(3)}\|^2 &= \left\|\frac{d}{\delta_t}\left((f(\hat{x}_t, y_t^{(K)})v_t - \tilde{f}(\hat{x}_t)v_t) - (\tilde{f}(x_t,\tilde{y}_t^{(K)})v_t - \tilde{f}(x_t)v_t)\right)\right\|^2 \\ &\leqslant 2\frac{d^2}{\delta_t^2}\left(\left\|f(\hat{x}_t, y_t^{(K)}) - \tilde{f}(\hat{x}_t, \mathsf{br}(\hat{x}_t))\right\|^2 + \left\|f(x_t,\tilde{y}_t^{(K)}) - \tilde{f}(x_t,\mathsf{br}(x_t))\right\|^2\right) \\ &\leqslant 2\frac{d^2}{\delta_t^2}L_2^2\left(\underbrace{\|y_t^{(K)} - \mathsf{br}(\hat{x}_t)\|^2}_{\text{Term A}} + \underbrace{\|\tilde{y}_t^{(K)} - \mathsf{br}(x_t)\|^2}_{\text{Term B}}\right).\end{aligned} \tag{J.9}$$

Recall, from Assumption 11.3.2 it holds that

$$\begin{aligned}\text{Term A} &= \|y_t^{(K)} - \mathsf{br}(\hat{x}_t)\|^2 \leqslant \alpha\|y_t^{(0)} - \mathsf{br}(\hat{x}_t)\|^2 \\ &= \alpha\|\tilde{y}_{t-1}^{(K)} - \mathsf{br}(\hat{x}_t)\|^2 = \alpha\|\tilde{y}_{t-1}^{(K)} - \mathsf{br}(x_{t-1}) + \mathsf{br}(x_{t-1}) - \mathsf{br}(\hat{x}_t)\|^2 \\ &= 2\alpha(\|\tilde{y}_{t-1}^{(K)} - \mathsf{br}(x_{t-1})\|^2 + L_S\|\hat{x}_t - x_{t-1}\|^2) \\ &\leqslant 2\alpha(\|\tilde{y}_{t-1}^{(K)} - \mathsf{br}(x_{t-1})\|^2 + 2L_S\|x_t - x_{t-1}\|^2 + 2L_S\|\delta_t v_t\|^2) \\ &\leqslant 2\alpha(\|\tilde{y}_{t-1}^{(K)} - \mathsf{br}(x_{t-1})\|^2 + 2L_S\|\delta_t v_t\|^2 \\ &\qquad + 2L_S\eta_{t-1}^2\frac{d^2}{\delta_{t-1}^2}\|f(\hat{x}_{t-1}, y_{t-1}^{(K)}) - f(x_{t-1},\tilde{y}_{t-1}^{(K)})\|^2) \\ &\leqslant 2\alpha(\|\tilde{y}_{t-1}^{(K)} - \mathsf{br}(x_{t-1})\|^2 + 2L_S\delta_{t-1}^2 \\ &\qquad + 2L_S\eta_{t-1}^2\frac{d^2}{\delta_{t-1}^2}\underbrace{\|f(\hat{x}_{t-1}, y_{t-1}^{(K)}) - f(x_{t-1},\tilde{y}_{t-1}^{(K)})\|^2}_{\text{Term C}}),\end{aligned} \tag{J.10}$$

where the last inequality follows by noting that (δ_t) is a decreasing sequence.

Similarly, we note that

$$\begin{aligned}
&\text{Term B} = \|\tilde{y}_t^{(K)} - \mathsf{br}(x_t)\|^2 \leqslant \alpha\|\tilde{y}_{t-1}^{(K)} - \mathsf{br}(x_t)\|^2 \\
&= \alpha\|\tilde{y}_{t-1}^{(K)} - \mathsf{br}(x_{t-1}) + \mathsf{br}(x_{t-1}) - \mathsf{br}(x_t)\|^2 \\
&= 2\alpha(\|\tilde{y}_{t-1}^{(K)} - \mathsf{br}(x_{t-1})\|^2 + L_S\|x_t - x_{t-1}\|^2) \\
&\leqslant 2\alpha\Bigg(\|\tilde{y}_{t-1}^{(K)} - \mathsf{br}(x_{t-1})\|^2 \\
&\qquad + L_S\eta_{t-1}^2\frac{d^2}{\delta_{t-1}^2}\underbrace{\|f(\hat{x}_{t-1}, y_{t-1}^{(K)}) - f(x_{t-1}, \tilde{y}_{t-1}^{(K)})\|^2}_{\text{Term C}}\Bigg).
\end{aligned} \tag{J.11}$$

To finish the bounds for Term A and Term B, we need to bound Term C

$$\begin{aligned}
&\text{Term C} = \|f(\hat{x}_{t-1}, y_{t-1}^{(K)}) - f(x_{t-1}, \tilde{y}_{t-1}^{(K)})\|^2 \\
&= \|f(\hat{x}_{t-1}, y_{t-1}^{(K)}) - f(x_{t-1}, y_{t-1}^{(K)}) + f(x_{t-1}, y_{t-1}^{(K)}) - f(x_{t-1}, \tilde{y}_{t-1}^{(K)})\|^2 \\
&\leqslant 2(\|f(\hat{x}_{t-1}, y_{t-1}^{(K)}) - f(x_{t-1}, y_{t-1}^{(K)})\|^2 + \|f(x_{t-1}, y_{t-1}^{(K)}) - f(x_{t-1}, \tilde{y}_{t-1}^{(K)})\|^2) \\
&= 2(L_1^2\|\hat{x}_{t-1} - x_{t-1}\|^2 + L_2^2\|y_{t-1}^{(K)} - \tilde{y}_{t-1}^{(K)}\|^2) \\
&= 2(L_1^2\delta_{t-1}^2 + L_2^2\underbrace{\|y_{t-1}^{(K)} - \tilde{y}_{t-1}^{(K)}\|^2}_{\text{Term D}}).
\end{aligned} \tag{J.12}$$

Note that from Assumption 11.3.1 it follows that Term D $\leqslant C^2\delta_{t-1}^2$. Consequently, we obtain the following bound

$$\begin{aligned}
\text{Term C} &= \|f(\hat{x}_{t-1}, y_{t-1}^{(K)}) - f(x_{t-1}, \tilde{y}_{t-1}^{(K)})\|^2 \\
&\leqslant 2(L_1^2\delta_{t-1}^2 + C_1^2\delta_{t-1}^2),
\end{aligned} \tag{J.13}$$

where $C_1 := L_2^2C^2$. Consequently, we can bound (J.10) as

$$\begin{aligned}
\|y_t^{(K)} - \mathsf{br}(\hat{x}_t)\|^2 &\leqslant 2\alpha\Bigg(\|\tilde{y}_{t-1}^{(K)} - \mathsf{br}(x_{t-1})\|^2 \\
&\qquad + L_S\eta_{t-1}^2\frac{d^2}{\delta_{t-1}^2}\left(2(L_1^2\delta_{t-1}^2 + C_1\delta_{t-1}^2)\right) + 2L_S\delta_{t-1}^2\Bigg) \\
&= 2\alpha\left(\|\tilde{y}_{t-1}^{(K)} - \mathsf{br}(x_{t-1})\|^2 + 2L_S\eta_{t-1}^2d^2(L_1^2 + C_1) + 2L_S\delta_{t-1}^2\right)
\end{aligned}$$

Define $e_t = \|y_t^{(K)} - \mathsf{br}(\hat{x}_t)\|^2$, $\tilde{e}_t = \|\tilde{y}_t^{(K)} - \mathsf{br}(\hat{x}_t)\|^2$. Then we have

$$e_t \leqslant \bar{\alpha}\left(\tilde{e}_{t-1} + C_2d^2\eta_{t-1}^2 + C_3d^2\eta_{t-1}^2 + C_4\delta_{t-1}^2\right), \tag{J.14}$$

$$\tilde{e}_t \leqslant \bar{\alpha}\left(\tilde{e}_{t-1} + C_2d^2\eta_{t-1}^2 + C_3d^2\eta_{t-1}^2\right), \tag{J.15}$$

where $\bar{\alpha} = 2\alpha, C_2 = 2L_S L_1^2, C_3 = 2L_S C_1$, $C_4 = 4L_S$ and $C_5 = 2L_S$. Consequently, it holds that

$$\begin{aligned}\tilde{e}_t &\leqslant \bar{\alpha}\left(\tilde{e}_{t-1} + C_2 d^2 \eta_{t-1}^2 + C_3 d^2 \eta_{t-1}^2\right) \\ &\leqslant \bar{\alpha}^t \tilde{e}_0 + \sum_{k=0}^{t-1} \bar{\alpha}^{t-k}\left(C_2 d^2 \eta_k^2 + C_3 d^2 \eta_k^2\right) \\ &\leqslant \bar{\alpha}^t \tilde{e}_0 + C_6 d^2 \sum_{k=0}^{t-1} \bar{\alpha}^{t-k} \eta_k^2, \end{aligned} \tag{J.16}$$

where $C_6 = C_2 + C_3$. Moreover, we also note that

$$\begin{aligned} e_t &\leqslant \bar{\alpha}(\tilde{e}_{t-1} + C_2 d^2 \eta_{t-1}^2 + C_3 d^2 \eta_{t-1}^2 + C_4 \delta_{t-1}^2) \\ &\leqslant \bar{\alpha}(\bar{\alpha}^{t-1}\tilde{e}_0 + C_6 d^2 \sum_{k=0}^{t-2} \bar{\alpha}^{t-1-k}\eta_k^2 + C_6 d^2 \eta_{t-1}^2 + C_4 \delta_{t-1}^2) \\ &\leqslant \bar{\alpha}^t \tilde{e}_0 + C_6 d^2 \sum_{k=0}^{t-1} \bar{\alpha}^{t-k}\eta_k^2 + \bar{\alpha} C_4 \delta_{t-1}^2. \end{aligned} \tag{J.17}$$

Combining (J.17) and (J.16) in (J.9). We obtain that

$$\|\mathcal{E}_t^{(3)}\|^2 \leqslant \frac{d^2}{\delta_t^2} L_2^2 \left(2\bar{\alpha}^t e_0 + 2C_6 \sum_{k=0}^{t-1} \bar{\alpha}^{t-k}\eta_k^2 + C_4 \sum_{k=0}^{t-1} \bar{\alpha}^{t-k}\delta_k^2\right). \tag{J.18}$$

Lemma 3 (Restated)**.** *The trajectory $z_s(\cdot)$ is an asymptotic pseudotrajectory of* (UL)*. That is, for any positive integer S*

$$\lim_{t\to\infty} \sup_{s\in[0,S]} \mathbb{E}\left[\|x_{t+s} - z_s(\hat{x}_t)\|^2\right] = 0$$

Proof. For any t, we see that

$$\begin{aligned} &\|x_{t+s+1} - z_{s+1}(x_t)\|^2 = \|x_{t+s} - \eta_{t+s}\hat{F}(x_t, \delta_t, v_t) - z_s(\hat{x}_t) + \eta_{t+s}\nabla\tilde{f}(z_s(\hat{x}_t))\|^2 \\ &\leqslant 2\|x_{t+s} - z_s(\hat{x}_t)\|^2 + 2\eta_{t+s}^2\|\hat{F}(x_{t+s}, \delta_{t+s}, v_{t+s}) - \nabla\tilde{f}(z_s(\hat{x}_t))\|^2 \\ &= 2\|x_{t+s} - z_s(\hat{x}_t)\|^2 + 2\eta_{t+s}^2\|\hat{F}(x_{t+s}, \delta_{t+s}, v_{t+s}) - \nabla\tilde{f}(x_{t+s}) + \nabla\tilde{f}(x_{t+s}) - \nabla\tilde{f}(z_s(\hat{x}_t))\|^2 \\ &\leqslant 2(1 + 4\eta_{t+s}^2\tilde{\ell}^2)\|x_{t+s} - z_s(\hat{x}_t)\|^2 + 4\eta_{t+s}^2\|\hat{F}(x_{t+s}, \delta_{t+s}, v_{t+s}) - \nabla\tilde{f}(x_{t+s})\|^2 \\ &= C_1\|x_{t+s} - z_s(\hat{x}_t)\|^2 + C_2\eta_{t+s}^2\|\hat{F}(x_{t+s}, \delta_{t+s}, v_{t+s}) - \nabla\tilde{f}(x_{t+s})\|^2 \\ &\leqslant C_1^s\|x_t - z_0(\hat{x}_t)\|^2 + \sum_{k=1}^{s} C_2 C_1^k \eta_{t+s-k}^2\|\hat{F}(x_{t+s-k}, \delta_{t+s-k}, v_{t+s-k}) - \nabla\tilde{f}(x_{t+s-k})\|^2 \end{aligned}$$

where $C_1 = 2(1 + 4(\bar{\eta}^2/d^2)\tilde{\ell}^2), C_2 = 4$. Next we note that $x_t = z_0(x_t)$. Consequently,

$$\mathbb{E}\left[\|x_{t+s+1} - z_{s+1}(\hat{x}_t)\|^2\right] \leqslant \sum_{k=1}^{s} C_1^{s-k}\eta_{t+k}^2\left(\mathbb{E}\left[\|\mathcal{E}_{t+k}^{(1)}\|^2 + \|\mathcal{E}_{t+k}^{(2)}\|^2 + \|\mathcal{E}_{t+k}^{(3)}\|^2\right]\right).$$

Therefore,

$$\begin{aligned}
&\sup_{s\in[0,S]} \mathbb{E}\left[\|x_{t+s+1} - z_{s+1}(\hat{x}_t)\|^2\right] \\
&\leqslant C_1^S \sum_{k=1}^{S} \eta_{t+k}^2 \Bigg(D_1\delta_{t+k}^2 + D_2 \\
&\qquad + \frac{d^2}{\delta_{t+k}^2} L_2^2 \Bigg(2\alpha^{t+k} e_0 + 2C_6 \sum_{p=0}^{t+k-1} \alpha^{t+k-p}\eta_p^2 + C_4 \sum_{p=0}^{t+k-1} \alpha^{t+k-p}\delta_p^2 \Bigg) \Bigg) \\
&\leqslant \mathcal{O}\Bigg(C_1^S \Bigg(\eta_t^2\delta_t^2 S + \eta_t^2 S + \eta_t\alpha^t S + \underbrace{\eta_t \sum_{k=1}^{S} \frac{\eta_{t+k}}{\delta_{t+k}^2} \sum_{p=0}^{t+k-1} \alpha^{t+k-p}\eta_p^2}_{\text{Term A}} \\
&\qquad + \underbrace{\eta_t \sum_{k=1}^{S} \frac{\eta_{t+k}}{\delta_{t+k}^2} \sum_{p=0}^{t+k-1} \alpha^{t+k-p}\delta_p^2}_{\text{Term B}} \Bigg) \Bigg).
\end{aligned}$$

Next, we analyze Term A and Term B by substituting $\eta_t = \bar{\eta}(t+1)^{-1/2}d^{-1}, \delta_t = \bar{\delta}(t+1)^{-1/4}d^{-1/2}$. First, we see that

$$\begin{aligned}
\text{Term A} &= \eta_t \sum_{k=1}^{S} \frac{\eta_{t+k}}{\delta_{t+k}^2} \sum_{p=0}^{t+k-1} \alpha^{t+k-p}\eta_p^2 \\
&= \frac{\bar{\eta}}{\bar{\delta}^2}\eta_t \sum_{k=1}^{S} \sum_{p=0}^{t+k-1} \alpha^{t+k-p}\eta_p^2 \\
&\leqslant \frac{\bar{\eta}}{\bar{\delta}^2}\eta_t\alpha^t \sum_{k=1}^{S} \alpha^k \sum_{p=0}^{t+k-1} \eta_p^2 \\
&\leqslant \frac{\bar{\eta}^3}{d^2\bar{\delta}^2}\eta_t\alpha^t \sum_{k=1}^{S} \alpha^k \sqrt{t+k} \\
&\leqslant \frac{\bar{\eta}^3}{d^2\bar{\delta}^2}\eta_t \frac{\alpha^{t+1}}{1-\alpha}\sqrt{t+S}.
\end{aligned}$$

Next, we analyze Term B

$$\begin{aligned}
\text{Term B} &= \eta_t \sum_{k=1}^{S} \frac{\eta_{t+k}}{\delta_{t+k}^2} \sum_{p=0}^{t+k-1} \alpha^{t+k-p} \delta_p^2 \\
&= \frac{\bar{\eta}}{\bar{\delta}^2} \eta_t \sum_{k=1}^{S} \sum_{p=0}^{t+k-1} \alpha^{t+k-p} \delta_p^2 \\
&\leqslant \frac{\bar{\eta}}{\bar{\delta}^2} \eta_t \alpha^t \sum_{k=1}^{S} \alpha^k \sum_{p=0}^{t+k-1} \delta_p^2 \\
&\leqslant \frac{\bar{\eta}}{d} \eta_t \alpha^t \sum_{k=1}^{S} \alpha^k \left(t+k\right)^{3/4} \\
&\leqslant \frac{\bar{\eta}^3}{d^2 \bar{\delta}^2} \eta_t \frac{\alpha^{t+1}}{1-\alpha} \left(t+S\right)^{3/4}.
\end{aligned}$$

To summarize, we obtain

$$\begin{aligned}
&\sup_{s \in [0,S]} \mathbb{E}\left[\|x_{t+s+1} - z_{s+1}(\hat{x}_t)\|^2\right] \\
&\leqslant \mathcal{O}\left(C_1^S \left(\eta_t^2 \delta_t^2 S + \eta_t^2 S + \eta_t \alpha^t S + \eta_t \alpha^{t+1} \sqrt{t+S} + \eta_t \alpha^{t+1} \left(t+S\right)^{3/4}\right)\right) \\
&= \mathcal{O}\left(C_1^S \left(\frac{S}{(t+1)^{3/2}} + \frac{S}{(t+1)} + \frac{\alpha^t S}{\sqrt{t+1}} + \frac{\sqrt{t+S}}{\sqrt{t+1}} \alpha^{t+1} + \frac{\left(t+S\right)^{3/4}}{\sqrt{t+1}} \alpha^{t+1}\right)\right).
\end{aligned}$$

Taking limit $t \to \infty$ we obtain the desired conclusion.

When does Assumption 11.3.1 hold?

Consider the scenario of bi-level optimization when the lower level problem is just a convex optimization problem with objective function $g(x, \cdot)$ a popular choice of H is projected gradient descent:

$$y^{(k+1)}(x) = H(y^{(k)}; x) = \mathcal{P}_Y(y^{(k)}(x) - \gamma \nabla_y g(x, y^{(k)}(x)). \tag{J.19}$$

Proposition 4. *Consider the bilevel optimization problem with convex-lower level optimization problem* (J.19)*. Then for any $x \in X$ the difference between $y^{(K)}(x)$ and $\tilde{y}^{(K)}(x)$ is bounded as*

$$\|y^{(K)}(\hat{x}) - y^{(K)}(x)\| \leqslant K L_3 \gamma \delta_t \exp(\gamma L_4 K) \tag{J.20}$$

Proof. We note that for any $k \in [K]$

$$\begin{aligned}
y^{(k)}(\hat{x}) &= \mathcal{P}_Y\left(y^{(k-1)}(\hat{x}) - \gamma \nabla g(\hat{x}, y^{(k-1)}(\hat{x}))\right), \\
y^{(k)}(x) &= \mathcal{P}_Y\left(y^{(k-1)}(x) - \gamma \nabla g(x, y^{(k-1)}(x))\right),
\end{aligned}$$

where we impose that $y^{(0)}(x) = y^{(0)}(\hat{x})$ due to Algorithm 10. Then

$$\begin{aligned}
&\|y^{(k)}(\hat{x}) - y^{(k)}(x)\| \\
&= \|\mathcal{P}_Y(y^{(k-1)}(\hat{x}) - \gamma\nabla g(\hat{x}, y^{(k-1)}(\hat{x}))) - \mathcal{P}_Y(y^{(k-1)}(x) - \gamma\nabla g(x, y^{(k-1)}(x))\| \\
&\leqslant \|y^{(k-1)}(\hat{x}) - y^{(k-1)}(x)\| + \gamma\|\nabla g(\hat{x}, y^{(k-1)}(\hat{x})) - \nabla g(x, y_t^{(k-1)}(x))\| \\
&\leqslant \gamma \sum_{\ell=0}^{k-1} \|\nabla g(\hat{x}, y^{(\ell)}(\hat{x})) - \nabla g(x, y^{(\ell)}(x))\| \\
&\leqslant \gamma \sum_{\ell=0}^{k-1} \Big(\|\nabla g(\hat{x}, y^{(\ell)}(\hat{x})) - \nabla g(\hat{x}, y^{(\ell)}(x))\| + \|\nabla g(\hat{x}, y^{(\ell)}(x)) - \nabla g(x, y^{(\ell)}(x))\|\Big) \\
&\leqslant \gamma k L_3\delta + \gamma L_4 \sum_{\ell=0}^{k-1} \|y^{\ell}(\hat{x}) - y^{\ell}(x)\| \\
&\leqslant KL_3\gamma\delta + \gamma L_4 \sum_{\ell=0}^{k-1} \|y^{\ell}(\hat{x}) - y^{\ell}(x)\|
\end{aligned}$$

By discrete Gronwall inequality stated in Lemma J.1.2 we obtain

$$\|y^{(k)}(\hat{x}) - y^{(k)}(x)\| \leqslant KL_3\gamma\delta\exp(\gamma L_4 k), \quad \forall\ k \in [K].$$

J.3 Proof of Main Results

The following appendices contain the proofs of the main two theorems.

Proof of Theorem 11.3.1

From Lemma 1, we know that $\tilde{f}(\cdot)$ approximately decreases along the trajectory of (UL). That is,

$$\begin{aligned}
&\mathbb{E}\left[\tilde{f}(x_{t+1})\right] \\
&\leqslant \mathbb{E}\left[\tilde{f}(x_t)\right] - \frac{\eta_t}{2}\mathbb{E}\left[\|\nabla\tilde{f}(x_t)\|^2\right] + \eta_t\mathbb{E}\left[\|\mathcal{E}_t^{(1)}\|^2\right] + \eta_t\mathbb{E}\left[\|\mathcal{E}_t^{(3)}\|^2\right] + \tilde{\ell}\eta_t^2\mathbb{E}\left[\|\mathcal{E}_t^{(2)}\|^2\right] \quad \text{(J.21)}
\end{aligned}$$

Using the bounds on error terms from Lemma 2, we obtain

$$\begin{aligned}
\mathbb{E}\left[\tilde{f}(x_{t+1})\right] \leqslant\ &\mathbb{E}\left[\tilde{f}(x_t)\right] - \frac{\eta_t}{2}\mathbb{E}\left[\|\nabla\tilde{f}(x_t)\|^2\right] + \eta_t\frac{\tilde{\ell}^2\delta_t^2 d^2}{4} \\
&+ \eta_t\left(\frac{d^2}{\delta_t^2}L_2^2\left(2\alpha^t e_0 + 2C_6 d^2\sum_{k=0}^{t-1}\alpha^{t-k}\eta_k^2 + C_4\sum_{k=0}^{t-1}\alpha^{t-k}\delta_k^2\right)\right) + 4d^2\tilde{L}^2\tilde{\ell}\eta_t^2.
\end{aligned}$$

Re-arranging the terms and adding and subtracting the term $\tilde{f}(x^*) = \min_x \tilde{f}(x)$ we obtain

$$\frac{\eta_t}{2}\mathbb{E}\left[\|\nabla \tilde{f}(x_t)\|^2\right] \leqslant \mathbb{E}\left[\tilde{f}(x_t)\right] - \tilde{f}(x^*) - \left(\mathbb{E}\left[\tilde{f}(x_{t+1})\right] - \tilde{f}(x^*)\right) + \eta_t \frac{\tilde{\ell}^2 \delta_t^2 d^2}{4} + \eta_t \frac{d^2}{\delta_t^2} L_2^2 \left(2\alpha^t e_0 + 2C_6 d^2 \sum_{k=0}^{t-1} \alpha^{t-k}\eta_k^2 + C_4 \sum_{k=0}^{t-1} \alpha^{t-k}\delta_k^2\right) + 4d^2 \tilde{L}^2 \tilde{\ell} \eta_t^2.$$

Summing the previous equation over time step t, we obtain

$$\sum_{t\in[T]} \eta_t \mathbb{E}\left[\|\nabla \tilde{f}(x_t)\|^2\right] \leqslant \left(\tilde{f}(x_0) - \tilde{f}(x^*)\right) + \frac{\tilde{\ell}^2 d^2}{4} \sum_{t\in[T]} \eta_t \delta_t^2 + 2e_0 d^2 L_2^2 \sum_{t\in[T]} \frac{\eta_t}{\delta_t^2}\alpha^t + \underbrace{2C_6 d^4 L_2^2 \sum_{t\in[T]} \frac{\eta_t}{\delta_t^2} \sum_{k=0}^{t-1} \alpha^{t-k}\eta_k^2}_{\text{Term E}} + \underbrace{C_4 L_2^2 d^2 \sum_{t\in[T]} \frac{\eta_t}{\delta_t^2} \sum_{k=0}^{t-1} \alpha^{t-k}\delta_k^2}_{\text{Term F}} + 4d^2 \tilde{L}^2 \tilde{\ell} \sum_{t\in[T]} \eta_t^2. \tag{J.22}$$

Setting $\eta_t = \eta_0 (t+1)^{-1/2} d^{-1}$ and $\delta_t = \delta_0 (t+1)^{-1/4} d^{-1/2}$ we obtain

$$\frac{1}{\sum_{t\in[T]}\eta_t} \sum_{t\in[T]} \eta_t \mathbb{E}\left[\|\nabla \tilde{f}(x_t)\|^2\right] \leqslant \frac{d}{\eta_0\sqrt{T}}\left(\tilde{f}(x_0) - \tilde{f}(x^*)\right) + \frac{\tilde{\ell} d \log(T)\delta_0^2}{4\sqrt{T}} + \frac{2e_0 d^3 L_2^2 \alpha}{(1-\alpha)\eta_0\sqrt{T}} + \frac{1}{\sum_{t\in[T]}\eta_t} \underbrace{2C_6 d^4 L_2^2 \sum_{t\in[T]} \frac{\eta_t}{\delta_t^2} \sum_{k=0}^{t-1} \alpha^{t-k}\eta_k^2}_{\text{Term E}} + \frac{1}{\sum_{t\in[T]}\eta_t} \underbrace{C_4 L_2^2 d^2 \sum_{t\in[T]} \frac{\eta_t}{\delta_t^2} \sum_{k=0}^{t-1} \alpha^{t-k}\delta_k^2}_{\text{Term F}} + \frac{4\tilde{L}^2 \tilde{\ell}^2 \eta_0 \log(T) d}{\sqrt{T}}$$

Let's consider the following term by defining $C_7 = 2C_6L_2^2, C_8 = C_4L_2^2$

$$\begin{aligned}
\text{Term E } + \text{Term F} &= \sum_{t=1}^{T} \frac{\eta_t}{\delta_t^2} \sum_{k=0}^{t-1} \alpha^{t-k} \left(C_7 d^4 \eta_k^2 + C_8 d^2 \delta_k^2 \right) \\
&= \frac{\eta_0}{\delta_0^2} \sum_{t=1}^{T} \sum_{k=0}^{t-1} \alpha^t \frac{\Theta_k}{\alpha^k} \\
&= \frac{\eta_0}{\delta_0^2} \sum_{k=0}^{T-1} \frac{\Theta_k}{\alpha^k} \sum_{t=k+1}^{T} \alpha^t \\
&\leqslant \frac{\eta_0}{\delta_0^2} \sum_{k=0}^{T-1} \frac{\Theta_k}{\alpha^k} \frac{\alpha^{k+1}}{1-\alpha} \\
&= \frac{\eta_0}{\delta_0^2} \frac{\alpha}{1-\alpha} \sum_{k=0}^{T-1} \Theta_k \\
&= \frac{\eta_0}{\delta_0^2} \frac{\alpha}{1-\alpha} \left(C_7 d^2 \eta_0^2 \log(T) + C_8 d^2 \delta_0^2 \sqrt{T} \right),
\end{aligned}$$

where in second equality $\Theta_k := (C_7 d^4 \eta_k^2 + C_8 d^2 \delta_k^2)$.

Thus from (J.22) we obtain

$$\begin{aligned}
\frac{1}{\sum_{t\in[T]} \eta_t} \sum_{t\in[T]} \eta_t \mathbb{E}\left[\|\nabla \tilde{f}(x_t)\|^2 \right] &\leqslant \frac{d}{\eta_0 \sqrt{T}} \left(\tilde{f}(x_0) - \tilde{f}(x^*) \right) + \frac{\tilde{\ell} d \log(T) \delta_0^2}{4\sqrt{T}} \\
&+ \frac{2e_0 d^3 L_2^2 \alpha}{(1-\alpha)\eta_0 \sqrt{T}} + \frac{d}{\delta_0^2 \sqrt{T}} \frac{\alpha}{1-\alpha} \left(C_7 d^2 \eta_0^2 \log(T) + C_8 d^2 \delta_0^2 \sqrt{T} \right) \\
&+ \frac{4\tilde{L}^2 \tilde{\ell}^2 \eta_0 \log(T) d}{\sqrt{T}}.
\end{aligned}$$

Thus, overall we obtain

$$\frac{1}{\sum_{t\in[T]} \eta_t} \sum_{t\in[T]} \eta_t \mathbb{E}\left[\|\nabla \tilde{f}(x_t)\|^2 \right] \leqslant \tilde{O}\left(\frac{d}{\sqrt{T}} + \frac{\alpha}{1-\alpha} d^3 \right).$$

Proof of Theorem 11.3.2

The proof follows by contradiction. Suppose there exists a saddle point x^* such that

$$\lim_{t\to\infty} \mathbb{E}\left[\|x_t - x^*\|^2 \right] = 0.$$

This implies that for any $\epsilon > 0$ there exists an integer T_ϵ such that for all $t \geqslant T_\epsilon$ it holds that

$$\mathbb{E}\left[\|x_{t+s} - x^*\|^2 \right] \leqslant \epsilon/4 \quad \forall s \geqslant 0.$$

Moreover, from Lemma 11.3.3 we know that for any S there exists $\tilde{T}_{\epsilon,S}$ such that

$$\sup_{s\in[0,S]} \mathbb{E}\left[\|z_s(\hat{x}_t) - x_{t+s}\|^2\right] \leqslant \epsilon/4 \quad \forall\ t \geqslant \tilde{T}_{\epsilon,S}.$$

Finally, note that

$$\|z_s(\hat{x}_t) - x^*\|^2 \leqslant 2\|z_s(\hat{x}_t) - x_{t+s}\|^2 + 2\|x_{t+s} - x^*\|^2.$$

Therefore, for any S and $t \geqslant \max\{T_\epsilon, T_{\epsilon,S}\}$

$$\|z_s(\hat{x}_t) - x^*\|^2 \leqslant \epsilon, \quad \forall\ s \in [0, S].$$

But from [235] we know that for gradient descent with random initialization[1] there exists S_ϵ such that for all $s \geqslant S_\epsilon$ it holds that

$$\|z_s(\hat{x}_t) - x^*\|^2 \geqslant 2\epsilon.$$

This establishes contradiction.

[1]Specifically, we use the results from [235, Proposition 8]. Even though the results in [235, Proposition 8] hold for gradient descent update with constant step-size, we can use this result for decaying step size as well. This is because the proof of [235, Proposition 8] only requires each step of the gradient update to be diffeomorphism, which holds in our setting as the step-sizes are constantly decaying.

Appendix K

Appendix for Chapter 12

K.1 Counter-example.

In this section, we present a non-atomic game, where the standard gradient-based incentive design approach would fail. Specifically, we will show that the gradient of equilibrium strategy with respect to incentive is singular and the equilibrium social cost function is non-convex in the incentive. Furthermore, we show that the fixed points of the gradient-based incentive update is non-unique, and almost all fixed points fail to induce a socially efficient outcome. In contrast, the fixed point of our externality-based incentive update is unique and results in a socially optimal outcome.

Consider a non-atomic routing game, comprising of two nodes and two edges connecting them. This network is used by one unit of travelers traveling from the source node $\mathcal{S}$ to the destination node $\mathcal{D}$. The latency function of two edges are denoted in Figure K.1. In this

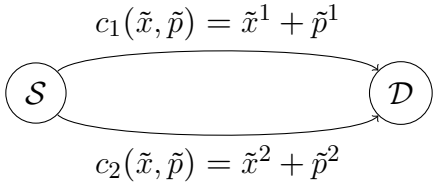


Figure K.1: Two-link routing game.

game, the strategy set is $\tilde{X} = \{\tilde{x} \in \mathbb{R}^2 : \tilde{x}_1 + \tilde{x}_2 = 1\}$. The equilibrium congestion levels on the two edges is obtained by computing the minimizer of the following function [435]:

$$\Phi(\tilde{x}, \tilde{p}) = \frac{1}{2}\tilde{x}_1^2 + \frac{1}{2}\tilde{x}_2^2 + \tilde{p}_1\tilde{x}_1 + \tilde{p}_2\tilde{x}_2.$$

Thus, for any toll vector $\tilde{p}$, the Nash equilibrium $\tilde{x}^*(\tilde{p}) = \arg\min_{\tilde{x}\in\tilde{X}} \Phi(\tilde{x}, \tilde{p})$ satisfies $\tilde{x}_1^*(\tilde{p}) = \mathcal{P}_{[0,1]}\left(\frac{\tilde{p}_2 - \tilde{p}_1 + 1}{2}\right)$, $\tilde{x}_2^*(\tilde{p}) = \mathcal{P}_{[0,1]}\left(\frac{\tilde{p}_1 - \tilde{p}_2 + 1}{2}\right)$, where for any scalar $x \in \mathbb{R}$, $\mathcal{P}_{[0,1]}(x)$

denotes its projection onto the line segment $[0, 1]$. The gradient of equilibrium strategy with respect to incentive is a singular matrix for all incentives $\tilde{p} = (\tilde{p}_1, \tilde{p}_2) \in \mathbb{R}^2$ such that $|\tilde{p}_1 - \tilde{p}_2| > 1$.

The equilibrium social cost function is:

$$\begin{aligned}\tilde{\Phi}(\tilde{x}^*(\tilde{p})) &= \tilde{x}_1^*(\tilde{p})\tilde{\ell}_1(\tilde{x}_1^*(\tilde{p})) + \tilde{x}_2^*(\tilde{p})\tilde{\ell}_2(\tilde{x}_2^*(\tilde{p})) \\ &= \begin{cases} \frac{(\tilde{p}_1-\tilde{p}_2)^2+1}{2}, & \text{if } |\tilde{p}_1 - \tilde{p}_2| \leqslant 1, \\ 1, & \text{otherwise.} \end{cases}\end{aligned}$$

Note that the equilibrium social cost function $\tilde{\Phi}(\tilde{x}^*(\tilde{p}))$ is non-convex in $\tilde{p}$, which contradicts the assumption commonly adopted in gradient-based incentive learning literature [253, 243, 302]. Furthermore, the gradient-based update[1] for this function takes the following form:

$$\tilde{p}_{k+1} = \tilde{p}_k - \beta_k \partial\Phi(\tilde{x}^*(\tilde{p}_k)),$$

where

$$\partial\Phi(\tilde{x}^*(\tilde{p})) \in \begin{cases} \left\{ \begin{bmatrix} \tilde{p}_1 - \tilde{p}_2 \\ \tilde{p}_2 - \tilde{p}_1 \end{bmatrix} \right\}, & \text{if } |\tilde{p}_1 - \tilde{p}_2| < 1, \\ \mathsf{conv}\left\{ \begin{bmatrix} \pm 1 \\ \mp 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}, & \text{if } \tilde{p}_1 - \tilde{p}_2 = \pm 1, \\ \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}, & \text{otherwise.} \end{cases}$$

Consequently, the set of fixed points for the gradient-based incentive update (i.e., where the gradient is zero) is given by:

$$\{(\tilde{p}_1, \tilde{p}_2) \in \mathbb{R}^2 : |\tilde{p}_1 - \tilde{p}_2| \in \{0\} \cup \{[1, \infty)\}\}. \tag{K.1}$$

On the other hand, the set of socially optimal tolls that minimize $\tilde{\Phi}(\tilde{x}^*(\tilde{p}))$ is given by $\{(\tilde{p}_1, \tilde{p}_2) \in \mathbb{R}^2 : \tilde{p}_1 = \tilde{p}_2\}$, which has measure zero within the set of fixed points of the gradient-based update (K.1).

In contrast, the fixed point of our externality-based incentive mechanism (cf. (12.10)) is unique $\tilde{p}_1^\dagger = \tilde{p}_2^\dagger = 1/2$ and minimizes the social cost.

K.2 Proofs and Additional Results on Aggregative Game in Section 12.3

In this section, we present the proofs of Propositions 12.3.1 and 12.3.2. Additionally, we introduce a generalization of the results in Section 12.3.

[1]Since this function is non-differentiable, it is common to use Clarke's subdifferential to study gradient-based updates [99].

Proof of Proposition 12.3.1

First, we show that $x^*(p) = -M^{-1}p$ for any $p \in \mathbb{R}^{|I|}$. Note that the cost function $c_i(x_i, x_{-i}, p)$ is strongly convex in x_i and that the strategy space X_i is unconstrained, ensuring that the game is strongly convex game. Therefore, $x^*(p)$ is Nash equilibrium if and only if $\nabla_{x_i} c_i(x^*(p), p) = 0$, for every $i \in I$. Consequently, using (12.29), we obtain

$$q_i x_i^*(p) + \alpha(Ax^*(p))_i + p_i = 0, \quad \forall i \in I. \tag{K.2}$$

Stacking (K.2) in vector form yields $Mx^*(p) = -p$.

Next, we show that $P^\dagger$ is a singleton set. Note that

$$P^\dagger = \{p^\dagger \in \mathbb{R}^{|I|} : x_i^*(p^\dagger) = \zeta_i, \ \ \forall\, i \in I\}. \tag{K.3}$$

The proof concludes by noting that $x^*(p) = -M^{-1}p$.

Proof of Proposition 12.3.2

Here, we verify the requirements (R1') and (R2) of Proposition 12.2.3. We start with verifying (R1'). We define a Lyapunov function candidate $V(p) = (p - p^\dagger)^\top M^{-\top}(p - p^\dagger)$ for the dynamical system (12.20). Note that $V(p^\dagger) = 0$ and $V(p) > 0$ for all $p \neq p^\dagger$. Next, we show that $\nabla V(p)^\top (e(x^*(p)) - p) < 0$, for every $p \neq p^\dagger$. Indeed,

$$\begin{aligned}\nabla V(p)^\top (e(x^*(p)) - p) &= 2(p - p^\dagger)^\top M^{-\top}(x^*(p) - \zeta) \\ &= -2(x^*(p) - x^*(p^\dagger))(x^*(p) - x^*(p^\dagger)) < 0, \quad \forall\, p \neq p^\dagger.\end{aligned}$$

This shows that $p^\dagger$ is globally asymptotically stable for (12.20).

Next, for requirement (R2) of Proposition 12.2.3, we verify sufficient conditions for the boundedness of iterates in two-timescale approximation theory [225]. In particular, using [225, Theorem 10], it is sufficient to show that the following two conditions are satisfied:

(a) The function $f_c(x, p) := \frac{1}{c}(f(cx, cp) - cx)$ satisfies $f_c \to f_\infty$ as $c \to \infty$, uniformly on compact sets, for some f_∞. Also, for every incentive vector $p \in \mathbb{R}^{|I|}$, $x^*(p)$ is the globally asymptotically stable fixed point of the following continuous-time dynamical system:

$$\dot{x}(t) = f_\infty(x(t), p). \tag{K.4}$$

Furthermore, $x^*(0) = 0$, and the system $\dot{x}(t) = f_\infty(x(t), 0)$ has the origin as its globally asymptotically stable fixed point.

(b) The function $h_c(p) := \frac{1}{c}(e(cx^*(p)) - cp)$ satisfies $h_c \to h_\infty$ as $c \to \infty$, uniformly on compact sets, for some h_∞. Also, the origin is a globally asymptotically stable fixed point of the dynamical system:

$$\dot{p}(t) = h_\infty(p(t)). \tag{K.5}$$

Condition (a) is satisfied due to Proposition 12.3.2-(ii) and the fact that $x^*(p) = -M^{-1}p$ in the atomic aggregative game. Condition (b) holds since $h_\infty(p) = -(Q + \alpha A)p$. Moreover, since $Q + \alpha A$ is symmetric positive definite, the origin is a globally asymptotically stable fixed point of (K.5).

Additional Results

Here, we consider a more general social cost function than (12.30). Specifically, we consider

$$\Phi(x) = \sum_{i \in I} h_i(x_i), \tag{K.6}$$

where, for every $i \in I$, the function $h_i : \mathbb{R} \to \mathbb{R}$ satisfies the following assumption:

Assumption K.2.1. *For every $i \in I$, $h_i(\cdot)$ is a strictly convex function with a Lipschitz continuous gradient. Furthermore, we assume the existence of $y^\dagger \in \mathbb{R}^{|I|}$ such that $\nabla h_i(y_i^\dagger) = 0$ for every $i \in I$.*

Proposition K.2.1. *Suppose that Assumption K.2.1 holds and $M := Q + \alpha A$ is invertible. Then, the Nash equilibrium $x^*(p) = M^{-1}p$ for any $p \in \mathbb{R}^{|I|}$. Furthermore, the set $P^\dagger$ is singleton.*

Proof. The proof that $x^*(p) = -M^{-1}p$ follows exactly as in Proposition 12.3.2. Next, we show that $P^\dagger$ is a singleton. Using (12.8a), (12.29), and (K.6), the externality is given by

$$e_i(x) = \nabla h_i(x_i) - q_i x_i - \alpha \sum_{j \in I} A_{ij} x_j, \quad \forall i \in I. \tag{K.7}$$

Combining (K.2) and (K.7), we obtain

$$e_i(x^*(p)) = \nabla h_i(x_i^*(p)) + p_i, \quad \forall i \in I. \tag{K.8}$$

Consequently, using (12.10), we have

$$P^\dagger = \{p^\dagger \in \mathbb{R}^{|I|} : \nabla h_i(x_i^*(p^\dagger)) = 0, \ \forall i \in I\}. \tag{K.9}$$

Since h_i is strictly convex, it follows from Assumption K.2.1 that there exists a unique $y^\dagger$ such that $\nabla h_i(y_i^\dagger) = 0$ for every $i \in I$. Therefore, for every $p^\dagger \in P^\dagger$, it must hold that $x^*(p^\dagger) = y^\dagger$. Since $x^*(p) = -M^{-1}p$, it follows that $p^\dagger = -My^\dagger$, establishing the uniqueness of $P^\dagger$. □

Next, we provide sufficient conditions to ensure the convergence of (x-update)-(p-update) to the fixed points. In particular, we present two sets of conditions: the first set establishes global convergence guarantees, while the second set ensures local convergence guarantees.

Proposition K.2.2. *Consider the updates* (x-update)-(p-update) *associated with the aggregative game G. Suppose that Assumptions 12.2.1, 12.2.2, and K.2.1 hold, and that*

$$\sup_{k\in\mathbb{N}}(\|x_k\|+\|p_k\|)<+\infty.$$

Additionally,

(i) If $M := Q+\alpha A$ is symmetric positive definite, then the discrete-time updates (x-update) *and* (p-update) *globally converge to the fixed point $(x^\dagger, p^\dagger)$ in the sense of Definition 12.2.2.*

(ii) If $M := Q+\alpha A$ is invertible with non-negative entries, M^{-1} has strictly negative off-diagonal entries, and there exists a vector $y^\dagger \in \mathbb{R}^{|I|}_-$ such that $\nabla h_i(y_i^\dagger)=0$ for every $i \in I$ [2] then the discrete-time updates (x-update) *and* (p-update) *locally converge to the fixed point $(x^\dagger, p^\dagger)$ in the sense of Definition 12.2.2.*

Propositions K.2.2-(i) and 12.3.2 are related but differ in two key aspects. First, the social cost function in Proposition 12.3.2 is a special case of the more general function in (K.6). Second, Proposition K.2.2-(i) directly assumes boundedness of iterates, $\sup_{k\in\mathbb{N}}(\|x_k\|+\|p_k\|)<+\infty$, whereas Proposition 12.3.2 instead relies on the global convergence of the limiting dynamical system associated with strategy updates (cf. (12.31)). The simpler social cost function (12.30) in Proposition 12.3.2 allows us to use stability results from two-timescale approximation theory [225, Theorem 10] to establish boundedness. Extending this approach to Proposition K.2.2 would require imposing global convergence of suitably defined limiting dynamical systems (cf. (K.4)-(K.5)). To maintain clarity, we impose $\sup_{k\in\mathbb{N}}(\|x_k\|+\|p_k\|)<+\infty$ directly in Proposition K.2.2.

The conditions imposed on the matrix M in Proposition K.2.2-(i) and (ii) are not directly comparable; neither necessarily implies the other [3].

Proof of Proposition K.2.2: We prove Proposition K.2.2(i)-(ii) in order.

(a) Proposition K.2.2-(i) follows by verifying the requirements (R1')-(R2) of Proposition 12.2.3. We only need to verify (R1') as (R2) is satisfied due to the assumption that $\sup_k \|x_k\|+\|p_k\|<\infty$. We define a Lyapunov function candidate $V(p)=$

[2] A similar statement can be obtained for the case when $y^\dagger \in \mathbb{R}^{|I|}_+$, but we omit it for brevity.

[3] For example, consider aggregative games with parameters (Q_1, A_1) and (Q_2, A_2) such that $M_1 = Q_1+\alpha A_1$, $M_2 = Q_2+\alpha A_2$ and

$$M_1=\begin{bmatrix}1 & 0.1\\ 1 & 1\end{bmatrix},\quad M_2=\begin{bmatrix}1 & -0.1\\ -0.1 & 1\end{bmatrix}.$$

The matrix M_1 satisfies the conditions in Proposition K.2.2-(ii) but does not satisfy the conditions in Proposition K.2.2-(i). On the other hand, the matrix M_2 satisfies the conditions in Proposition K.2.2-(i) but does not satisfy the conditions in Proposition K.2.2-(ii).

$(p-p^\dagger)^\top M^{-\top}(p-p^\dagger)$ for the dynamical system (12.20). Note that $V(p^\dagger)=0$ and $V(p)>0$ for all $p\neq p^\dagger$. Next, we show that $\nabla V(p)^\top(e(x^*(p))-p)<0$ for every $p\neq p^\dagger$. Indeed,

$$\begin{aligned}
&\nabla V(p)^\top(e(x^*(p))-p) = 2(p-p^\dagger)^\top M^{-\top}\nabla h(x^*(p))\\
&= -2(x^*(p)-x^*(p^\dagger))\nabla h(x^*(p))\\
&\overset{(K.9)}{=} -2(x^*(p)-x^*(p^\dagger))\left(\nabla h(x^*(p))-\nabla h(x^*(p^\dagger))\right)\\
&= -2(x^*(p)-x^*(p^\dagger))\left(\nabla h(x^*(p))-\nabla h(x^*(p^\dagger))\right)\\
&< 0, \quad \forall\, p\neq p^\dagger,
\end{aligned}$$

where the last equality follows from the strict convexity of h_i for each $i\in I$, completing the proof.

(b) Proposition K.2.2-(ii) follows by verifying conditions (R1) and (R2) of Proposition 12.2.3. Given that $\sup_k \|x_k\|+\|p_k\|<\infty$, it suffices to verify (R1). This follows since condition (C1) in Lemma 1 holds under Proposition 12.3.2-(ii) and Assumption K.2.1.

First, we show that for $i,j\in I$ with $i\neq j$, it holds that $\frac{\partial e_i(x^*(p))}{\partial p_j}>0$. Indeed,

$$\begin{aligned}
\frac{\partial e_i(x^*(p))}{\partial p_j} &= \nabla^2 h_i(x_i^*(p))\frac{\partial x_i^*(p)}{\partial p_j}\\
&= \nabla^2 h_i(x_i^*(p))(-M^{-1})_{ij} > 0,
\end{aligned}$$

where the inequality follows from the strict convexity of h_i and the fact that $(M^{-1})_{ij}<0$. Second, we show that condition (C1)-(i) in Lemma 1 holds. First, we establish that $e_i(x^*(0))\geqslant 0$ for every $i\in I$. From (K.8), we note that $e_i(x^*(0))=\nabla h_i(0)$ for every $i\in I$. Therefore, it suffices to show that $\nabla h_i(0)\geqslant 0$ for all $i\in I$. By Assumption K.2.1, $\nabla h_i(\cdot)$ is strictly increasing, and for each $i\in I$, there exists a unique $y_i^\dagger\leqslant 0$ such that $\nabla h_i(y_i^\dagger)=0$. This implies that $\nabla h_i(0)\geqslant 0$, for every $i\in I$. Next, we verify that $p^\dagger\in\mathbb{R}_+^{|I|}$. From Proposition K.2.1, $p^\dagger=-My^\dagger$. Since M has non-negative entries and $y^\dagger\in\mathbb{R}_-^{|I|}$, it follows that $p^\dagger\in\mathbb{R}_+^{|I|}$. Finally, we show the other condition in (C1)-(i) in Lemma 1, which requires that for any $p\in\mathbb{R}_+^{|I|}$, there exists $p'\in\mathbb{R}_+^{|I|}$ such that for every $i\in I$, $p_i'>p_i$ and $e_i(x^*(p'))-p_i'\leqslant 0$, for all $i\in I$. To show this, we define $p^\epsilon=-(1+\epsilon)My^\dagger$ for every $\epsilon>0$. Note that $p^\epsilon\in\mathbb{R}_+^{|I|}$ and for any $p\in\mathbb{R}_+^{|I|}$, we can select $\epsilon>0$ such that $p_i^\epsilon>p_i$ for every $i\in I$. Therefore, we show that for every $\epsilon>0$,

$$e_i(x^*(p^\epsilon))-p^\epsilon\leqslant 0, \quad \forall i\in I. \tag{K.10}$$

From (K.8), we note that $e_i(x^*(p^\epsilon))-p^\epsilon=\nabla h_i(x_i^*(p^\epsilon))$, for every $i\in I$. Therefore, to show (K.10), it is sufficient to show that

$$\nabla h_i(x_i^*(p^\epsilon))\leqslant 0, \quad \forall i\in I,\epsilon>0. \tag{K.11}$$

Indeed, for every $i \in I$ and $\epsilon > 0$,

$$\begin{aligned} 0 &< (\nabla h_i(x_i^*(p^\epsilon)) - \nabla h_i(y_i^\dagger))(x_i^*(p^\epsilon) - y_i^\dagger) \\ &= \nabla h_i(x_i^*(p^\epsilon))\epsilon y_i^\dagger, \end{aligned}$$

where we note that $x^*(p^\epsilon) = (1+\epsilon)y^\dagger$. To conclude, (K.11) follows because $y_i^\dagger \leqslant 0$ and $\epsilon > 0$.

K.3 Proofs of Results in Section 12.3

Proof of Proposition 12.3.3.

First, we show that $\tilde{\mathbf{P}}^\dagger$ is non-empty. This can be shown analogously to the proof of existence in Proposition 12.2.1 by using the Schauder fixed-point theorem and the continuity of the function $\tilde{w}^*(\cdot)$. We omit the details of this proof for the sake of brevity.

Next, we show that any $p^\dagger \in \tilde{\mathbf{P}}^\dagger$ aligns the Nash equilibrium with social optimality, i.e. $\tilde{w}(p^\dagger) = w^\dagger$. For any $p^\dagger \in \tilde{\mathbf{P}}^\dagger$, we have $\tilde{p}_a^\dagger = \tilde{w}_a^*(\tilde{p}^\dagger)\nabla l_a(\tilde{w}_a^*(\tilde{p}^\dagger))$ for every $a \in \tilde{\mathcal{E}}$. This implies, for every $a \in \tilde{\mathcal{E}}$,

$$\frac{\partial}{\partial \tilde{w}_a}\Big(\tilde{w}_a(\tilde{p}^\dagger) l_a(\tilde{w}_a(\tilde{p}^\dagger))\Big) = l_a(\tilde{w}_a(\tilde{p}^\dagger)) + \tilde{p}_a^\dagger. \tag{K.12}$$

Note that for any arbitrary edge toll $\tilde{p} \in \mathbb{R}^{|\tilde{\mathcal{E}}|}$, $\tilde{w}^*(\tilde{p})$ is the unique solution to the following strictly convex optimization problem [361].

$$\min_{\tilde{w} \in \tilde{W}} \quad \tilde{T}(\tilde{w}) = \sum_{a \in \tilde{\mathcal{E}}} \int_0^{\tilde{w}_a} l_a(\tau)\, d\tau + \sum_{a \in \tilde{\mathcal{E}}} \tilde{p}_a \tilde{w}_a. \tag{K.13}$$

Therefore, $\tilde{w}^*(\tilde{p})$ is a Nash equilibrium if and only if

$$\sum_{a \in \tilde{\mathcal{E}}} \Big(l_a\big(\tilde{w}_a(\tilde{p}) + \tilde{p}_a\big)\big(\tilde{w}_a - \tilde{w}_a(\tilde{p})\big)\Big) \geqslant 0, \ \forall\ \tilde{w} \in \tilde{W}. \tag{K.14}$$

Combining (K.12) and (K.14), we conclude that for every $\tilde{w} \in \tilde{W}$,

$$\sum_{a \in \tilde{\mathcal{E}}} \frac{\partial}{\partial \tilde{w}_a}(\tilde{w}_a(\tilde{p}^\dagger) l_a(\tilde{w}_a(\tilde{p}^\dagger)))(\tilde{w}_a - \tilde{w}_a^*(\tilde{p}^\dagger)) \geqslant 0. \tag{K.15}$$

Further, from the first-order conditions of optimality for the social cost function, we know that $\tilde{x}^\dagger$ is socially optimal if and only if, for every $\tilde{x} \in \tilde{X}$,

$$\sum_{i \in \tilde{I}} \sum_{j \in \mathbf{R}_i} \frac{\partial \Phi(\tilde{x}^\dagger)}{\partial \tilde{x}_i^j}(\tilde{x}_i^j - \tilde{x}_i^{\dagger j}) \geqslant 0. \tag{K.16}$$

Using Lemma 3 in Chapter K.4, we can equivalently write (K.16) in terms of edge flows as follows

$$\sum_{a\in\tilde{\mathcal{E}}} \frac{\partial}{\partial \tilde{w}_a}(\tilde{w}_a^\dagger l_a(\tilde{w}_a^\dagger))(\tilde{w}_a - \tilde{w}_a^\dagger) \geqslant 0 \quad \forall\ \tilde{w} \in \tilde{W}, \tag{K.17}$$

where $w^\dagger$ is the edge flow corresponding to the route flow $x^\dagger$. Comparing (K.15) with (K.17), we note that $\tilde{w}^*(p^\dagger)$ is the minimizer of social cost function $\tilde{\mathbf{\Phi}}$. Therefore, $\tilde{w}^*(\tilde{p}^\dagger) = \tilde{w}^\dagger$.

The proof that $\tilde{\mathbf{P}}^\dagger$ is a singleton follows by contradiction, which is analogous to that in Proposition 12.2.1. We omit the details for the sake of brevity.

Proof of Proposition 12.3.4

The proof follows by verifying the requirements (R1)-(R2) in Proposition 12.2.3. Requirement (R2) holds since the strategy space is a compact set. It suffices to show that requirement (R1) holds. Towards this goal, we define a Lyapunov function candidate $V(\tilde{p}) = (\tilde{p} - \tilde{p}^\dagger)^\top \Delta(\tilde{p} - \tilde{p}^\dagger)$ for the dynamical system (12.38), where $\Delta \in \mathbb{R}^{|\tilde{\mathcal{E}}|\times|\tilde{\mathcal{E}}|}$ is a diagonal matrix defined in (12.37). Due to the strict monotonicity and convexity of $l_a(\cdot)$, it follows that $\Delta_{a,a} > 0$ for every $a \in \tilde{\mathcal{E}}$. Consequently, the Lyapunov function candidate is positive definite.

We show that there exists a positive scalar r such that for any $\tilde{p} \in \mathcal{B}_r(\tilde{p}^\dagger)$, the following holds:

$$\sum_{a\in\tilde{\mathcal{E}}} \nabla_{\tilde{p}_a} V(\tilde{p})^\top \left(\tilde{w}_a^*(\tilde{p})\nabla l_a(\tilde{w}_a^*(\tilde{p})) - \tilde{p}_a\right) < -2V(\tilde{p}). \tag{K.18}$$

Indeed, we note that

$$\begin{aligned}
&\sum_{a\in\tilde{\mathcal{E}}} \nabla_{\tilde{p}_a} V(\tilde{p}) \left(\tilde{w}_a^*(\tilde{p})\nabla l_a(\tilde{w}_a^*(\tilde{p})) - \tilde{p}_a\right) \\
&= 2\sum_{a\in\tilde{\mathcal{E}}} \Delta_{a,a}(\tilde{p}_a - \tilde{p}_a^\dagger) \left(\tilde{w}_a^*(\tilde{p})\nabla l_a(\tilde{w}_a^*(\tilde{p})) - \tilde{p}_a\right) \\
&= 2\sum_{a\in\tilde{\mathcal{E}}} \Delta_{a,a}(\tilde{p}_a - \tilde{p}_a^\dagger) \left(\tilde{w}_a^*(\tilde{p})\nabla l_a(\tilde{w}_a^*(\tilde{p})) - \tilde{p}_a^\dagger + \tilde{p}_a^\dagger - \tilde{p}_a\right) \\
&= -2V(\tilde{p}) + 2\sum_{a\in\tilde{\mathcal{E}}} \Delta_{a,a}(\tilde{p}_a - \tilde{p}_a^\dagger) \left(\phi_a(\tilde{p}) - \phi_a(\tilde{p}^\dagger)\right),
\end{aligned}$$

where for every $a \in \tilde{\mathcal{E}}$, $\phi_a(\tilde{p}) := \tilde{w}_a^*(\tilde{p})\nabla l_a(\tilde{w}_a^*(\tilde{p}))$. Thus, to show local convergence, it suffices to show that there exists $r > 0$ such that

$$\sum_{a\in\tilde{\mathcal{E}}} \Delta_{a,a}(\tilde{p}_a - \tilde{p}_a^\dagger) \left(\phi_a(\tilde{p}) - \phi_a(\tilde{p}^\dagger)\right) \leqslant 0, \quad \forall \tilde{p} \in \mathcal{B}_r(\tilde{p}^\dagger).$$

To show this, we note that due to condition (12.36), the function ϕ is differentiable in a neighborhood of $\tilde{p}^\dagger$ (cf. [435, Chapter 4]). Consequently, using Lemma 5 in Chapter K.4, it is sufficient to show that

$$\sum_{a,a' \in \tilde{\mathcal{E}}} z_a \Delta_{a,a} \frac{\partial \phi_a(\tilde{p}^\dagger)}{\partial \tilde{p}_{a'}} z_{a'} \leqslant 0, \quad \forall\, z \in \mathbb{R}^{|\tilde{\mathcal{E}}|}. \tag{K.19}$$

Indeed, by the design of Δ, it holds that

$$\Delta_{a,a} \frac{\partial \phi_a(\tilde{p}^\dagger)}{\partial \tilde{p}_{a'}} = \frac{\partial \tilde{w}_a^*(\tilde{p}^\dagger)}{\partial \tilde{p}_{a'}}, \quad \forall\, a, a' \in \tilde{\mathcal{E}}. \tag{K.20}$$

Furthermore, Lemma 4 and Lemma 5 in Chapter K.4 guarantee that

$$\sum_{a,a' \in \tilde{\mathcal{E}}} z_a \frac{\partial \tilde{w}_a^*(\tilde{p}^\dagger)}{\partial \tilde{p}_{a'}} z_{a'} \leqslant 0, \quad \forall\, z \in \mathbb{R}^{|\tilde{\mathcal{E}}|}. \tag{K.21}$$

The proof concludes by noting that (K.20) and (K.21) imply (K.19).

K.4 Auxiliary Results

Lemma 1. *Requirement (R1) of Proposition 12.2.3 is satisfied if either one of the following conditions holds:*

(C1) $\frac{\partial e_i(x^*(p))}{\partial p_j} > 0$ *for all* $p \in \mathbb{R}^n$ *and all* $i \neq j$*, and at least one of the following conditions holds:*

- *(i)* $e_i(x^*(0)) \geqslant 0$ *for every* $i \in I$*,* $p^\dagger \in \mathbb{R}_+^{|I|}$*, and for any* $p \in \mathbb{R}_+^{|I|}$*, there exists* $p' \in \mathbb{R}_+^{|I|}$ *such that* $p'_i > p_i$ *and* $e_i(x^*(p')) - p'_i \leqslant 0$ *for every* $i \in I$*. Moreover,* $x_0 \in X, p_0 \in \mathbb{R}_+^{|I|}$*.*
- *(ii)* $e_i(x^*(0)) \leqslant 0$ *for every* $i \in I$*,* $p^\dagger \in \mathbb{R}_-^{|I|}$*, and for any* $p \in \mathbb{R}_-^{|I|}$*, there exists* $p' \in \mathbb{R}_-^{|I|}$ *such that* $p'_i < p_i$ *and* $e_i(x^*(p')) - p'_i \geqslant 0$ *for every* $i \in I$*. Moreover,* $x_0 \in X, p_0 \in \mathbb{R}_-^{|I|}$*.*

(C2) There exists a set $\mathsf{dom}(V) \subset \mathbb{R}^{|I|}$ *and a continuously differentiable function* $V : \mathsf{dom}(V) \to \mathbb{R}_+$ *such that* $V(p^\dagger) = 0$ *and* $V(p) > 0$ *for all* $p \neq p^\dagger$*. Moreover, for every* $p \neq p^\dagger$*,* $\nabla V(p)^\top \left(e(x^*(p)) - p\right) < 0$*.*

Proof. Conditions (C1) and (C2) above are based on results from non-linear dynamical systems which ensure convergence of (12.19). In particular, (C1)-(i) (resp. (C1)-(ii)) builds on cooperative dynamical systems theory [181], which ensures that $\mathbb{R}_+^{|I|}$ (resp. $\mathbb{R}_-^{|I|}$) is positively invariant for (12.19) and $p^\dagger \in \mathbb{R}_+^{|I|}$ (resp. $p^\dagger \in \mathbb{R}_-^{|I|}$) is asymptotically stable. On the other hand, condition (C2) ensures the existence of a Lyapunov function that is strictly positive everywhere except at $p^\dagger$ and decreases along any trajectory of (12.19) (cf. [366]). □

Lemma 2. *For every* $i \in \tilde{I}, j \in \boldsymbol{R}_i$, $\tilde{e}_i^j(\tilde{x}) = \sum_{a \in j} \tilde{w}_a \nabla l_a(\tilde{w}_a)$.

Proof. Using (12.34), we note that

$$\tilde{e}_i^j(\tilde{x}) = \sum_{i' \in \tilde{I}} \sum_{j' \in \mathbf{R}_i} \tilde{x}_{i'}^{j'} \frac{\partial \tilde{\ell}_{i'}^{j'}(\tilde{x})}{\partial \tilde{x}_i^j} \overset{(a)}{=} \sum_{i' \in \tilde{I}} \sum_{j' \in \mathbf{R}_i} \tilde{x}_{i'}^{j'} \sum_{a \in \tilde{\mathcal{E}}} \mathbb{1}(a \in j') \nabla l_a(\tilde{w}_a) \frac{\partial \tilde{w}_a}{\partial \tilde{x}_i^j}$$
$$\overset{(b)}{=} \sum_{i' \in \tilde{I}} \sum_{j' \in \mathbf{R}_i} \tilde{x}_{i'}^{j'} \sum_{a \in \tilde{\mathcal{E}}} \mathbb{1}(a \in j') \nabla l_a(\tilde{w}_a) \mathbb{1}(a \in j) \overset{(c)}{=} \sum_{a \in j} \nabla l_a(\tilde{w}_a) \tilde{w}_a,$$

where (a) follows by expanding out the expression of route costs in terms of edge costs and using the chain rule, (b) follows by the definition of edge flows, and (c) follows by changing the order of summations and using the definition of edge flows. This completes the proof. □

Lemma 3. $x^\dagger$ *that satisfies* (K.16) *if and only if* (K.17).

Proof. First, we show that $\boldsymbol{\Phi}(\tilde{x}) = \sum_{a \in \tilde{\mathcal{E}}} \tilde{w}_a l_a(\tilde{w}_a)$.

$$\boldsymbol{\Phi}(\tilde{x}) \overset{(12.33)}{=} \sum_{i \in \tilde{I}} \sum_{j \in \mathbf{R}_i} \tilde{x}_i^j \tilde{\ell}_i^j(\tilde{x}) = \sum_{i \in \tilde{I}} \sum_{j \in \mathbf{R}_i} \tilde{x}_i^j \sum_{a \in \tilde{\mathcal{E}}} \mathbb{1}(a \in j) l_a(\tilde{w}_a) = \sum_{a \in \tilde{\mathcal{E}}} l_a(\tilde{w}_a) \tilde{w}_a.$$

Next, observe that

$$\sum_{i \in \tilde{I}} \sum_{j \in \mathbf{R}_i} \frac{\partial \boldsymbol{\Phi}(\tilde{x}^\dagger)}{\partial \tilde{x}_i^j} (\tilde{x}_i^j - \tilde{x}_i^{\dagger j}) = \sum_{i \in \tilde{I}} \sum_{j \in \mathbf{R}_i} \sum_{a \in \tilde{\mathcal{E}}} \frac{\partial}{\partial \tilde{x}_i^j} \left(\tilde{w}_a l_a(\tilde{w}_a)\right) (\tilde{x}_i^j - \tilde{x}_i^{\dagger j})$$
$$= \sum_{i \in \tilde{I}} \sum_{j \in \mathbf{R}_i} \sum_{a \in \tilde{\mathcal{E}}} \frac{\partial}{\partial \tilde{w}_a} \left(\tilde{w}_a l_a(\tilde{w}_a)\right) \mathbb{1}(a \in j) (\tilde{x}_i^j - \tilde{x}_i^{\dagger j}) = \sum_{a \in \tilde{\mathcal{E}}} \frac{\partial}{\partial \tilde{w}_a} \left(\tilde{w}_a l_a(\tilde{w}_a)\right) (\tilde{w}_a - \tilde{w}_a^\dagger).$$

This concludes the proof. □

Lemma 4. *Following inequality holds:*

$$\sum_{a \in \tilde{\mathcal{E}}} (\tilde{p}_a - \tilde{p}'_a) \left(\tilde{w}_a^*(\tilde{p}_a) - \tilde{w}_a^*(\tilde{p}'_a)\right) \leqslant 0, \quad \forall\ \tilde{p}, \tilde{p}' \in \mathbb{R}^{|\tilde{\mathcal{E}}|}. \tag{K.22}$$

Proof. To prove this result, we first show that

$$\sum_{i \in \tilde{I}} \sum_{j \in \mathbf{R}_i} (\tilde{P}_i^j - \tilde{P}_i^{\prime j})(\tilde{x}_i^{*j}(\tilde{P}) - \tilde{x}_i^{*j}(\tilde{P}')) \leqslant 0, \tag{K.23}$$

where $\tilde{P}$ and $\tilde{P}'$ are the route tolls associated with edge tolls $\tilde{p}$ and $\tilde{p}'$, respectively, through (12.32). Let the feasible set of route flows in the optimization problem (K.13) be denoted by $\mathcal{F}$. Using the first-order conditions of optimality for the strictly convex optimization problem (K.13), we obtain:

$$\sum_{i \in \tilde{I}} \sum_{j \in \mathbf{R}_i} \left(\tilde{c}_i^j(\tilde{x}^*(\tilde{P}), \tilde{P})\right) \cdot \left(\tilde{y}_i^j - \tilde{x}_i^{*j}(\tilde{P})\right) \geqslant 0, \forall \quad \tilde{y} \in \mathcal{F}, \tag{K.24}$$

where $\tilde{P}$ is the route toll associated with edge toll $\tilde{p}$. Rewriting (K.24) for edge tolls $\tilde{p}'$ we obtain

$$\sum_{i\in\tilde{I}}\sum_{j\in\mathbf{R}_i}\left(\tilde{c}_i^j(\tilde{x}^*(\tilde{P}'),\tilde{P}')\right)\cdot\left(\tilde{y}_i'^j-\tilde{x}_i^{*j}(\tilde{P}')\right)\geqslant 0,\forall\quad \tilde{y}'\in\mathcal{F}, \tag{K.25}$$

where $\tilde{P}'$ is the route toll associated with edge toll $\tilde{p}'$.

Next, we prove (K.23). Note that

$$\begin{aligned}
&\sum_{i\in\tilde{I}}\sum_{j\in\mathbf{R}_i}(\tilde{P}_i^j-\tilde{P}_i'^j)(\tilde{x}_i^{*j}(\tilde{P})-\tilde{x}_i^{*j}(\tilde{P}'))\overset{(a)}{\leqslant}\sum_{i\in\tilde{I}}\sum_{j\in\mathbf{R}_i}(\tilde{\ell}_i^j(\tilde{x}^*(\tilde{P}'))-\tilde{\ell}_i^j(\tilde{x}^*(\tilde{P})))(\tilde{x}_i^{*j}(\tilde{P})-\tilde{x}_i^{*j}(\tilde{P}'))\\
&\overset{(b)}{=}\sum_{i\in\tilde{I}}\sum_{j\in\mathbf{R}_i}(\tilde{x}_i^{*j}(\tilde{P})-\tilde{x}_i^{*j}(\tilde{P}'))\cdot\sum_{a\in\tilde{\mathcal{E}}}(l_a(\tilde{w}_a^*(\tilde{p}'))-l_a(\tilde{w}_a^*(\tilde{p})))\mathbb{1}(a\in j)\\
&\overset{(c)}{=}\sum_{a\in\tilde{\mathcal{E}}}(l_a(\tilde{w}_a^*(\tilde{p}'))-l_a(\tilde{w}_a^*(\tilde{p})))\cdot\sum_{i\in\tilde{I}}\sum_{j\in\mathbf{R}_i}(\tilde{x}_i^{*j}(\tilde{P})-\tilde{x}_i^{*j}(\tilde{P}'))\mathbb{1}(a\in j)\\
&\overset{(d)}{=}\sum_{a\in\tilde{\mathcal{E}}}(l_a(\tilde{w}_a^*(\tilde{p}'))-l_a(\tilde{w}_a^*(\tilde{p})))(\tilde{w}_a^*(\tilde{p})-\tilde{w}_a^*(\tilde{p}'))\overset{(e)}{\leqslant}0,
\end{aligned}$$

where we obtain (a) by adding (K.24), evaluated at $\tilde{y}=\tilde{x}^*(\tilde{P}')$, and (K.25), evaluated at $\tilde{y}'=\tilde{x}^*(\tilde{P})$, (b) holds by the definition of the route loss function, (c) holds by interchange of summation, (d) holds by the definition of edge flows, and (e) holds due to the monotonicity of edge latency functions. This proves (K.23).

Finally, we prove (K.22). Note that

$$\begin{aligned}
&\sum_{a\in\tilde{\mathcal{E}}}(\tilde{p}_a-\tilde{p}_{a'})(\tilde{w}_a^*(\tilde{p})-\tilde{w}_a^*(\tilde{p}'))\overset{(a)}{=}\sum_{a\in\tilde{\mathcal{E}}}(\tilde{p}_a-\tilde{p}_{a'})\sum_{i\in\tilde{I}}\sum_{j\in\mathbf{R}_i}(\tilde{x}_i^{*j}(\tilde{P})-\tilde{x}_i^{*j}(\tilde{P}'))\mathbb{1}(a\in j)\\
&\overset{(b)}{=}\sum_{i\in\tilde{I}}\sum_{j\in\mathbf{R}_i}(\tilde{x}_i^{*j}(\tilde{P})-\tilde{x}_i^{*j}(\tilde{P}'))\sum_{a\in\tilde{\mathcal{E}}}(\tilde{p}_a-\tilde{p}_{a'})\mathbb{1}(a\in j)\\
&\overset{(c)}{=}\sum_{i\in\tilde{I}}\sum_{j\in\mathbf{R}_i}(\tilde{x}_i^{*j}(\tilde{P})-\tilde{x}_i^{*j}(\tilde{P}'))(\tilde{P}_i^j-\tilde{P}_i'^j)\overset{(d)}{\leqslant}0,
\end{aligned}$$

where (a) holds due to the definition of edge flows, (b) holds due to interchange of summation, (c) holds due to the definition of route tolls, and (d) holds due to (K.23). This concludes the proof. □

Lemma 5 ([135])**.** *For any fixed p' and continuously differentiable function $\phi:\mathbb{R}^e\to\mathbb{R}^{\tilde{\mathcal{E}}}$, the condition*

$$\langle\phi(p)-\phi(p'),p-p'\rangle\leqslant 0\quad\forall\ p\in\mathcal{B}_r(p')$$

for some $r>0$, holds if and only if

$$\sum_{i,j\in|\tilde{\mathcal{E}}|}z_iz_j\frac{\partial\phi_i(p')}{\partial p_j}\leqslant 0,\quad\forall\ z\in\mathbb{R}^{|\tilde{\mathcal{E}}|}.$$

Appendix L

Appendix for Chapter 14

Section L.1 presents a simple example to illustrate the time-extended graph and other constraints in the optimization problems discussed in Chapter 14. Section L.2 contains the proofs of all theoretical results discussed in Chapter 14. A detailed explanation of the ADMM formulation is provided in Section L.3. Section L.4 explores an additional AAM scenario involving vertiport reservation for air-taxi services in Northern California.

L.1 A Simple Example

In this section, we explain the time-extended graph (Definition 14.1.1) along with constraints (14.2b)-(14.2c) through a simple example comprising of 3 regions, denoted by $\{A, B, C\}$.

Time-extended graph The time-extended graph (for $T = 5$) corresponding to our scenario is shown in Figure L.1. In the time-extended graph $\tilde{\mathcal{G}} = (\tilde{\mathcal{R}}, \tilde{\mathcal{E}})$, at every time t each region r is replicated into three regions $t : v(r,t)$, $v^{\mathsf{arr}}(r,t)$, $v^{\mathsf{dep}}(r,t)$. For conciseness, we will only discuss one edge corresponding to each of the four types $\tilde{\mathcal{E}}^{(1)}, \tilde{\mathcal{E}}^{(2)}, \tilde{\mathcal{E}}^{(3)}, \tilde{\mathcal{E}}^{(4)}$. The (red) edge $(v^{\mathsf{arr}}(B,2), v(B,2))$ is an edge of type $\tilde{\mathcal{E}}^{(1)}$. The (red) edge $(v(A,1), v^{\mathsf{dep}}(A,1))$ is an edge of type $\tilde{\mathcal{E}}^{(2)}$. The (black) edge $(v(A,1), v(A,2))$ is an edge of type $\tilde{\mathcal{E}}^{(3)}$. The (green) edge $(v^{\mathsf{dep}}(A,2), v^{\mathsf{arr}}(B,3))$ is an edge of type $\tilde{\mathcal{E}}^{(4)}$.

Constraint (14.2b)-(14.2c) Here, we illustrate the constraints (14.2b)-(14.2c) through an example. Consider an AAM vehicle u that wants to travel from region A to region C. Suppose

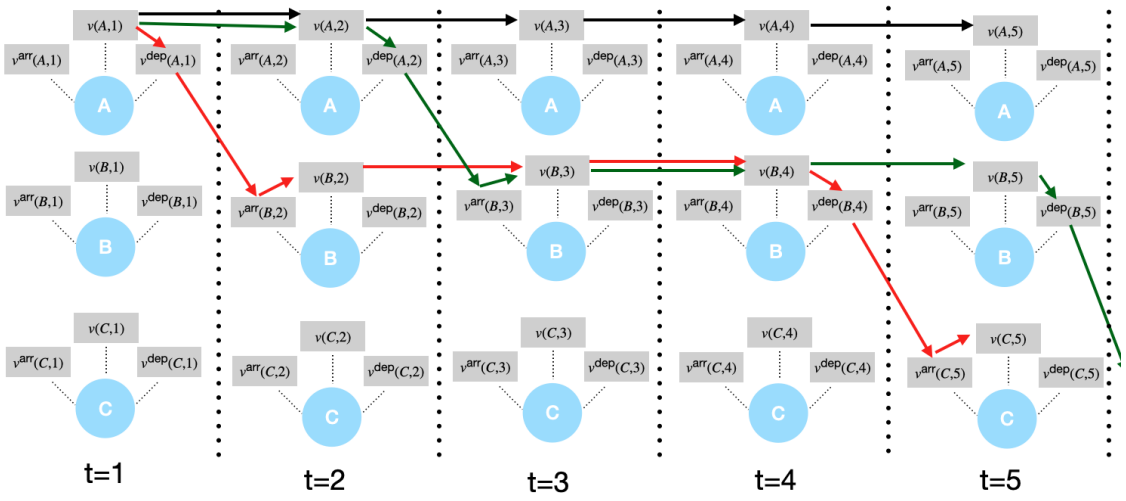


Figure L.1: Time Extended Graph: From left to right, we show a sequence of time steps and different color-coded trajectories that an AAM vehicle can request. The red trajectory shows an AAM vehicle traveling from region A to C while transiting from region B. The green trajectory represents the same trajectory as the red path but is delayed by one unit of time. The black trajectory denotes an option where the AAM vehicle stays parked at the origin region. To simplify the visualization, we have not shown all possible edges on this time-extended graph.

the menu of that AAM vehicle is comprised of the routes s_1, s_2, s_3, as described below:

$$s_1 = \{(v(A,1), v(A,2)), (v(A,2), v(A,3)), (v(A,3), v(A,4)), (v(A,4), v(A,5))\}$$
Black-path in Fig. L.1

$$s_2 = \{(v(A,1), v^{\mathsf{dep}}(A,1)), (v^{\mathsf{dep}}(A,1), v^{\mathsf{arr}}(B,2)), (v^{\mathsf{arr}}(B,2), v(B,2)), (v(B,2), v(B,3)),$$
$$(v(B,3), v(B,4)), (v(B,4), v^{\mathsf{dep}}(B,4)), (v^{\mathsf{dep}}(B,4), v^{\mathsf{arr}}(C,5)),$$
$$(v^{\mathsf{arr}}(C,5), v(C,5))\}\text{Red-path in Fig. L.1}$$

$$s_3 = \{(v(A,1), v(A,2)), (v(A,2), v^{\mathsf{dep}}(A,2)), (v^{\mathsf{dep}}(A,2), v^{\mathsf{arr}}(B,3)), (v^{\mathsf{arr}}(B,3), v(B,3)),$$
$$(v(B,3), v(B,4)), (v(B,4), v(B,5)), (v(B,5), v^{\mathsf{dep}}(B,5)), (v^{\mathsf{dep}}(B,5), v^{\mathsf{arr}}(C,6))\}$$
Green-path in Fig. L.1

Here

$$e^*(s_1) = (v(A,1), v(A,2)),$$
$$e^*(s_2) = (v(A,1), v(A,1)^{\mathsf{dep}}), e^*(s_3) = (v(A,2), v^{\mathsf{dep}}(A,2)).$$

The menu M_u of AAM vehicle u is given by $M_u = \{s_1, s_2, s_3, \varnothing\}$.

Consequently, the constraint (14.2b) for this AAM vehicle is given by $x_{u,e^*(s_1)} + x_{u,e^*(s_2)} + x_{u,e^*(s_3)} + x_{u,\varnothing} = 1$. Additionally, the constraint (14.2c) for this AAM vehicle contains two

types of constraints: (1) the flow balance constraints:

$$
\begin{aligned}
&x_{u,(v(A,1),v(A,2))} = x_{u,(v(A,2),v^{\mathsf{dep}}(A,2))} + x_{u,(v(A,2),v(A,3))}\\
&x_{u,(v(A,2),v(A,3))} = x_{u,(v(A,3),v(A,4))}\\
&x_{u,(v(A,3),v(A,4))} = x_{u,(v(A,4),v(A,5))}\\
&x_{u,(v(A,2),v^{\mathsf{dep}}(A,2))} = x_{u,(v^{\mathsf{dep}}(A,2),v^{\mathsf{arr}}(B,3)}\\
&x_{u,(v^{\mathsf{dep}}(A,2),v^{\mathsf{arr}}(B,3)} = x_{u,(v^{\mathsf{arr}}(B,3),v(B,3))}\\
&x_{u,(v^{\mathsf{arr}}(B,3),v(B,3))} + x_{u,(v(B,2),v(B,3))} = x_{u,(v(B,3),v(B,4))}\\
&x_{u,(v(B,3),v(B,4))} = x_{u,(v(B,4),v(B,5))} + x_{u,(v(B,4),v^{\mathsf{dep}}(B,4))}\\
&x_{u,(v(B,4),v(B,5))} = x_{u,(v(B,2),v^{\mathsf{dep}}(B,5))}\\
&x_{u,(v(B,4),v^{\mathsf{dep}}(B,4))} = x_{u,(v^{\mathsf{dep}}(B,4),v^{\mathsf{arr}}(C,5))}\\
&x_{u,(v^{\mathsf{dep}}(B,4),v^{\mathsf{arr}}(C,5))} = x_{u,(v^{\mathsf{arr}}(C,5),v(C,5))}\\
&x_{u,(v^{\mathsf{dep}}(A,1),v^{\mathsf{arr}}(B,2))} = x_{u,(v^{\mathsf{arr}}(B,2),v(B,2))}\\
&x_{u,(v^{\mathsf{arr}}(B,2),v(B,2))} = x_{u,(v(B,2),v(B,3))}.
\end{aligned}
$$

and (2) additional constraints:

$$
\begin{aligned}
&x_{u,(v(A,1),v^{\mathsf{dep}}(A,1))} = x_{u,(v^{\mathsf{dep}}(B,4),v^{\mathsf{arr}}(C,5))}\\
&x_{u,(v(A,2),v^{\mathsf{dep}}(A,2))} = x_{u,(v^{\mathsf{dep}}(B,5),v^{\mathsf{arr}}(C,6))}.
\end{aligned}
$$

These constraints ensure that the path flows allocated on the departing edge, as per (14.2b), result in unique edge flows on the entire network.

L.2 Proof of Theoretical Results

Proof of Proposition 14.3.1

Before presenting the proof, let us recall some important mathematical definitions and results which are crucial for the proof. First, we recall the definition of upper semicontinuous and lower semicontinuous correspondences.

Definition L.2.1. *A correspondence $f : X \rightrightarrows Y$ is* upper semicontinuous *if for every sequence $x_n \in X$ (with limit x) and the sequence $y_n \in f(x_n)$ which has a limit, then there exists $y \in f(x)$ such that $y = \lim_n y_n$.*

Definition L.2.2. *A correspondence $f : X \rightrightarrows Y$ is* lower semicontinuous *if for every sequence $x_n \in X$ (with limit x) and $y \in f(x)$, then there exists a convergent sequence $y_n \in f(x_n)$ with limit y.*

Next, we recall the Kakutani fixed point theorem.

Theorem L.2.1 (Kakutani Fixed Point Theorem)**.** *Suppose X is a non-empty, convex, and compact subset of $\mathbb{R}^n$ and $f : X \rightrightarrows X$ is a non-empty, closed-valued, convex-valued, and upper semicontinuous correspondence. Then f has a fixed point.*

Finally, we recall Berge's maximum theorem.

Theorem L.2.2 (Berge's Maximum Theorem)**.** *Consider the optimization problem*

$$\max_{x \in A(\theta)} F(x, \theta).$$

Let $X(\theta)$ be the set of solutions of the preceding problem. If F is continuous in (x, θ) and $\theta \rightrightarrows A(\theta)$ is a non-empty, compact-valued, and continuous correspondence, then $X(\theta)$ is a non-empty, compact-valued, and upper semicontinous correspondence.

Proof of Proposition 14.3.1. The proof builds on a result about the existence of a competitive equilibrium in Fisher markets with auxiliary inequality constraints [193]. Particularly, our proof accounts for auxiliary *equality* constraints (resulting due to (14.2b)-(14.2c)).

In this result, we consider a relaxation of (14.2), by converting the integrality constraint to the positivity constraint. Consider the relaxed individual optimization problem of every agent stated below:

$$\begin{aligned}
\max_{\bar{\mathbf{x}}_u} \quad & f_u(\bar{\mathbf{x}}_u) && \text{(L.1a)}\\
\text{s.t.} \quad & \mathbf{p}^\top \mathbf{x}_u + p_{\mathsf{o}} x_{u,\mathsf{o}} = w_u && \text{(L.1b)}\\
& \tilde{\mathbf{a}}_u^\top \mathbf{x}_u + x_{u,\varnothing} = 1 && \text{(L.1c)}\\
& \tilde{\mathbf{A}}_u \mathbf{x}_u = \mathbf{0} && \text{(L.1d)}\\
& x_{u,\mathsf{o}} \geqslant 0, x_{u,\varnothing} \geqslant 0, x_{u,e} \geqslant 0 \quad \forall e \in \tilde{\mathcal{E}}. && \text{(L.1e)}
\end{aligned}$$

To prove the existence result, we scale the problem such that the total budget of all agents is 1 and the capacity of each good is 1. To do this, for every $e \in \tilde{\mathcal{E}}, u \in U$, we scale any allocation $x_{u,e}$ to $x_{u,e}/\ell_e$, scale $\tilde{\mathbf{a}}_{u,e}$ to $\tilde{\mathbf{a}}_{u,e} \cdot \ell_e$, scale $\tilde{\mathbf{A}}_u[:, e]$ to $\tilde{\mathbf{A}}_u[:, e] \cdot \ell_e$, p_e to $p_e \ell_e / W$, and w_u to w_u / W, where $W = \sum_u w_u$. Note that under this change the solution of (L.1) does not change. Furthermore, due to the condition that $v_{u,\mathsf{o}} \geqslant 0$ and the variable $x_{u,\mathsf{o}}$ does not enter in the constraint (L.1c)-(L.1d), it is ensured that the budget constraint (L.1b) hold with equality.

Define $\Delta_{|\tilde{\mathcal{E}}|} = \{\mathbf{p} \in \mathbb{R}^{|\tilde{\mathcal{E}}|} : \sum_{e \in \tilde{\mathcal{E}}} p_e = 1, p_e \geqslant 0 \ \forall e \in \tilde{\mathcal{E}}\}$. Moreover, for every UAV u, define $Y_u = \{\bar{\mathbf{x}}_u \in \mathbb{R}_{\geqslant 0}^{|\tilde{\mathcal{E}}|+2} : \tilde{\mathbf{a}}_u^\top \mathbf{x}_u + x_{u,\varnothing} = 1, \quad \bar{\mathbf{A}}_u \mathbf{x}_u = 0\}$, and $Q_u = \{\bar{\mathbf{x}}_u \in \mathbb{R}_{\geqslant 0}^{|\tilde{\mathcal{E}}|+2} : x_{u,r} \leqslant \Omega, \quad \forall\, r \in \tilde{\mathcal{E}} \cup \{\mathsf{o}, \varnothing\}\}$ for some $\Omega > 1$. Define $X = \Pi_{u \in U} Q_u$.

Define $B_u(\mathbf{p}) = \{\bar{\mathbf{x}}_u \in Y_i : \mathbf{p}^\top \mathbf{x}_u + p_{\mathsf{o}} x_{u,\mathsf{o}} = w_u\}$. Note that this set is non-empty, so we can always choose $x_{u,\varnothing}$ to ensure that $\mathbf{x}_u = \mathbf{0}$ and spend all the budget in the outside

option o. Define

$$\tilde{\mathbf{x}}_u(\mathbf{p}) = \underset{\bar{\mathbf{x}}_u \in Q_u \cap B_u(\mathbf{p})}{\arg\max} \; f_u(\bar{\mathbf{x}}_u), \tag{L.2}$$

$$\tilde{\mathbf{p}}(\mathbf{x}) = \underset{\mathbf{p} \in \Delta_{|\tilde{\mathcal{E}}|}}{\arg\max} \; \mathbf{p}^\top \left(\sum_{u \in U} \mathbf{x}_u - \mathbf{1} \right). \tag{L.3}$$

Using the above definitions, define a correspondence $h(\mathbf{x}, \mathbf{p}) = ((\tilde{\mathbf{x}}_u(\mathbf{p}))_{u \in U}, \tilde{\mathbf{p}}(\mathbf{x}))$. We shall show that a fixed point of this mapping exists and is a fractional competitive equilibrium.

Existence of a Fixed Point We show that h satisfies the condition of the Kakutani fixed point theorem (cf. Theorem L.2.1), which ensures the existence of a fixed point. First, note that the domain of h, i.e. $X \times \Delta_{|\tilde{\mathcal{E}}|}$, is non-empty, compact and convex.

Next, we show that h is a non-empty, closed-valued, convex-valued, and upper semi-continuous correspondence. It is enough to show that $\tilde{\mathbf{x}}_u(\mathbf{p})$ and $\tilde{\mathbf{p}}(\mathbf{x})$ are non-empty, convex-valued and upper semicontinuous correspondences.

From (L.3), we observe that $\tilde{\mathbf{p}}(\mathbf{x})$ is non-empty and convex-valued and is an optimal solution to a linear program with a non-empty, convex, and compact feasible set. All conditions for Berge's maximum theorem (cf. Theorem L.2.2) are satisfied, and therefore $\tilde{\mathbf{p}}(\mathbf{x})$ is also compact-valued and upper semicontinuous.

Next, we show that $\tilde{\mathbf{x}}_u(\mathbf{p})$ is non-empty and convex-valued as it is the optimizer of a linear function on a non-empty, convex, and compact set. Next, we leverage Theorem L.2.2 to show that this map is compact-valued and upper semicontinuous. First, we need to show that the correspondence $g_u : \mathbf{p} \rightrightarrows Q_u \cap B_u(\mathbf{p})$ is a compact-valued and continuous correspondence. Compactness follows by construction, so the only thing remaining to show is continuity. To show continuity, it is enough to show that the mapping is upper semicontinuous and lower semicontinuous.

To show g_u is upper semicontinuous, consider a sequence $(\bar{\mathbf{x}}_u^n, \mathbf{p}^n)$ such that $\bar{\mathbf{x}}_u^n \in Q_u \cap B_u(\mathbf{p}^n)$, which has limit $(\bar{\mathbf{x}}_u, \mathbf{p})$. Then, it is sufficient to establish that $\bar{\mathbf{x}}_u \in Q_u \cap B_u(\mathbf{p})$. Note that Q_u is compact, so if $\bar{\mathbf{x}}_u^n \in Q_u$, for every $n \in \mathbb{N}$, it follows that $\bar{\mathbf{x}}_u \in Q_u$. Furthermore, since $\bar{\mathbf{x}}_u^n \geqslant 0$, for every $n \in \mathbb{N}$, it follows that $\bar{\mathbf{x}}_u \geqslant 0$. Additionally, for every $n \in \mathbb{N}$, $\tilde{\mathbf{a}}_u^\top \mathbf{x}_u^n + x_{u,\varnothing}^n = 1, \quad \tilde{\mathbf{A}}_u \mathbf{x}_u^n = \mathbf{0}$, it follows that $\tilde{\mathbf{a}}_u^\top \mathbf{x}_u + x_{u,\varnothing} = 1, \quad \tilde{\mathbf{A}}_u \mathbf{x}_u = \mathbf{0}$. Moreover, the continuity of product ensures that $\mathbf{p}^{n\top} \mathbf{x}_u^n + p_{\mathsf{o}} x_{u,\mathsf{o}}^n = w_u$, for every $n \in \mathbb{N}$, implies $\mathbf{p}^\top \mathbf{x}_u + p_{\mathsf{o}} x_{u,\mathsf{o}} = w_u$. This ensures that g_u is upper semicontinuous.

Next, we show that g_u is lower semicontinuous. To show this, it is sufficient to show that for any sequence $\mathbf{p}^n$ with limit $\mathbf{p}$ and any point $\bar{\mathbf{x}}_u \in Q_u \times B_u(\mathbf{p})$ there is a sequence $\bar{\mathbf{x}}_u^n \in Q_u \cap B_u(\mathbf{p}^n)$ such that $\lim_{n \to \infty} \bar{\mathbf{x}}_u^n = \bar{\mathbf{x}}_u$. Towards this goal, for every $u \in U, e \in \tilde{\mathcal{E}}$, we define $\bar{\mathbf{x}}^n$ such that

$$x_{u,e}^n = \min \left\{ 1, \frac{w_u}{\mathbf{p}^{n\top} \mathbf{x}_u + p_{\mathsf{o}} x_{u,\mathsf{o}}} \right\} x_{u,e}, \quad x_{u,\varnothing}^n = 1 - \tilde{\mathbf{a}}_u^\top \mathbf{x}_u^n, \quad x_{u,\mathsf{o}}^n = \frac{1}{p_{\mathsf{o}}} \left(w_u - \mathbf{p}^{n\top} \mathbf{x}_u^n \right).$$

It is easy to check that the $\lim_{n\to\infty} \bar{\mathbf{x}}_u^n = \bar{\mathbf{x}}_u$. It remains to demonstrate that $\bar{\mathbf{x}}_u^n \in Q_u \cap B_u(\mathbf{p}^n)$. First, note that $\tilde{\mathbf{a}}_u^\top \mathbf{x}_u^n + x_{u,\varnothing}^n = 1$ follows by construction. Next, we show that $x_{u,\varnothing}^n \geqslant 0$. This is because

$$\begin{aligned}\tilde{\mathbf{a}}_u^\top \mathbf{x}_u^n &= \min\left\{1, \frac{w_u}{\mathbf{p}^{n\top}\mathbf{x}_u + p_{\mathsf{o}} x_{u,\mathsf{o}}}\right\} \tilde{\mathbf{a}}_u^\top \mathbf{x}_u \leqslant \tilde{\mathbf{a}}_u^\top \mathbf{x}_u = 1 - x_{u,\varnothing} \\ &\implies x_{u,\varnothing}^n = 1 - \tilde{\mathbf{a}}_u^\top \mathbf{x}_u^n \geqslant x_{u,\varnothing} \geqslant 0,\end{aligned}$$

where the inequality follows as $\tilde{a}_{u,e} \geqslant 0, x_{u,e}^n \geqslant 0$. Similarly, one can show that $\tilde{\mathbf{A}}_u \mathbf{x}_u^n = \mathbf{0}$. Next, we note that budget constraints are satisfied by the construction of $x_{u,\mathsf{o}}^n$. Finally, we show that $x_{u,\mathsf{o}}^n \geqslant 0$. Indeed,

$$\mathbf{p}^{n\top}\mathbf{x}_u^n = \min\left\{1, \frac{w_u}{\mathbf{p}^{n\top}\mathbf{x}_u + p_{\mathsf{o}} x_{u,\mathsf{o}}}\right\} p^{n\top}\mathbf{x}_u \leqslant w_u.$$

Thus, we conclude that g_u is a compact-valued continuous correspondence. Thus, from Theorem L.2.1, we conclude that there exists $(\bar{\mathbf{x}}^*, \mathbf{p}^*)$ such that $\bar{\mathbf{x}}_u^* = \tilde{\mathbf{x}}_u(\mathbf{p}^*), \quad \mathbf{p}^* = \tilde{\mathbf{p}}(\bar{\mathbf{x}}^*), \quad \forall\, u \in U$.

Existence of a Fractional Competitive Equilibrium We show that any fixed point corresponds to a fractional competitive equilibrium. First, using (L.2), we conclude that $\bar{\mathbf{x}}_u^*$ is an optimal solution to (L.1). Second, note that $\mathbf{p}^* \in \mathbb{R}_{\geqslant 0}^{|\tilde{\mathcal{E}}|}$ by construction. Next, we show that the capacity constraints are satisfied. We show this by contradiction. Suppose there exists an edge $e' \in \tilde{\mathcal{E}}$ such that $\sum_{u\in U} x_{u,e'}^* > 1$. Then by (L.3), it must hold that

$$\sum_{e\in\tilde{\mathcal{E}}} p_e^* \left(\sum_{u\in U} x_{u,e}^* - 1\right) \geqslant \sum_{e\in\tilde{\mathcal{E}}} p_e \left(\sum_{u\in U} x_{u,e}^* - 1\right), \quad \forall\, \mathbf{p} \in \Delta_{|\tilde{\mathcal{E}}|}.$$

We claim that $\sum_{e\in\tilde{\mathcal{E}}} p_e^*(\sum_{u\in U} x_{u,e}^* - 1) = 0$. Indeed,

$$\sum_{e\in\tilde{\mathcal{E}}} p_e^*(\sum_{u\in U} x_{u,e}^* - 1) = \sum_{u\in U}\sum_{e\in\tilde{\mathcal{E}}} p_e^* x_{e,u}^* - \sum_{e\in\tilde{\mathcal{E}}} p_e^* = \sum_{u\in U} w_u - 1 = 0.$$

Thus, we conclude that

$$0 \geqslant \sum_{e\in\tilde{\mathcal{E}}} p_e \left(\sum_{u\in U} x_{u,e}^* - 1\right), \quad \forall\, \mathbf{p} \in \Delta_{|\tilde{\mathcal{E}}|}. \tag{L.4}$$

Since $\sum_{u\in U} x_{u,e'}^* > 1$, we can select $p_{e'} = 1$ and 0 otherwise, which would violate the above inequality, a contradiction.

Next, we show that if $p_e^* > 0$ then $\sum_{u\in U} x_{u,e}^* = 1$. This follows immediately from the fact that capacity constraints are satisfied and the fact that $\sum_{e\in\tilde{\mathcal{E}}} p_e^*(\sum_{u\in U} x_{u,e}^* - 1) = 0$. This completes the proof. □

Proof of Proposition 14.3.3

Observe that for any fixed value of $\omega \in \mathbb{R}^{|U|}$, the optimization problem (14.3) is a convex optimization problem. Define the Lagrangian as follows.

$$\begin{aligned}\mathcal{L}_{\mathbf{P}} = &\sum_{u\in U}(w_u+\omega_u)\log\left(f_u(\bar{\mathbf{x}}_u)\right) - \sum_{u\in U} p_{\mathsf{o}} x_{u,\mathsf{o}} - \mathbf{p}^\top\left(\sum_{u\in U}\mathbf{x}_u - \ell\right)\\ &- \sum_{u\in U}\lambda_u(\tilde{\mathbf{a}}_u^\top \mathbf{x}_u + x_{u,\varnothing} - 1) - \sum_{u\in U}\kappa_u^\top \tilde{\mathbf{A}}_u \mathbf{x}_u + \sum_{u\in U}\mu_u^\top \bar{\mathbf{x}}_u,\end{aligned}$$

where $\mathbf{p} \in \mathbb{R}^{|\tilde{\mathcal{E}}|}_{\geqslant 0}$ is the Lagrange multiplier corresponding to constraint (14.3b), $\lambda = (\lambda_u)_{u\in U} \in \mathbb{R}^{|U|}$ is the Lagrange multiplier corresponding to (14.3c), $\kappa = (\kappa_u)_{u\in U} \in \mathbb{R}^{K|U|}$ is the Lagrange multiplier corresponding to (14.3d), and $\mu = (\mu_u)_{u\in U} \in \mathbb{R}^{|U||\tilde{\mathcal{E}}|}_{\geqslant 0}$ is the Lagrange multiplier corresponding to (14.3e).

We observe that, for a given ω, any optimal solution $\bar{\mathbf{x}}^\dagger$ of (14.3) with optimal dual multipliers $(\mathbf{p}^\dagger, \lambda^\dagger, \kappa^\dagger, \mu^\dagger)$ will satisfy the following first order conditions of optimality.

$$0 \geqslant \begin{cases} \frac{(w_u+\omega_u)}{f_u(\bar{\mathbf{x}}_u^\dagger)} v_{u,e} - p_e^\dagger - \tilde{a}_{u,e}\lambda_u^\dagger - (\tilde{\mathbf{A}}_u^\top \kappa_u^\dagger)_e & \text{if } e \in \tilde{\mathcal{E}} \\ \frac{(w_u+\omega_u)}{f_u(\bar{\mathbf{x}}_u^\dagger)} v_{u,\mathsf{o}} - p_{\mathsf{o}} & \text{if } e = \mathsf{o} \\ \frac{(w_u+\omega_u)}{f_u(\bar{\mathbf{x}}_u^\dagger)} v_{u,\varnothing} - \lambda_u^\dagger & \text{if } e = \varnothing. \end{cases} \tag{L.5}$$

Furthermore, the complementary slackness conditions are given by

$$\begin{aligned} 0 &= \begin{cases} \frac{(w_u+\omega_u)}{f_u(\bar{\mathbf{x}}_u^\dagger)} v_{u,e} x_{u,e}^\dagger - p_e^\dagger x_{u,e}^\dagger - \tilde{a}_{u,e} x_{u,e}^\dagger \lambda_u^\dagger - (\tilde{\mathbf{A}}_u^\top \kappa_u^\dagger)_e x_{u,e}^\dagger & \text{if } e \in \tilde{\mathcal{E}} \\ \frac{(w_u+\omega_u)}{f_u(\bar{\mathbf{x}}_u^\dagger)} v_{u,\mathsf{o}} x_{u,\mathsf{o}}^\dagger - p_{\mathsf{o}}^\dagger x_{u,\mathsf{o}}^\dagger & \text{if } e = \mathsf{o} \\ \frac{(w_u+\omega_u)}{f_u(\bar{\mathbf{x}}_u^\dagger)} v_{u,\varnothing} x_{u,\varnothing}^\dagger - \lambda_u x_{u,\varnothing}^\dagger & \text{if } e = \varnothing \end{cases}, \\ 0 &= p_e^\dagger\Big(\sum_{u\in U} x_{u,e}^\dagger - \ell_e\Big), \quad \forall\, e \in \tilde{\mathcal{E}}. \end{aligned} \tag{L.6}$$

Similarly, the Lagrangian of the (relaxed) individual optimization problem (L.1) is given by

$$\mathcal{L}_{\mathbf{I}} = f_u(\bar{\mathbf{x}}_u) - \tilde{\omega}_u\left(\mathbf{p}^\top \mathbf{x}_u + p_{\mathsf{o}} x_{u,\mathsf{o}} - w_u\right) - \tilde{\lambda}_u(\tilde{\mathbf{a}}_u^\top \mathbf{x}_u + x_{u,\varnothing} - 1) - \sum_{u\in U}\tilde{\kappa}_u^\top \tilde{\mathbf{A}}_u \mathbf{x}_u + \tilde{\mu}_u^\top \bar{\mathbf{x}}_u,$$

where $\tilde{\omega}_u \in \mathbb{R}$ is the Lagrange multiplier corresponding to the budget constraint, $\tilde{\lambda}_u \in \mathbb{R}$ is the Lagrange multiplier corresponding to (L.1c), $\tilde{\kappa}_u \in \mathbb{R}^K$ is the Lagrange multiplier corresponding to (L.1d), and $\tilde{\mu}_u \in \mathbb{R}^{|\tilde{\mathcal{E}}|}_{\geqslant 0}$ is the Lagrange multiplier corresponding to the positivity constraint (L.1e).

We observe that, for a given $\mathbf{p}$, any optimal solution $\bar{\mathbf{x}}^{\ddagger}$ of (L.1) with optimal dual multipliers $(\tilde{\omega}^{\ddagger}, \tilde{\lambda}^{\ddagger}, \tilde{\kappa}^{\ddagger}, \tilde{\mu}^{\ddagger})$ satisfies the following first order conditions of optimality.

$$0 \geqslant \begin{cases} v_{u,e} - \tilde{\omega}_u^{\ddagger} p_e - \tilde{\lambda}_u^{\ddagger} \tilde{a}_{u,e} - (\tilde{\mathbf{A}}_u^{\top} \tilde{\kappa}_u^{\ddagger})_e & \text{if } e \in \tilde{\mathcal{E}} \\ v_{u,\mathsf{o}} - \tilde{\omega}_u^{\ddagger} p_{\mathsf{o}} & \text{if } e = \mathsf{o} \\ v_{u,\varnothing} - \tilde{\lambda}_u^{\ddagger} & \text{if } e = \varnothing. \end{cases} \tag{L.7}$$

Furthermore, using the complementary slackness condition, we obtain

$$0 = \begin{cases} v_{u,e} x_{u,e}^{\ddagger} - \tilde{\omega}_u^{\ddagger} p_e x_{u,e}^{\ddagger} - \tilde{\lambda}_u^{\ddagger} \tilde{a}_{u,e} x_{u,e}^{\ddagger} - (\tilde{\mathbf{A}}_u^{\top} \tilde{\kappa}_u^{\ddagger})_e x_{u,e}^{\ddagger} & \text{if } e \in \tilde{\mathcal{E}} \\ v_{u,\mathsf{o}} x_{u,\mathsf{o}}^{\ddagger} - \tilde{\omega}_u^{\ddagger} p_{\mathsf{o}} x_{u,\mathsf{o}}^{\ddagger} & \text{if } e = \mathsf{o} \\ v_{u,\varnothing} x_{u,\varnothing}^{\ddagger} - \tilde{\lambda}_u^{\ddagger} x_{u,\varnothing}^{\ddagger} & \text{if } e = \varnothing. \end{cases} \tag{L.8}$$

In order to prove Proposition 14.3.3, we show that if there exists ω^* such that $\omega^* = \lambda^{\dagger}(\omega^*)$ then $(\bar{\mathbf{x}}^{\dagger}(\omega^*), \mathbf{p}^{\dagger}(\omega^*))$ is a fractional-competitive equilibrium. It is sufficient to verify the following:

(i) By fixing the prices to $\mathbf{p}^{\dagger}(\omega^*)$, $\bar{\mathbf{x}}_u^{\dagger}(\omega^*)$ is an optimal solution of (L.1), for every $u \in U$;

(ii) the capacity constraints are satisfied at every resource;

(iii) $p_e^{\dagger}(\omega^*) \geqslant 0$ for every $e \in \tilde{\mathcal{E}}$; and

(iv) if $p_e^{\dagger}(\omega^*) > 0$ for some $e \in \tilde{\mathcal{E}}$, then $\sum_{u \in U} x_{u,e}^{\dagger}(\omega^*) = \ell_e$.

It is immediate to note that $(ii) - (iv)$ are satisfied due to dual and primal feasibility conditions of (14.3). It only remains to show (i).

To show (i), it is sufficient to show that the there exists $\tilde{\omega}_u^{\ddagger}, \tilde{\lambda}_u^{\ddagger}, \tilde{\kappa}_u^{\ddagger}$ such that the tuple $(\bar{\mathbf{x}}_u^{\dagger}(\omega^*), \tilde{\omega}_u^{\ddagger}, \tilde{\lambda}_u^{\ddagger}, \tilde{\kappa}_u^{\ddagger})$ satisfies the conditions (L.7)-(L.8), and the budget constraint in (L.1b) holds.

Setting the optimal Lagrange variable of (14.3b) with $\omega_u = \lambda_u$ then the optimal solution x^* of (14.3) is the solution of individual optimization problem for all players with price $\mathbf{p}^*$.

By primal optimality conditions in (L.5), we obtain

$$0 \geqslant \begin{cases} v_{u,e} - \frac{f_u(\bar{\mathbf{x}}_u^{\dagger})}{(w_u + \omega_u^*)} p_e^{\dagger} - \frac{f_u(\bar{\mathbf{x}}_u^{\dagger})}{(w_u + \omega_u^*)} \tilde{a}_{u,e} \lambda_u^{\dagger} - \frac{f_u(\bar{\mathbf{x}}_u^{\dagger})}{(w_u + \omega_u^*)} (\tilde{\mathbf{A}}_u^{\top} \kappa_u^{\dagger})_e & \text{if } e \in \tilde{\mathcal{E}} \\ v_{u,\mathsf{o}} - \frac{f_u(\bar{\mathbf{x}}_u^{\dagger})}{(w_u + \omega_u^*)} p_{\mathsf{o}} & \text{if } e = \mathsf{o} \\ v_{u,\varnothing} - \frac{f_u(\bar{\mathbf{x}}_u^{\dagger})}{(w_u + \omega_u^*)} \lambda_u^{\dagger} & \text{if } e = \varnothing. \end{cases} \tag{L.9}$$

The preceding equation is equivalent to the primal optimality condition of individual optimization problem in (L.7) if we select $\tilde{\lambda}_u^{\ddagger} = \frac{f_u(\bar{\mathbf{x}}_u^{\dagger})}{(w_u + \omega_u^*)} \lambda_u^{\dagger}$, $\tilde{\omega}_u = \frac{f_u(\bar{\mathbf{x}}_u^{\dagger})}{(w_u + \omega_u^*)}$ and $\tilde{\kappa}_u^{\ddagger} = \frac{f_u(\bar{\mathbf{x}}_u^{\dagger})}{(w_u + \omega_u^*)} \kappa_u^{\dagger}$.

Similarly, (L.8) is also satisfied with the same choice. Finally, we show that individual budget constraint (L.1b) holds. For this we use the complementary slackness condition in (L.6) by summing all three cases in (L.6). For every $u \in U$, we obtain

$$0 = \frac{(w_u + \omega_u^*)}{f_u(\bar{\mathbf{x}}_u^\dagger)} f_u(\bar{\mathbf{x}}_u^\dagger) - \mathbf{p}^{\dagger\top}\mathbf{x}_u^\dagger - p_{\mathsf{o}}\mathbf{x}_{u,\mathsf{o}}^\dagger - \lambda_u^\dagger(\tilde{\mathbf{a}}_u^\top \mathbf{x}_u^\dagger + x_{u,\varnothing}^\dagger) - \kappa_u^{\dagger\top}\tilde{\mathbf{A}}_u^\top \mathbf{x}_u^\dagger.$$

Consequently, using (14.3c)-(14.3d) we obtain

$$\begin{aligned} 0 &= (w_u + \omega_u^*) - \mathbf{p}^{\dagger\top}\mathbf{x}_u^\dagger - p_{\mathsf{o}}x_{u,\mathsf{o}}^\dagger - \lambda_u^\dagger \\ &= w_u - \mathbf{p}^{\dagger\top}\mathbf{x}_u^\dagger - p_{\mathsf{o}}x_{u,\mathsf{o}}^\dagger, \end{aligned}$$

where in the last equation we used the fact that $\omega^* = \lambda^\dagger$. This completes the proof.

L.3 Derivation of Inner Loop Updates in Algorithm 11

The updates in the inner loop in Algorithm 11 is derived based on ADMM updates for (14.4). We review the basic structure of the ADMM algorithm in Section L.3 and then derive the inner loop updates in Section L.3.

Review of ADMM algorithm

The Alternative Direct Method of Multipliers (ADMM) is a distributed convex optimization algorithm that decomposes a problem into smaller subproblems, solves them in parallel, and coordinates to find a global solution via dual updates [63, 62]. It is built on dual ascent and augmented lagrangian methods.

Consider the following optimization problem with separable cost structure:

$$\begin{aligned} \max_{\mathbf{x}\in X, \mathbf{y}\in Y} \; & h(\mathbf{x},\mathbf{y}) = h_1(\mathbf{x}) + h_2(\mathbf{y}) \\ \text{s.t. } & A\mathbf{x} + B\mathbf{y} = \mathbf{c}, \end{aligned} \tag{L.10}$$

where

(i) $X \subset \mathbb{R}^a, Y \subset \mathbb{R}^b$ are closed convex sets,

(ii) $h_1 : \mathbb{R}^a \to \mathbb{R}, h_2 : \mathbb{R}^b \to \mathbb{R}$,

(iii) $A \in \mathbb{R}^{s\times a}, B \in \mathbb{R}^{s\times b}, \mathbf{c} \in \mathbb{R}^s$.

Let $\mu \in \mathbb{R}^s$ be the dual multiplier of constraint in (L.10). Consider the following augmented Lagrangian function for (L.10) for some parameter $\beta > 0$

$$L_\beta(\mathbf{x}, \mathbf{y}) = h_1(\mathbf{x}) + h_2(\mathbf{y}) - \mu^\top (A\mathbf{x} + B\mathbf{y} - \mathbf{c}) - \frac{\beta}{2}\|A\mathbf{x} + B\mathbf{y} - \mathbf{c}\|^2.$$

The ADMM algorithm is a discrete-time algorithm, indexed by k, given as follows

$$\begin{aligned} \mathbf{x}^{(n+1)} &= \arg\max_{\mathbf{x}\in X} L_\beta(\mathbf{x}, \mathbf{y}^{(n)}) \\ \mathbf{y}^{(n+1)} &= \arg\max_{\mathbf{y}\in Y} L_\beta(\mathbf{x}^{(n+1)}, \mathbf{y}) \\ \mu^{(n+1)} &= \mu^{(n)} + \beta(A\mathbf{x}^{(n+1)} + B\mathbf{y}^{(n+1)} - \mathbf{c}). \end{aligned} \tag{L.11}$$

The parameter β is also referred to as the step-size parameter for the ADMM algorithm.

ADMM Updates for Inner Loop Solving

The inner loop in Algorithm 11 is nothing but the ADMM algorithm applied to (14.4).

For any $\beta > 0$, we form the augmented Lagrangian $\mathcal{L}(\bar{\mathbf{x}}, \mathbf{y}, \mathbf{z}, \lambda, \mathbf{p}, \tilde{\mathbf{p}})$ for (14.4) as follows

$$\begin{aligned} L_\beta(\bar{\mathbf{x}}, \mathbf{y}, \mathbf{z}, \lambda, \mathbf{p}, \tilde{\mathbf{p}}) = &\sum_{u\in U} (w_u + \omega_u) \log(f_u(\bar{\mathbf{x}}_u)) - \sum_{u\in U} p_{\mathsf{o}} x_{u,\mathsf{o}} \\ &- \sum_{u\in U} \tilde{\mathbf{p}}_u^\top (\mathbf{x}_u - \mathbf{y}_u) - \mathbf{p}^\top \left(\sum_{u\in U} \mathbf{y}_u + \mathbf{z} - \ell \right) - \sum_{u\in U} \lambda_u (\tilde{\mathbf{a}}_u^\top \mathbf{x}_u + x_{u,\varnothing} - 1) \\ &- \frac{\beta}{2} \sum_{u\in U} \|\mathbf{x}_u - \mathbf{y}_u\|^2 - \frac{\beta}{2} \left\| \sum_{u\in U} \mathbf{y}_u + \mathbf{z} - \ell \right\|^2 . \end{aligned} \tag{L.12}$$

The ADMM algorithm (as per (L.11)) are given as follows:

$$\begin{aligned}
\bar{\mathbf{x}}^{(n+1)} &= \underset{\bar{\mathbf{x}},\ \text{s.t. (14.3d)}-\text{(14.3e) hold}}{\arg\max} L_\beta(\bar{\mathbf{x}}, \mathbf{y}^{(n)}, \mathbf{z}^{(n)}, \lambda^{(n)}, \mathbf{p}^{(n)}, \tilde{\mathbf{p}}^{(n)}) \\
&= \underset{\bar{\mathbf{x}},\ \text{s.t. (14.3d)}-\text{(14.3e) hold}}{\arg\max} \sum_{u\in U} (w_u + \omega_u) \log(f_u(\bar{\mathbf{x}}_u)) - \sum_{u\in U} p_{\mathsf{o}} x_{u,\mathsf{o}} \\
&\qquad - \sum_{u\in U} \tilde{\mathbf{p}}_u^{(n)\top}(\mathbf{x}_u - \mathbf{y}_u^{(n)}) - \sum_{u\in U} \lambda_u^{(n)}(\tilde{\mathbf{a}}_u^\top \mathbf{x}_u + x_{u,\varnothing} - 1) \\
&\qquad - \frac{\beta}{2} \sum_{u\in U} \|\mathbf{x}_u - \mathbf{y}_u^{(n)}\|^2 \qquad \text{(L.13a)} \\
(\mathbf{y}^{(n+1)}, \mathbf{z}^{(n+1)}) &= \underset{\mathbf{y}\in\mathbb{R}^{U\times|\tilde{\mathcal{E}}|},\ \mathbf{z}\in\mathbb{R}_+^{|\tilde{\mathcal{E}}|}}{\arg\max} L_\beta(\bar{\mathbf{x}}^{(n+1)}, \mathbf{y}, \mathbf{z}, \lambda^{(n)}, \mathbf{p}^{(n)}, \tilde{\mathbf{p}}^{(n)}) \\
&= \underset{\mathbf{y}\in\mathbb{R}^{U\times|\tilde{\mathcal{E}}|},\ \mathbf{z}\in\mathbb{R}_+^{|\tilde{\mathcal{E}}|}}{\arg\max} - \sum_{u\in U} \tilde{\mathbf{p}}_u^{(n)\top}(\mathbf{x}_u^{(n+1)} - \mathbf{y}_u) - \mathbf{p}^{(n)\top}\left(\sum_{u\in U} \mathbf{y}_u + \mathbf{z} - \ell\right) \\
&\qquad - \frac{\beta}{2} \sum_{u\in U} \|\mathbf{x}_u^{(n+1)} - \mathbf{y}_u\|^2 - \frac{\beta}{2}\left\|\sum_{u\in U} \mathbf{y}_u + \mathbf{z} - \ell\right\|^2 \qquad \text{(L.13b)} \\
\lambda_u^{(n+1)} &= \lambda_u + \beta(\tilde{\mathbf{a}}_u^\top \mathbf{x}_u^{(n+1)} + x_{u,\varnothing}^{(n+1)} - 1), \quad \forall\ u \in U \qquad \text{(L.13c)} \\
\mathbf{p}^{(n+1)} &= \mathbf{p}_u^{(n)} + \beta(\sum_{u\in U} \mathbf{y}_u^{(n+1)} + \mathbf{z}^{(n+1)} - \ell) \qquad \text{(L.13d)} \\
\tilde{\mathbf{p}}_u^{(n+1)} &= \tilde{\mathbf{p}}_u^{(n)} + \beta(\mathbf{x}_u^{(n+1)} - \mathbf{y}_u^{(n+1)}), \quad \forall\ u \in U. \qquad \text{(L.13e)}
\end{aligned}$$

First, we claim the if $\mathbf{p}^{(0)} = \tilde{\mathbf{p}}_u^{(0)}$ for every $u \in U$, then $\mathbf{p}^{(n)} = \tilde{p}_u^{(n)}$ for every $u \in U$ and $n \in \mathbb{N}$. We prove this by induction. Suppose for some n, $\mathbf{p}^{(n)} = \tilde{p}_u^{(n)}$ for every $u \in U$ then we show that $\mathbf{p}^{(n+1)} = \tilde{p}_u^{(n+1)}$ for every $u \in U$. To see this, note from the first-order conditions of optimality for (L.13b) with respect to $\mathbf{y}$, we obtain

$$\tilde{\mathbf{p}}_u^{(n)} - \mathbf{p}_u^{(n)} + \beta(\mathbf{x}_u^{(n+1)} - \mathbf{y}^{(n+1)}) - \beta(\sum_{u\in U} \mathbf{y}_u^{(n+1)} + \mathbf{z}^{(n+1)} - \ell) = 0. \qquad \text{(L.14)}$$

Using preceding equation, we obtain

$$\tilde{\mathbf{p}}_u^{(n+1)} \underset{\text{(L.13e)}}{=} \tilde{\mathbf{p}}_u^{(n)} + \beta(\mathbf{x}_u^{(n+1)} - \mathbf{y}^{(n+1)}) \underset{\text{(L.14)}}{=} \mathbf{p}_u^{(n)} + \beta(\sum_{u\in U} \mathbf{y}_u^{(n+1)} + \mathbf{z}^{(n+1)} - \ell) \underset{\text{(L.13d)}}{=} \mathbf{p}_u^{(n+1)}.$$

This concludes our claim. Therefore, we get rid of notation $\tilde{\mathbf{p}}$ and work only with $\mathbf{p}$.

Finally, note that (L.13a) is separable in $\bar{\mathbf{x}}_u$ for every $u \in U$. Against the preceding backdrop, (L.13) can be re-written as

$$\bar{\mathbf{x}}_u^{(n+1)} = \underset{\bar{\mathbf{x}}_u,\ \text{s.t. (14.3d)}-\text{(14.3e) hold}}{\arg\max} \quad (w_u + \omega_u)\log(f_u(\bar{\mathbf{x}}_u)) - p_{\mathsf{o}} x_{u,\mathsf{o}} - \mathbf{p}_u^{(n)\top}\mathbf{x}_u - \lambda_u^{(n)}(\tilde{\mathbf{a}}_u^\top \mathbf{x}_u + x_{u,\varnothing} - 1) - \frac{\beta}{2}\|\mathbf{x}_u - \mathbf{y}_u^{(n)}\|^2 \tag{L.15a}$$

$$(\mathbf{y}^{(n+1)}, \mathbf{z}^{(n+1)}) = \underset{\mathbf{y}\in\mathbb{R}^{U\times|\tilde{\mathcal{E}}|},\ \mathbf{z}\in\mathbb{R}_+^{|\tilde{\mathcal{E}}|}}{\arg\max} \quad -\mathbf{p}^{(n)\top}\mathbf{z} - \frac{\beta}{2}\sum_{u\in U}\|\mathbf{x}_u^{(n+1)} - \mathbf{y}_u\|^2 - \frac{\beta}{2}\left\|\sum_{u\in U}\mathbf{y}_u + \mathbf{z} - \ell\right\|^2 \tag{L.15b}$$

$$\lambda_u^{(n+1)} = \lambda_u + \beta(\tilde{\mathbf{a}}_u^\top \mathbf{x}_u^{(n+1)} + x_{u,\varnothing}^{(n+1)} - 1), \quad \forall\ u \in U \tag{L.15c}$$

$$\mathbf{p}^{(n+1)} = \mathbf{p}_u^{(n)} + \beta(\sum_{u\in U}\mathbf{y}_u^{(n+1)} + \mathbf{z}^{(n+1)} - \ell). \tag{L.15d}$$

Updates (L.15) correspond to the inner loop updates in Algorithm 11, where (L.15a) is implemented locally by different AAM vehicles and (L.15b)-(L.15d) are implemented by service provider.

L.4 Vertiport Reservation Mechanism in Northern California

In this section, we study a scenario of vertiport reservation for (hypothesized) air taxi services in Northern California. We simulate a scenario where different air taxis request access to air routes to transport people at an urban and regional level. The vertiports in this simulation are located in various cities in Northern California as shown in the map in Fig. L.2. For simplicity, we are modeling linear trajectories and assuming a maximum travel range of 100 miles.

In this example, 20 air taxis request a departure, air route, and landing clearances among seven vertiport destinations during a 10-minute auction window. The requests by the air taxis, the final allocation of routes, and payments to the SP are presented in Table L.4.1 along with the maximum capacity in every segment of the desired routes, the utility of the air taxis for a given path, and their initial air credits. We set $\beta = 50, p_{\mathsf{o}} = 10, v_{u,\mathsf{o}} = 1, v_{u,\varnothing} = 1, N = 2, \mathsf{tol} = 1 \times 10^{-4}$. The *Maximum Capacity* column in Table L.4.1 specifies the maximum number of vehicles that can traverse a travel segment at any given time. These values help the reader identify contested travel segments and understand why certain agents must compete for access. In these tables, air taxis sharing the same color represent those that simultaneously requested the same trajectory slot, leading to a constraint violation. As a result, only a subset of these air taxis were granted their preferred route, while the others were denied access. We also present the rank number of these agents representing the order in which each agent computed their integral allocation, as outlined in Algorithm 2.

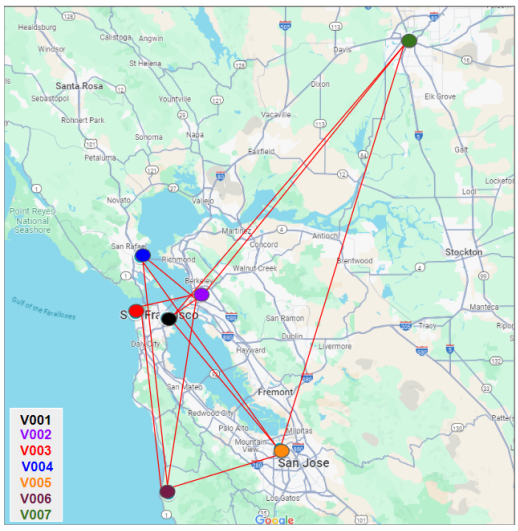


Figure L.2: Northern California Vertiport Map. This map, adapted from a Google Maps image, highlights seven distinct vertiports using unique color codes and displays the example routes as red lines.

Below, we highlight the main observations from our numerical study.

(i) At time step 16, AC003, AC004, and AC015 request departure from V002, which has a departure capacity constraint of one. Consequently, only AC004 is allocated to depart at this time step due to its higher air credits, while AC003 and AC015 are delayed. Naturally, these air taxis would prefer to depart at the next time step; however, they now compete for departure from V002 with AC013 and AC018 at time steps 17 and 18, respectively. Notably, Algorithm 12 prioritizes AC013 and AC018, resulting in further delays for AC003 and AC015. At time step 19, AC003 and AC015 compete again, with AC003 receiving priority due to its higher air credits in Algorithm 12.

(ii) AC002 and AC011 request landing slots at V004 at the same time, which results in delay for AC011. This is because AC011 has both a lower budget and lower utility in comparison to AC002.

(iii) AC009 and AC010 request departure from V001 at the same time, which results in delay for AC009. This is because AC010 has a significantly higher budget than AC009.

(iv) Air taxis that are delayed are charged less than those who are allocated their preferred routes.

Table L.4.1: Results of the allocation of air taxis to the desired routes, payments to the SP, utility, initial air credits, and maximum capacity in the en-route travel segment

Aircraft	**Req. Route (Orig., Dest.)**	**Req. Time (Arr, Dep)**	**Max. Capacity (Dep, Route, Arr)**	**Allocated Time (Arr, Dep)**	**Status**	**Price ($)**	**Initial Air Credits**	**Utility**	**Rank**
AC001	(V007, V002)	(16, 54)	(2,4,1)	(16, 54)	on-time	0.0	125	118	6
AC002	(V005, V004)	(19, 47)	(4,5,1)	(19, 47)	on-time	5.73	90	171	7
AC011	**(V006, V004)**	**(19, 47)**	**(1,2,1)**	**(20, 48)**	**delayed**	**1.53**	**78**	**135**	**19**
AC003	**(V002, V001)**	**(16, 21)**	**(1,1,2)**	**(19, 24)**	**delayed**	**3.71**	**135**	**172**	**18**
AC004	(V002, V001)	(16, 21)	(1,1,2)	(16, 21)	on-time	20.36	154	133	13
AC015	**(V002, V001)**	**(16, 21)**	**(1,1,2)**	**(20, 25)**	**delayed**	**0.86**	**65**	**194**	**20**
AC013	(V002, V006)	(17, 41)	(1,4,3)	(17, 41)	on-time	11.11	55	147	16
AC018	(V002, V007)	(18, 56)	(1,2,3)	(18, 56)	on-time	8.46	103	165	8
AC005	(V003, V002)	(11, 19)	(1,5,1)	(11, 19)	on-time	0.0	83	177	4
AC006	(V005, V007)	(18, 68)	(4,3,3)	(18, 68)	on-time	0.0	199	148	15
AC007	(V003, V002)	(15, 23)	(1,5,1)	(15, 23)	on-time	0.0	100	183	5
AC008	(V007, V001)	(12, 54)	(2,3,2)	(12, 54)	on-time	0.0	104	155	10
AC009	**(V001, V005)**	**(13, 34)**	**(5,1,2)**	**(14, 35)**	**delayed**	**2.10**	**67**	**189**	**17**
AC010	(V001, V005)	(13, 34)	(5,1,2)	(13, 34)	on-time	5.75	114	163	3
AC012	(V005, V001)	(16, 37)	(4,3,2)	(16, 37)	on-time	0.0	90	124	12
AC014	(V001, V002)	(11, 24)	(5,2,1)	(11, 24)	on-time	0.0	64	174	9
AC016	(V007, V005)	(17, 67)	(2,5,2)	(17, 67)	on-time	0.0	109	189	14
AC017	(V004, V006)	(16, 44)	(5,3,3)	(16, 44)	on-time	0.0	155	149	11
AC019	(V004, V002)	(16, 35)	(5,5,2)	(16, 35)	on-time	0.0	104	147	1
AC020	(V003, V006)	(16, 38)	(1,2,3)	(16, 38)	on-time	0.11	96	146	2